Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

- General background
- Focused introduction (write as if 1-2 years ago and then introduce subsequent studies)
- State of the art
- (Contributions - what is each contribution and where located)
- Aims

1.1.1 Synthetic Biology

Synthetic biology is a newly emerging field within the scientific community which lies at the junctions of many disciplines. One of the key aims of synthetic biologists is to apply an engineering approach to the design and construction of biological systems. This approach has the potential to create a new wave of applications and technologies which *"will influence many other scientific and engineering disciplines, as well as affect various aspects of daily life and society"*[?]. However, as with the development of any other new scientific discipline or engineering field, there are numerous challenges which must be overcome before this potential can be reached.

The literature offers myriad definitions of synthetic biology, though a widely accepted one is that synthetic biology is A) *"the design and construction of new biological parts, devices,*

and systems," and B) *"the re-design of existing, natural biological systems for useful purposes"*[?]. Analogies between synthetic biology and electrical or computer engineering have been made[?]. Although there exist clear departures from this analogy, many parallels can be drawn between the two and it can be useful for people trying to understand what synthetic biology is, and the approach to its development, to keep this in mind. Figure 1.1 is taken from Andrianantoandro et al.[?] and shows a comparison between the hierarchical structure of synthetic biology and a similar hierarchy in computer engineering.

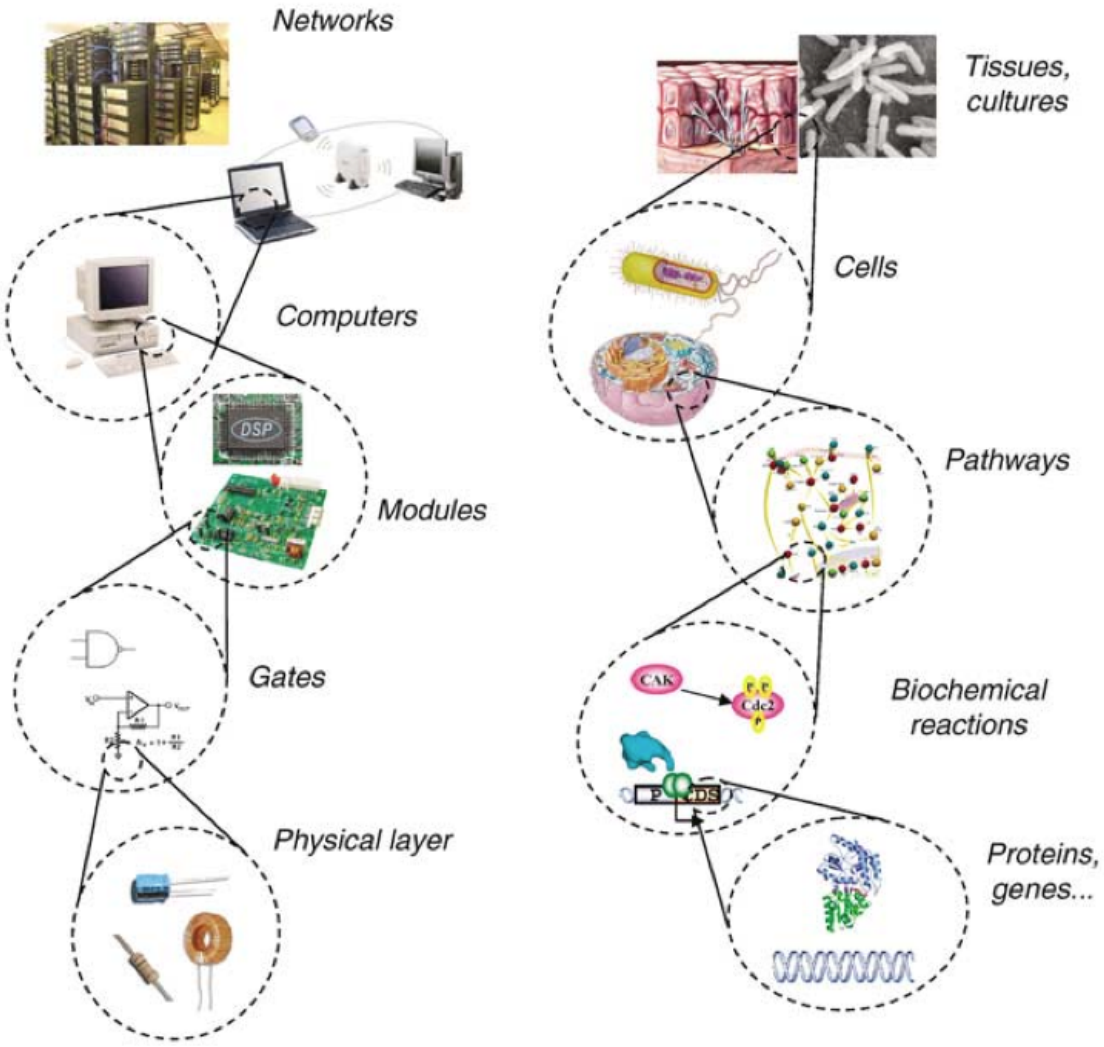


Figure 1.1: An analogy between synthetic biology and computer engineering[?]

The expression of the genes from a synthetic circuit can impart extra functionality on a cell, such as the ability to detect arsenic[] or perform logical operations[]. There is a diverse range of applications that affect a broad range of fields such as health, such as biofilm detection in catheters[] to energy where next generation biofuels are begin developed[] etc. The synthetic biology solutions to the problems within this wide range of fields of potential applications are

currently limited by a number of technical and scientific factors[?]. As a field that is now currently emerging from its infancy, there are a number of 'growing pains'. When the first synthetic biology research started to emerge almost a decade and a half ago, the circuits being reported were relatively simple and used a small range of biological parts[]. However, as the field has progressed there is an increased demand for better tools and understanding of the underlying biology to facilitate more advanced applications.

Synthetic biology research can be split into two main sub-fields: fundamentals and applications. The fundamentals side of synthetic biology concentrates on answering fundamental questions about the nature of life and the underlying processes which occur in biology[?] whilst the applications side uses this knowledge to either reengineer existing systems or to construct new systems in a way which allows for the development of novel functions and applications[?]. These two sides are intrinsically linked by the need for greater understanding to facilitate improved methods and techniques for the generation of applications. In this project we aim to gain a greater understanding of some of the most important fundamentals of implementing synthetic biology - the link between the chassis cell and a synthetic genetic circuit - with results which will directly impact upon the ability of researchers to create future applications.

Definitions

At this point it is important to introduce some definitions that are crucial for understanding synthetic biology and especially this study.

Synthetic Genetic Circuit A *synthetic genetic circuit* (or *synthetic circuit* in the context of synthetic biology) is a section of DNA which is inserted into the cell and contains sequence for the expression of certain mRNA or proteins. For *E. coli* his is usually inserted in the form of a plasmid, a circular strand of DNA which is 'separate from, and can replicate independently of, the chromosomal DNA'[?] or inserted directly into the genome of the cell.

Chassis/Host Cell The cell which contains this synthetic circuit is known in synthetic biology as the *chassis*, or *host*, cell. The synthetic circuit uses machinery from this cell in order to replicate itself and to express the genes it contains. Returning to the computer engineering

analogy, one can imagine the chassis cell as a computer and its operating system, while the synthetic circuit is the code for a programme on that computer.

Shared Resources The term *shared resources* refers to the machinery and building blocks (such as amino acids and nucleotides) required by both the host cell and synthetic circuit for the replication of DNA, transcription of RNA and translation of protein. Examples of these resources are ribosomes, RNA and DNA polymerases etc.

Burden The *burden* placed on a chassis cell's shared resources by a synthetic circuit refers to the resources that are used by the synthetic circuit and therefore not available for the cell to use for its own native processes. This burden on shared resources is one of many types of stress such as nutrient starvation, nitrogen starvation, physiological stresses etc. and in the same way that cells respond to these stresses, cells will respond in different ways to burden (such as the stringent response in some strains - see Section 1.1.5).

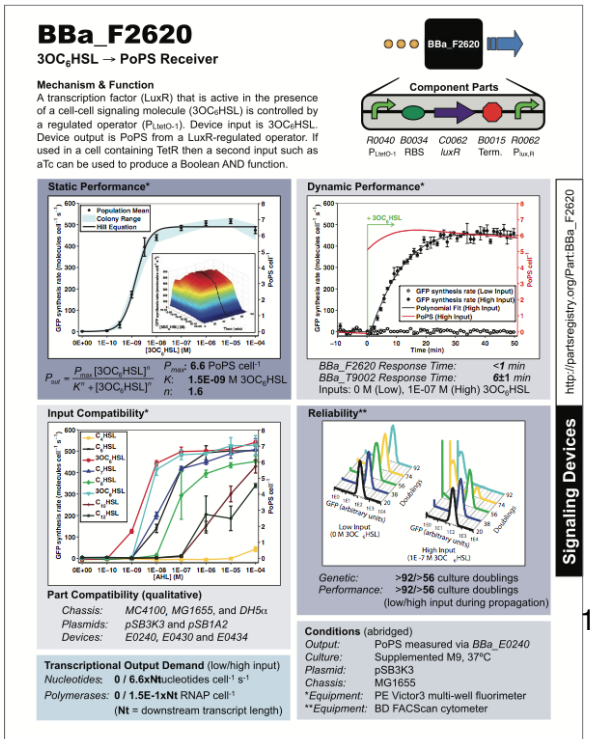
Capacity The *capacity* in a cell is the amount of free resources that are not being used by either a synthetic circuit or the chassis cell. An increased burden from a synthetic genetic circuit will, dependent on any cellular feedback in response to the burden, reduce the capacity in a cell. Certain strains of *E. coli* cell may have a larger 'unburdened' capacity than others.

Although synthetic biology promises much, there still remain many challenges to overcome[? ?]. One of these challenges regards standardisation and the manner in which it can be implemented within the field. Within biology there are many examples of standards, ranging from microarray data[?] to restriction endonuclease activities[?]. An attempt has been made to standardise the manner in which genetic parts are characterised in Canton et al.[?], where an example data-sheet akin to those seen in other engineering disciplines is shown[?] (see Figure 1.2). Although data about individual parts can be collected and displayed[?], this is not sufficient for making predictions. As Jay Keasling, one of the foremost experts in the field of synthetic biology, claims *"even if the function of each part is known, the parts may not work as expected when put together"*[?]. This principle is not only true for part-part combinations, but it has also been seen that a circuit may not act as predicted due to interactions with the host cell. This can be seen in numerous examples such as Tan et al.[?].

In Tan et al. they explore the interactions between a host cell and a synthetic circuit. They observe that these interactions can change the behaviour of a circuit and cause it to perform with characteristics that would not be possible when predicting circuit behaviour from a model of the circuit in isolation. They introduce a simple 2 gene non-cooperative feedback loop into *E. coli*. This control structure on its own does not allow for bistability, however the circuit is observed to have this behaviour. This is due to an additional negative feedback loop that is introduced through interactions between the cell and the circuit doe to changes in growth rate. This study shows that these interactions are not always negative, and indeed can be used for positive effects by enabling otherwise unachievable circuit behaviours.

An understanding of how a circuit interacts with other circuits and its host chassis cell may allow us to design circuits that have specific behaviours (such as introducing bistability through cell-circuit interactions[?]). This understanding may also allow circuits to be optimised so that the level of interactions with the cell is minimised for a given output. Alternatively, for a given 'budget' of burden the output of a desirable product may be optimised (such as the production of a valuable protein).

A key tenent of any engineering discipline is predictability, where the behaviour of a system can be reliably predicted from an understanding of its component parts[]. This equally applies to synthetic biology, though the large levels of complexity often means making reliable predictions can be difficult, if not impossible in some cases. Moving synthetic biology towards being a discipline where designs with predicable functionalities can be made requires an increased understanding of the underlying biology and a modelling framework to be in place.



The creation of models allows predictions to be made about biological systems and the functions of parts. This was shown in Ellis et al.[?] where a model-based approach was used in conjugation with the creation of a library of promoters of varying strength. This approach allowed a single test design to be used to inform a model which was able to accurately predict how a range of circuit versions would behave. A reliable ge-

netic timer switch was created where 'tweaks' in behaviour were able to be rationally designed rather than performed by retrofitting networks. It is important that we are able to expand and extend this model-based approach to the design of full large-scale genetic circuits and cell-circuit combinations.

The reason we are currently unable to make accurate predictions about circuits with expression levels where the cellular context is important is due to the fact that there is no explanation for how the data presented in data-sheets[?] or databases[?] can be used together to make these predictions. These 'rules of composition' are vital and, as claimed in Andrianantoandro et al., will *"help determine which device combinations yield the desired logic functions and, more importantly, how to match cellular or physical functions of devices"*[?].

There have been attempts to develop basic frameworks for predicting the behaviour of genetic parts when combined. Marchisio et al.[?] demonstrate a novel approach to this by modelling systems based on fluxes of cellular machinery such as polymerases, ribosomes, transcription factors and environmental signals[?]. This modelling framework was implemented within the ProMoT systems modelling and design tool and allows for individual parts to be composed into larger networks, something that is key in making synthetic biology design predictable and modular. The models used in this study are very simple and lack some key considerations (such as codon usage etc, which we discuss later in this introduction). There is also no provision of a methodology for characterising the individual parts to obtain the information required by the model to make predictions. However, an expansion of this approach may lead to an improved framework whereby predictions can be made about cell-circuit interactions.

Any part can be characterised in a number of different ways, and it is standardisation which will ensure this is done in a consistent manner for all parts. A number of approaches can be taken to part characterisation in terms of what data to collect, however it is vital that before a standard is implemented we understand what datasets will be required in order to use a model to make predictions. Figure 1.3 shows the different variables that can impact upon levels of commercial protein production, the simplest example of a gene expression application for bacterial synthetic biology.

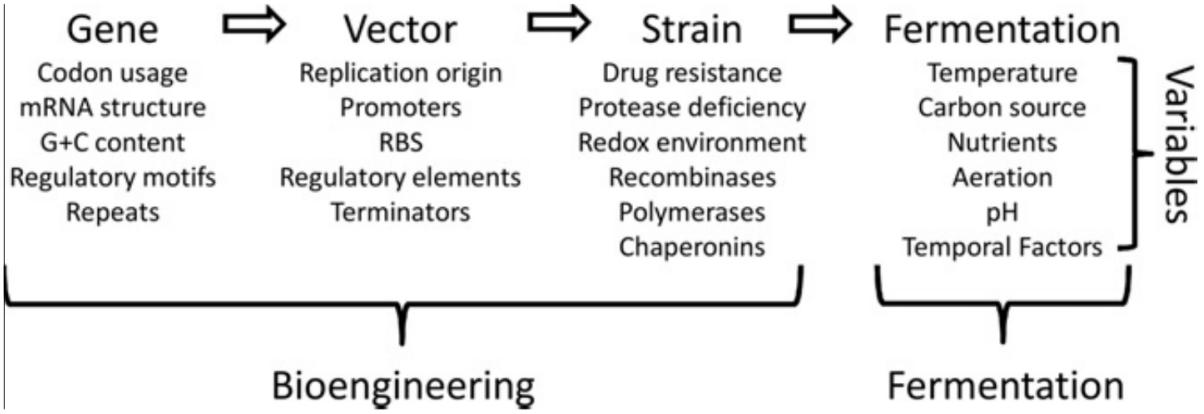


Figure 1.3: The factors impacting protein production levels can be split into different categories. We are interested in both gene and vector variables. Figure adapted from Gustafsson et al. **??**

Therefore, in order for standardisation to enable accurate and portable predictions it is necessary that standards are developed for both the collection and displaying of data as well as how this data from individual parts can be combined (and additionally how this data can be combined with species or strain characterisation to make predictions about how a circuit will act within different cellular contexts). These two types of standard must clearly be developed together as one loses much of its worth when not used with the other. This project aims to develop both a standard for characterising cells and circuits as well as a methodology for using this characterisation data to make accurate predictions about how a circuit and cell will behave when combined.

1.1.2 Cell-circuit and circuit-circuit interactions

As discussed above, the nature of biological systems is such that it is often not possible to insulate the activities of two or more biological processes. We use the term *isolate* in this context to mean that two circuits are isolate if each of their behaviour is independent of the presence of the other circuit. In Hajimorad et al. they express three proteins and investigate their interdependence through shared usage of transcriptional resources. This study shows that it is very difficult to isolate the behaviour of the devices[**?**].

It is often the case that the behaviour of one process is dependent on the behaviour of others[**?**]. The impact of cell-circuit interactions is neatly summed up by Cardinale et al. when they say that ”the last decade has shown that predictable engineering of cell functions is ham-

pered by ignorance of the host factors that affect and are affected by the engineered pathway and that lead to non-optimal or undesirable behavior" (where a pathway is a type of circuit)[]. These undesirable interactions are often unavoidable, and as such it is important that they are understood, able to be predicted and, if possible, their effects able to be minimised. In synthetic biology, the context of the cellular environment for a synthetic circuit and their interactions are very important and there are numerous ways in which the circuit can interact with its host cell. The effects of these interactions can be wide-ranging in their scale, from a linear relationship between gene copy-number where the interactions can be ignored whilst maintaining predictability in terms of cell and circuit activity[**?**], all the way to a situation where the presence of the synthetic circuit causes non-viability of the host cell[**?**].

Interactions between the cell and the circuit can be divided into two general types. Firstly there are the generic interactions that impact upon the behaviour of both the synthetic circuit and the host cell and are due to variations in global dynamics and shared resources. Secondly are the specific interactions that occur because of the choice of specific strains of host cells or parts within the synthetic circuit. These specific interactions are often avoidable through the choice of suitable combinations of host cells and synthetic circuits. Moser et al. show that the selection of both host strain and growth media can impact upon the behaviour of a synthetic circuit[]. This is due to the cell having different characteristics such as proteome (the entire set of proteins expressed by a cell at a certain point in time) and transcriptome (the set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA produced in a cell) when growing in different growth media.

In addition to cell-circuit interactions there is also the issue of circuit-circuit interactions whereby the combination of multiple parts within a single circuit, or alternatively the inclusion of a number of separate circuits, causes these parts or circuits to act differently to how they would in isolation, for example the addition of an extra gene within a circuit may cause the expression levels from the original circuit to drop compared to the original alone[**? ? ?**]. It is therefore important when attempting to make predictions about the behaviour of a system to understand how its components interact with each other.

Interactions that occur when introducing a synthetic genetic circuit into a chassis cell can occur in numerous ways such as: toxicity, where the proteins produced by the circuit are toxic to the host cell[**?**]; cross-talk, where the circuit expresses mRNAs or protein which interfere with

the native regulatory mechanisms of the host, or there may be molecules in the cell, such as transcription factors, which affect the behaviour of the circuit[?];and shared resource pools, where expression from the circuit requires the use of the same resources and machinery as native cellular processes[?]. Since toxicity and cross-talk are both types of specific, avoidable interactions (toxicity is only an issue when expressing toxic proteins and cross-talk is avoidable through the use of regulatory mechanisms orthogonal to those of the host cell) we will ignore these aspects of cell-circuit interaction from now on and concentrate on shared resource pools.

In 2000 two papers were published which are widely considered to be the first true synthetic biology papers, these were the repressilator circuit by Elowitz et al. (a three gene genetic oscillator)[?] and the toggle switch circuit by the Gardner et al.[?]. In Gardner et al.[?] the synthetic circuit constructed is referred to as an *'applet'*, a circuit designed to be a self-contained genetic circuit which ran with minimal impact on the native cellular processes[?]. Research in synthetic biology has since increasingly moved from looking at 'toy' circuits, such as logic gates[?] towards increasingly application driven projects where higher expression from circuits is often required, such as projects with metabolic pathways[?], or with large and complex signalling circuits[?]. Increased expression and well as an increase in the number of genes and parts correlates to increased 'burden' and the interactions between the circuit and its host cell become more significant.

This requirement for increased output from the genetic circuit, whilst maintaining the important characteristics of its behaviour, is not always trivial due to, amongst other things, the non-linearity of changes in the behaviour of both the cell and circuit due to their interactions at high levels of circuit expression[?]. A result of this limit is the relatively low maximum number of promoters observed (and thus genes) in circuits reported in synthetic biology publications before 2009[?] (see Figure 1.4). In order to effectively and predictably increase gene expression levels it becomes important to understand the mechanisms by which the cell and circuit interact, and to use this understanding to build up models of cell-circuit interactions which can reliably predict the behaviour of both.

Cells can react in a number of way to the introduction of a synthetic circuit. Different circuits will induce different reactions ranging from almost no change in the cell physiology to cell death[]. Often the growth rate of the cell will decrease in response to heterologous protein production due to the generic effects mentioned above. Evolution of the synthetic circuit may also occur,

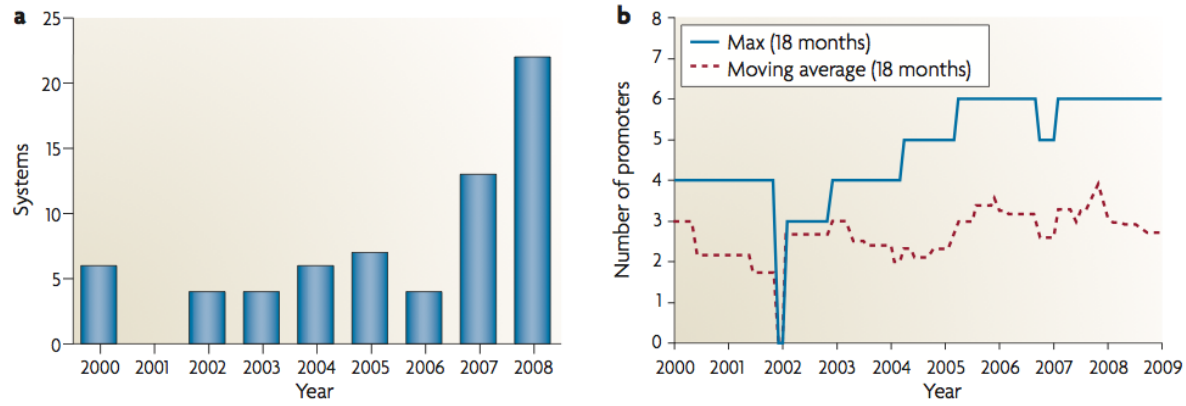


Figure 1.4: Purnick & Weiss[?] *"sampled publications that describe the construction and characterization of complete synthetic biological circuits from 2000 to 2008. Each publication can contain more than one circuit (each circuit is referenced here as a system).* ***a*** *— The number of synthetic systems in publication. The number of new synthetic biological systems increased moderately from 2000 to 2008.* ***b*** *— The complexity of synthetic systems in publication. For the purposes of this analysis, we define complexity as the number of regulatory regions (promoters) comprising any given synthetic system. Shown are 18-month moving window averages and maximum values. Although the overall number of synthetic systems has increased over a 9 year span (as shown in part a), the complexity of published systems seems to have reached a plateau (at least for now)."*

where cells have either evolved to grow without the circuit, or had recombination events occur where parts that are disadvantageous to cell growth have been mutated or deleted in response to burden[]. While these events are relatively rare, a beneficial impact on growth rate occurs and they quickly outcompete other cells and become dominant within a population, leading to a population mostly lacking the desired circuit[].

1.1.3 E. coli

As mentioned previously, the organism we are using in this study is *E. coli*, specifically the K-12 line of strains. *E. coli* is a widely used model organism that is a gram-negative bacterium. It is widely used in synthetic biology[] and has been widely characterised **??**. It is the dominant prokaryote model organism used for biological research. There are a large number of strains of *E. coli* that have been developed for different purposes, including research and industry[]. *E. coli* are typically rod-shaped, and are about 2.0 microns (μm) long and 0.5 μm in diameter, with a cell volume of 0.60.7 μm^3. The *E. coli* genome consists of approximately

The wild-type K-12 strain is MG1655, however, the strain used through most of this study is *DH10B*. This is a commonly used strain designed for the propagation of large insert DNA library

clones. It is widely used in research, where its properties such as high DNA transformation efficiency and the ability maintenance of large plasmids are taken advantage of[?]. DH10B has the *relA1* and *spoT1* allele which inactivates the protein responsible for ppGpp production, during a stringent response[?]. This means that the stringent response phenotype is not present in DH10B. The construction path of DH10B can be seen in Figure 1.5.

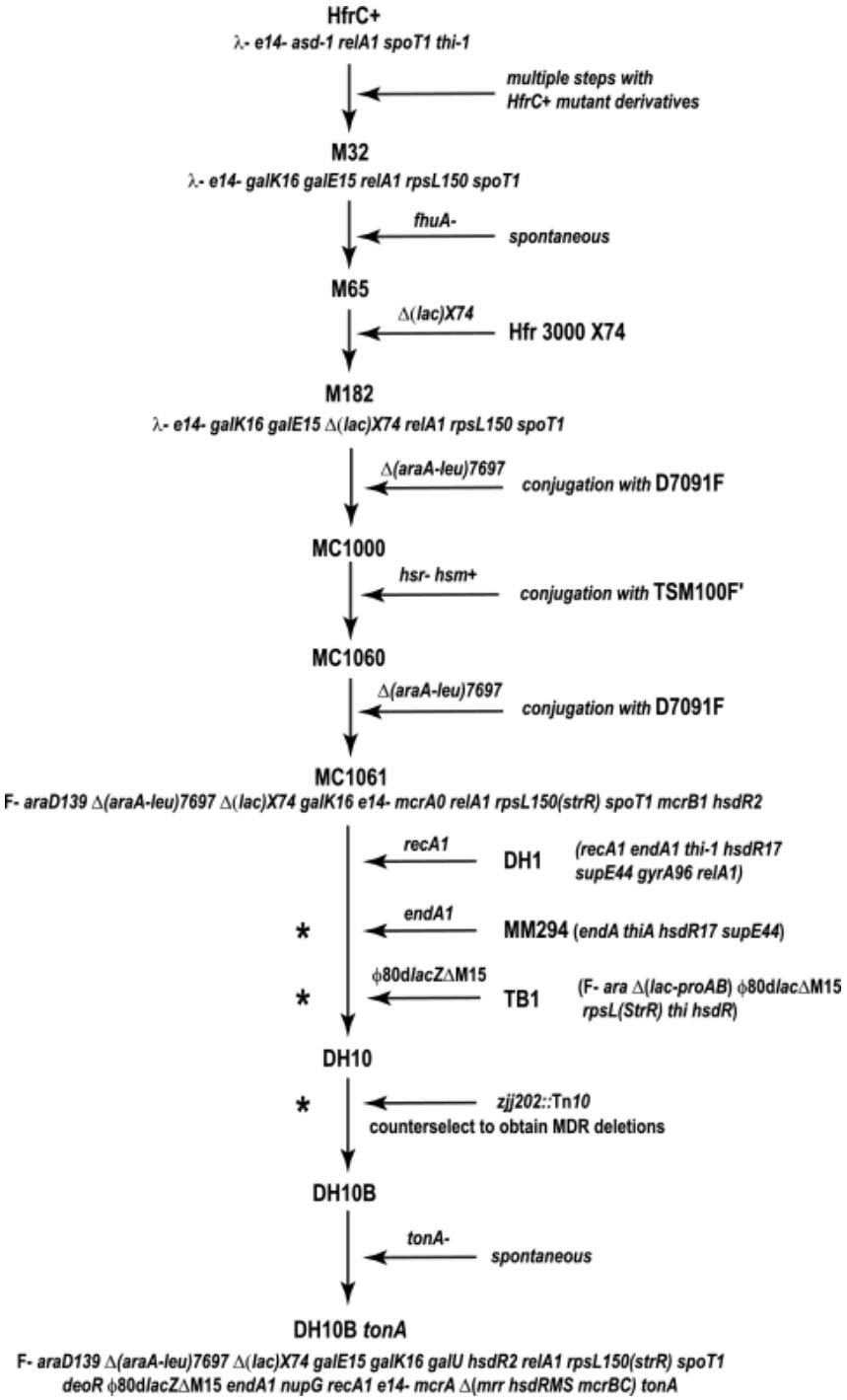


Figure 1.5: Construction of DH10B. The steps leading to the creation of DH10B from HfrC+ are outlined. From Durfee et al.[?].

1.1.4 Shared resource pools

The cellular genetics of individual species have been optimised through evolution to maximise fitness[]. In this means that the mechanisms for cell growth and maintenance are balanced such that there are enough resources available to the cell for maximal growth rate in desired environments whilst maintaining a limited level of *extra capacity* to deal with environmental perturbations such as change in carbon source which leads to new enzymes needing to be made and thus a change in transcriptome and proteome[].

Once again we return to the computing analogy. In computing there exist a wide range of programmes, ranging from small ones with very little impact upon the computer's resources such as RAM or processor availability, all the way to much larger ones which require a huge amount of the computers resources to run and can even cause the computer to crash or stall. In much the same way a synthetic circuit can have a whole range of impacts upon the host cell ranging from small background circuits which have no significant impact upon the host cell, to circuits which cause huge amounts of burden upon the cell's resources, sometimes even causing cell death CITE.

Upon the introduction of a synthetic genetic circuit into a chassis cell, the replication of and expression from the synthetic circuit requires the usage of native cellular resources and machinery. Replication of the synthetic DNA (DNAx) within the cell requires the same machinery used to replicate native cellular DNA (DNAc) such as DNA polymerase and deoxyribonucleotides. Depending on the system used to maintain the DNAx in the cell, there may also be competition for factors involved in the initiation of DNA replication[**?**]. There will also be competition between DNAx and DNAc for proteins which bind DNA, such a RNA polymerase (RNAP) which is central to transcription. RNA polymerase binds both specifically and non-specifically to DNA, both types of binding lead to the redistribution of RNAP between DNAs and DNAc on the basis of number of bases and promoter strength and number respectively. In the process of transcription RNA is produced by RNAP as well as a cohort of other enzymes and factors (such as ribonucleotides) associated with transcription, all of which are competed for in the production of RNA from DNAx and DNAc. Translation is also subject to the same types of competition for resources as transcription and replication. mRNA produced from the synthetic circuit (mRNAx) and mRNA derived from the native cell (mRNAc) are templates for protein production a process which uses ribosomes and amino acids, along with a number of other factors. The

balance of protein derived from the DNAx and DNAc will be a function of the balances of mRNAx and mRNAc as well as the affinities of the ribosome binding sites on these mRNAs for ribosomes[?].

This study focuses on cell-circuit interactions arising solely from the competitive interactions between the cell and circuit for shared resources. In order to correctly predict and simulate the response of the host cell to the insertion of a synthetic circuit it is important initially to understand and evaluate the kinetics of the processes common to expression from both DNAc and DNAx[?]. There are two key steps in this process; firstly to understand how the interactions occur which enables the creation of a mechanistic model followed by implementation and analysis of this model. Studies have been done to build up models of the link between synthetic circuit and the host cell through shared resource pools[? ? ? ? ? ?].

These differ in the approaches taken to the modelling process both in terms of the way cellular processes are described, as well as the depth to which these processes are described. It is important to understand the crucial interactions that cause changes in the behaviour of cells and circuits, as well as the important variables affecting these interactions. In Klumpp et al.[?] they model a direct link between gene expression levels and cell growth rates, however they do not include the mechanisms which link these two processes. Scott and Hwa argue that gene expression is 'intimately coupled to the growth state of the cell' CITE in the sense that knowing growth rate, the rate of gene expression can be readily predicted. However, they acknowledge that the study was only performed by media-based limits on growth rates and that 'relationships between ribosomal content, protein expression and growth rate must be characterised under other modes of growth inhibition'[].

Klumpp et al.[?] explore the relationship between growth rate and a range of cellular metrics that influence gene expression such as cellular RNA levels and protein production rates. They alter growth rates using a range of 5 different growth medias and show that there is a proportional relationship between the growth rate and levels or RNA in the cell. In addition they show that there is no growth rate dependence for the translation rate in cells (defined as total cellular protein divided by total cellular RNA). They argue that cells use feedback mechanisms involving growth rate to auto regulate certain key processes and also suggest the possibility for designing circuits which utilise growth feedback in there behaviour (as seen in Tan et al.[?]).

1.1.5 Translation and ribosomal usage

It has been indicated in previous research that ribosomes are the key factor in the limitation of growth rates due to over expression of proteins[? ? ?]. This suggests that an intelligent approach to understanding cell-circuit interactions would be to investigate, in depth, the process of translation. This would involve understanding how ribosomes are used by a synthetic circuit (including identifying the key variables which affect this usage) as well as how a change in the number of available ribosomes can affect native cellular processes and how these effects manifest themselves in terms of the cellular phenotype. An understanding of these processes may allow us to design circuits that make more efficient use of shared resources to produce proteins.

As mentioned before, the ribosome is the molecular machine which converts information encoded in the sequence of an mRNA into a chain of amino acids which forms a protein. It is composed of ribosomal RNA (rRNA, which is also involved in the regulation of ribosomal protein production) and ribosomal protein and is formed from a large and a small subunit (bacterial 50S and 30S respectively). As well as being sequestered by synthetic circuits, levels of available ribosomes in the cell can also change due to environmental variation and different nutrient availability[?]. Ribosomes form a very large portion of both the RNA and protein found in a cell and much of their activity involves generating new ribosomes[?].

Stringent Response

One of the key mechanisms through which ribosome numbers are regulated in *E. coli* is the *stringent response*[?] which recognises that there has been a drop in charged transfer RNA (these are type of RNA which are bound to an amino acid, or 'charged', and take part in the elongation stage of translation - see below) through an increase in the concentration of uncharged tRNA (transfer RNA that are not bound to an amino acid and need to be bound before they can take part in translation), and signals to a protein RelA to synthesise the *alarmone* ppGpp[?]. ppGpp[?] then causes a decrease in the amount of time most promoters are available for transcription to initiate, causing a down-regulation in synthesis of stable RNA, including rRNA, which in turn down-regulates the synthesis of the entire translation apparatus[?]. Thus we can see how over-expression of genes within a cell could lead to a systematic

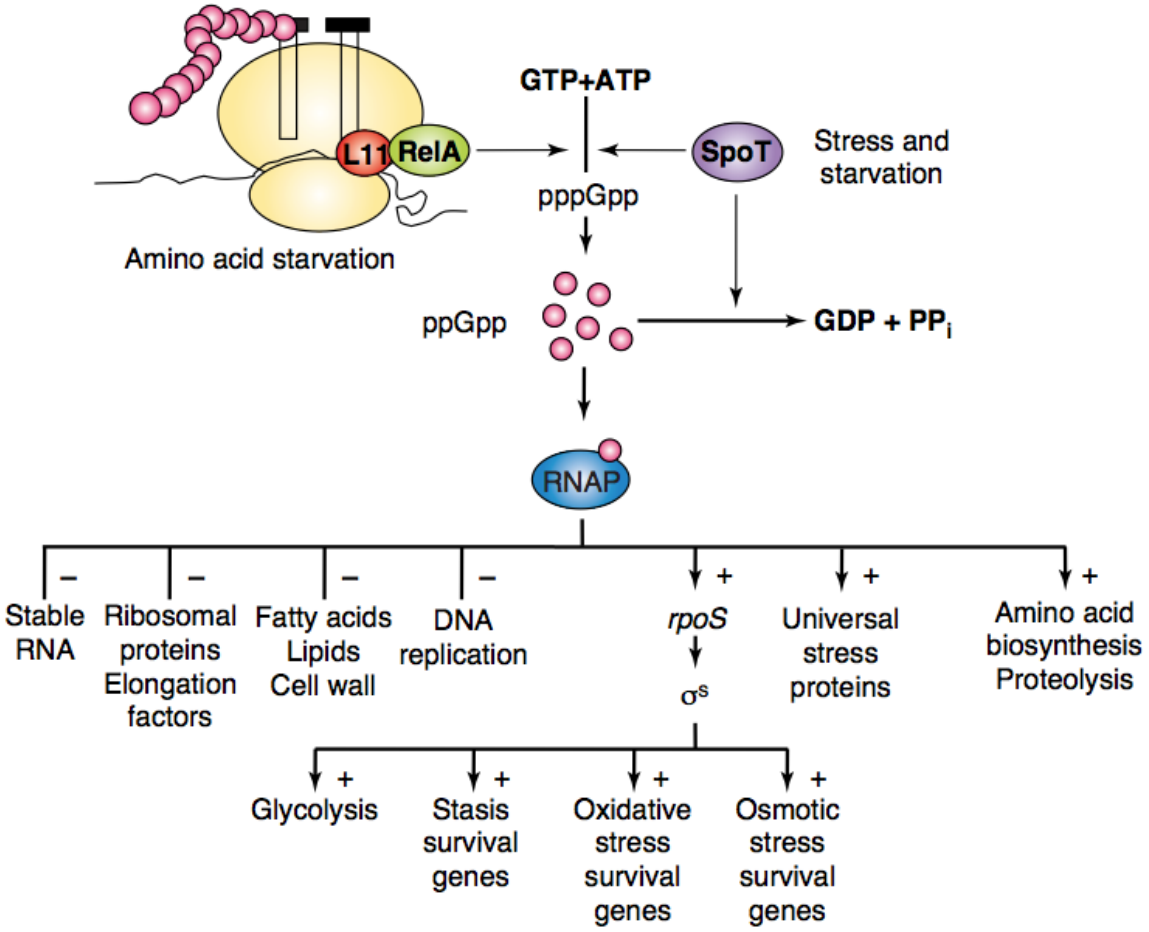


Figure 1.6: Diagram of the stringent response in bacterial cells, adapted from Magnusson et al.[?]

change in global translation rates through decreased ribosome levels, heavily impacting upon the cell. However, many *E. coli* strains used in modern biotechnology have the stringent response mechanisms disabled by deletion of the *relA* gene. One such commonly used strain is **DH10B**[?] which we will be using in this project. These experiments will be complemented by also using MG1655 (the K-12 wild type), which has the stringent response intact. Figure 1.6 shows a diagram of the stringent response, including which pathways are up-regulated and which are down-regulated.

Translation

Translation occurs in 4 phases: initiation, elongation, termination and ribosome turnover[?]. Initiation involves the assembling of the ribosomal subunits on the 5' end of the mRNA at the *ribosome binding site* (RBS). Elongation occurs as the ribosome moves along the mRNA and

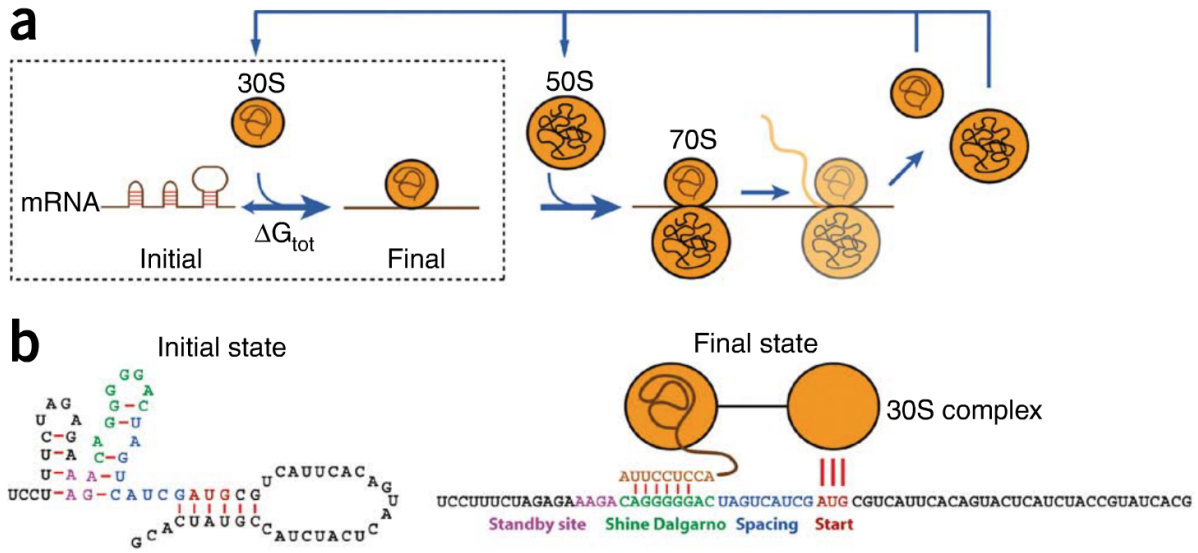


Figure 1.7: a) Translation is a multi-step process: the assembly of the 30S complex (box), initiation, elongation, termination, and the turnover of ribosomal subunits and other factors. b) mRNA folding, Shine-Dalgarno sequence and spacing distance all play important roles in the binding of the 30S subunit to the RBS. Figure adapted from Salis et al.[?].

a peptide chain is synthesised as transfer RNA (tRNA) that have been 'charged' with an amino acid read base triplets (codons) corresponding to certain amino acids. This process occurs until the ribosome reaches a stop codon and termination occurs, whereby the synthesis of the protein finishes and the unfolded protein is released from the ribosome. Ribosome turnover is the next stage as the ribosome breaks up off the mRNA and returns to the free ribosome pool[?]. Figure 1.7a shows an overview of the translation process.

Initiation is a complex process and the rate at which it occurs determines how quickly ribosomes will bind to an mRNA with an empty RBS (an RBS which does not have a ribosome bound to it). This rate depends on the sequence of the RBS and is affected by the binding affinity between the 16S rRNA and the RBS as well as secondary structures formed by the RBS and surrounding mRNA[?]. Salis et al. developed a statistical thermodynamic model of initiation to develop a calculator which can be used to predict the initiation rate of a given RBS sequence as well as design RBS sequences which will have a desired initiation rate[?]. Translational initiation is a highly complex process that involves multiple agents and factors. The 5' sequence of a CDS interacts with the RBS site and can encourage or discourage secondary mRNA structures that might affect ribosomal binding and the sequence of the RNA at this point governs the free energy of secondary structures where higher energies correspond to reduced time during which the the Shine-Dalgarno sequence is exposed. When the Shine-

Dalgarno sequence is unblocked, it interacts with the anti-SD sequence of the 16S rRNA[?]. The ejection of initiation factors then cause binding by the 50S subunit to form the 70S initiation complex which concludes the initiation phase. Initiation is governed by a number of factors including mRNA secondary structure, Shine-Dalgarno sequence and spacing distance between the Shine-Dalgarno sequence and start codon.

Each step in elongation (the β_is in figure 1.8) does not occur at the same rate and the dynamics of this process are dependent on a cohort of factors such as *"gene sequences, the tRNA pool of the organism and the thermodynamic stability of the mRNA transcripts"*[?]. These varying rates can have large effects on how efficiently ribosomes translate a protein and certain configurations of codons can cause ribosomal traffic jams where a section of codons can cause ribosomes to translate more slowly and cause other ribosomes to bunch up behind it[? ?]. The speed of codons is related to the availability of corresponding charged tRNAs[] and it has been shown that under nutrient starvation the relative speeds of codons can change[?].

Initiation and elongation rates affect both the protein production rate as well as the number of ribosomes sequestered on the transcripts at any given point in time[? ?]. It is clear these variables will affect ribosome availability within the host cell and therefore impact on its behaviour. Different genes place different levels of burden on the cell and have different production rates due to a range of factors such as promoter and RBS strengths. However, Welch et al.[?] have shown that even with the same promoter and RBS strengths, changing the coding sequence, even whilst conserving the amino acid sequence and only switching synonymous codons , can vastly change the amount of protein produced (from undetectable to 30% of total protein) and thus the impact on the host[?].

It has been shown in Tuller et al. that a certain profile of translational efficiency along mRNAs is conserved universally[?] in nature. In this profile codons which are translated with low efficiency are found with higher frequency among the first 30-50 codons. Tuller et al.[?] 'suggest that the slow 'ramp' at the beginning of mRNAs serves as a late stage of translation initiation, forming an optimal and robust means to reduce ribosomal traffic jams, thus minimising the cost of protein production' in terms of ribosomes sequestered[?]. Although this profile may cause protein to be produced at a lower rate, it may require less ribosomes to be sequestered per protein produced as they will not be stuck in large 'traffic jams' along the middle of the transcript. This lower output per transcript may be compensated for by a higher number of transcripts, which the

cell can do at little cost if RNAP is not limiting. These universally conserved motifs seen in nature can be used as inspiration for the analysis of mathematical models and may assist in discovering rules which can be applied to the optimisation of gene sequences.

An alternative theory for why certain codons are more prevalent at the start of a transcript has been suggested, where the codons used mean the mRNA is less likely to form strong secondary structures. This in turn means that the RBS sequence is unblocked and free for a ribosome to bind to it, thus increasing the likelihood of translational initiation[].

There are many factors which influence the speed and efficiency of the translational process, and a number of competing theories about the reasons certain genetic motifs are observed (such as 'slow' or 'rare' codons vs anti Shine-Dalgarno sequences). The slow codon argues that there is a lower abundance of the loaded tRNAs for certain 'slow' codons, meaning there is a lower likelihood that an elongation event for these codons occurs within any time interval. Li et al. show that part of the reduced codon speeds that occur around 'slow codons' is due to the increased likelihood of anti Shine-Dalgarno sequences which causes hybridisation between the 16S ribosomal RNA of the translating ribosome and the mRNA. The energetic interaction between the 16S ribosomal RNA of the translating ribosome and the mRNA means that an elongation event becomes less energetically favourable and is therefore less likely to occur. It is likely that all of these factors play some role but their importance is a function of multiple other variables.

Mitarai et al. model the translational process and treat it as a traffic problem and perform stochastic simulations. They observe 'traffic jams' of ribosomes on the mRNA where there are codons which are translated at a slower rate. They also report that codon usage impact the total number of ribosomes on a transcript. The effect of traffic james and slow codons can be minimised using an 'on ramp' of slower codons at the start of the transcript, as well as using a weaker RBS (though there is a threshold of RBS strength below which no additional benefit is gained). A ribosomal 'cost' of protein production per transcript is also discussed, and is a concept we are particularly interested in this project.

Several models have been constructed to try and link synthetic circuits with their host cell through ribosome usage. In Scott et al.[?] they show a model where there is a detailed description of how ribosomes and protein levels are linked through growth rates. In this study they hypothesise that cellular protein is posit into different fractions where a fixed fraction of protein

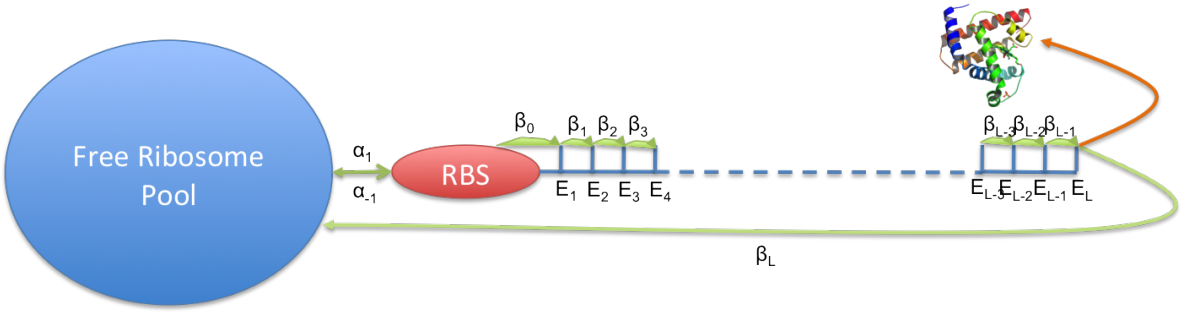


Figure 1.8: A representation of the flow of ribosomes in translation (green arrows). The blue scaffold represents the mRNA with E_is being codons along it, the green arrows represent the direction of movement of the ribosome with α_is and β_is being the rates and the orange arrow represents the production of a protein.

is allocated for core cellular maintenance functions and the remaining fraction is split between unnecessary, or heterologous, protein, ribosomal protein and 'other' protein. Capacity is a concept in this study and they attempt to draw analogies to physical laws such as with Ohm's Law. However, a key assumption in their model is that the cell is in *'balanced growth'* which means that total protein grows at the population growth rate and proportions of different types of protein (such as ribosomal protein and protein from the synthetic circuit) remain constant. This is not true in a large number of cases where the circuit has a dynamic output and therefore protein levels are not constant.

Carrera et al.[?] attempts to build a model which can predict the growth rate of cells based upon a number of parameters associated with the inserted synthetic circuit. This model was built using existing data from the literature. They then built two constructs with fluorescent proteins and inducible copy number. By predicting the amount of ribosomes being used by the circuit they were able to estimate the number of ribosomes available to the cell and from that predict the growth rate. However, in this model they assume universal elongation rates for translation for any circuit, which we have seen above is certainly not true. In addition, as we will see in this thesis, in different media a single synthetic circuit can place different levels of burden upon the cell. In order to improve this model it is necessary to look in more detail at the inserted synthetic circuit and make prediction about the elongation rates in the specific cellular context as well as understanding the link between the burden imposed and the environment in which the cell is growing.

1.2 Aims and Objectives

This project aims to gain a greater understanding of the interactions between a synthetic genetic circuit and its host chassis. We are interested in the concept of the 'burden' that synthetic circuits can place on a chassis cell through shared resources. We aim to develop a quantitative system that can interrogate the shared resource pool and inform us about the availability of shared resources within the cell. Through development of such a system we aim to be able to test how different circuits place different burdens on the cell and attempt to uncover a relationship between circuit design choices and the implications of these choices in terms of circuit output and the burden imposed. Complementing the experimental work we also aim to build a model of translation that can be used to make predictions about the implications of changing various control points in gene expression (e.g. promoter strength, RBS strength, codon usage).

This project has a number of milestones it will attempt to achieve, which are split between modelling and wet-lab. The project will develop and make use of two core experimental modules and a modelling module. The modules are as follows:

Module 1: Capacity Monitor We will build a synthetic device that is able to infer the availability of shared resources (capacity) within the cell. This circuit will run as a small 'background' process within the cell and have minimal impact on the shared resource pool. The shared resource pools will be inferred by monitoring levels of protein and OD from this circuit and then using these measurements to calculate rates of protein production. Using quantitative PCR techniques to quantify the levels we on specific mRNA per cell will also be able to infer the availability of shared resources associated with both transcription and translation. This system will also require robust testing to ensure that its behaviour is due to changes in shared resources from the production of heterologous proteins.

This system will be able to be used by synthetic biologists to investigate how any circuit (within the constraints our design will impose) impacts upon the cell in terms of shared resources. We will provide a standardised way of investigating the capacity available within a cell that can be used to characterise both cell strains and circuits. This will allow synthetic biology to move beyond the contextless characterisation of cell strains and circuits to characterising them in terms of the interface of shared resources. As mentioned above, this is a very important

consideration as circuits become larger and more complex and as greater accuracy in the predictability of behaviour is required.

Module 2: Investigating the Impact of Various Control Points We know that the expression of heterologous protein within cells can be controlled at a number of points by changing the DNA parts involved (copy number, promoter strength, RBS strength and codon usage). The current literature is full of investigations into how these control points impact upon the amount of protein produced[]. We consider an equally important consideration to be the amount of resources that a circuit uses. This module will consist of building a library of circuits that differ in these control points. This library will then be used to elucidate the impact of changing different control points and potentially allow us to uncover design principles that should be employed for optimal circuit behaviour, not just in terms of output but also in terms of resource usage.

The findings of this module have the potential to be high impact as they will have implications across all of genetic design, which in turn cascades down to applications in synthetic biology and biotechnology. Optimising resource usage by synthetic circuits will allow researchers to minimise deleterious effects on the chassis cell whilst maintaining desired circuit performance. This is of great importance as larger and more complex circuits are built as well as for industry where efficiencies and waste are of huge importance.

Module 3: Modelling the Interactions To complement the experimental data we will build a model of translation (identified within the literature to be the key bottleneck in protein production and cell-circuit interactions[]). This model will capture the impact of changing all of the control points mentioned above and will be able to simulate phenomena such as ribosomal stalling and 'traffic-jams'. Ideally this model should be able to reproduce the behaviours seen in our experimental data. Whilst all of the parameters required for such a model are not currently obtainable, we will be able to perform qualitative comparisons to understand why we see the behaviours from the experimental data.

This model will be able to be used by researchers to predict the impact of potential genetic designs on both protein production rates and resource usage. It will also be implemented in a programming language (Python) so that it can easily be used by researchers who want to simulate circuit behaviour using their own parameters as input.

Chapter 2

Materials and Methods

Move to Second Chapter

2.1 Polymerase Chain Reactions (PCR)

2.1.1 Reaction Protocol

Unless otherwise specified, all polymerase chain reactions (PCRs) were performed using the protocol detailed below.

Each reaction had a total volume of 50µl with the following amounts of reagents:

Buffer	10µl
dNTPs	1µl
Phusion enzyme	0.5µl
Primer 1 @ 10nM conc.	1µl
Primer 2 @ 10nM conc.	1µl
Template DNA @ 30ng/µl conc.	1µl
H_2O	35.5µl
TOTAL	50 µl

Table 2.1: PCR reaction recipe

The reactions were performed in an *GS482 Dual 48 Well Thermal Cycler* using the following protocol:

1. 1:20 @ 98℃

2. 1 cycles of:
    (a) 0:15 @ 98℃
    (b) 0:45 @ 67℃ dropping by 1℃ each cycle until reaching 58℃
    (c) 2:00 @ 72℃
3. 20 cycles of:
    (a) 0:15 @ 95℃
    (b) 0:45 @ 58℃
    (c) 2:00 @ 72℃
4. 5:00 @ 72℃
5. left to store @ 4℃

2.1.2 Primers

All primers were synthesised by *Integrated DNA Techologies (IDT)* and diluted to a working stock concentration of 10mM ready to be used directly in a reaction without further dilution. Both master stocks and working stocks of these primers were stored in a freezer at -20℃ and were thawed on ice when required in a reaction.

2.2 Digestions and Ligations

2.2.1 Digestions

All digestions were performed using *New England BioLabs (NEB)* enzymes according to the protocols detailed for the corresponding enzyme at the NEB website.

Unless otherwise specified, all digestion reactions had a total volume of 20μl. If the reaction was a single restriction digest, the following recipe was used:

If the reaction was a double restriction digest, the following amount of reagents were used:

TODO: Find out make and model of static incubator and heating block.

Buffer as determined by *NEB* protocol @10x conc.	2μl
BSA @10x conc. (if required, else H_2O)	2μl
Enzyme 1	0.25μl
Template DNA diluted in H_2O to appropriate conc.	15.75μl
TOTAL	20 μl

Table 2.2: Single restriction digest recipe

Buffer as determined by *NEB* protocol @10x conc.	2μl
BSA @10x conc. (if required, else H_2O)	2μl
Enzyme 1	0.25μl
Enzyme 2	0.25μl
Template DNA diluted in H_2O to appropriate conc.	15.5μl
TOTAL	20 μl

Table 2.3: Double restriction digest recipe

Digestions were done in 1.5μl volume tubes at the specified temperature. If the protocol required the reaction to take place at 37℃ then the reaction was placed in a TODO: static incubator at 37℃, otherwise it was placed in a TODO: heating block at the appropriate temperature. Digestion reactions took place for a minimum of 1 hour and maximum of 3 hours. Once complete, samples were denatured if required by the protocol and then placed on ice or put straight to ice if denaturation was not required.

2.2.2 Ligations

All ligation reactions were performed using the following protocol, unless otherwise stated.

Ligation reactions were done at a volume of 10μl with the following recipe:

NEB T4 Ligaze buffer	1μl
NEB T4 Ligaze	0.25μl
Input DNA diluted in H_2O to appropriate conc.	8.75μl
TOTAL	10 μl

Table 2.4: Ligation recipe

For ligations that contained separate backbone and insert DNA fragments, the molar concentrations were 1:3 respectively with backbone DNA concentration of 40-80ng/μl.

For ligations with a single linear DNA fragment to be circularised the DNA was at a concentration of 50-100ng/μl.

The following protocol was followed:

1. 1.5 hours @ room temperature (18-22℃)
2. 15 minutes @ 65℃
3. 5 minutes on ice

2.3 DNA Plasmid Extraction

All plasmid extractions from cells were performed using *QIAGEN QIAprep Spin Miniprep Kit*, eluted into 50μl of ddH_2O.

Unless otherwise stated, the following protocol was used:

1. Inoculate 5ml of LB media (with appropriate antibiotic) and grow-up overnight at 37℃
2. Spin 1.5ml of grow-up in TODO: bench-top centrifuge at 13,000rpm for 1 minute to form a pellet in 1.5ml tube.
3. Repeat previous step 2 more times so total amount of grow-up used is 4.5ml.
4. Remove supernatant.
5. Resuspend cells in 250μl of supplied *QIAGEN* P1 buffer.
6. Add 250μl of supplied *QIAGEN* P2 buffer and gently mix by inverting tube 5 times.
7. Add 350μl of supplied *QIAGEN* N3 buffer and gently mix by inverting tube 5 times.
8. Spin in TODO: bench-top centrifuge at 13,000rpm for 10 minutes to separate cell debris from supernatant.
9. Remove 800μl from tube and place into *QIAGEN* mini-prep column.
10. Spin in TODO: bench-top centrifuge at 13,000rpm for 1 minute, then discard flow-through.
11. Add 750μl of PE buffer to column and spin in TODO: bench-top centrifuge at 13,000rpm for 1 minute and discard flow-through.
12. Spin again at 13,000rpm for 1 minute.
13. Place column unit into new 1.5μl tube and add 50μl of ddH_2O to matrix.

14. Spin in TODO: bench-top centrifuge at 13,000rpm for 1 minute, collecting DNA suspended in ddH_2O in 1.5ml tube.
15. Measure DNA concentration on TODO: nanodrop.

2.4 DNA Purification

All purifications of DNA fragment products from PCRs, restriction enzyme digests etc were performed using the following protocol, unless otherwise stated:

1. Pour a 1% agarose gel (50ml TAE buffer containing 0.5g of agarose and 5μl of *Invitrogen SYBR® Safe DNA Gel Stain*).
2. Run gel in TODO: electrophoresis tank at 85V for 40 minutes.
3. Excise DNA band from gel using a scalpel on TODO: blue viewer.
4. Weight excised gel fragment.
5. Use *QIAGEN QIAquick Gel Extraction Kit* as follows:
   (a) Add 3x weight of gel fragment as volume of QG buffer.
   (b) Heat buffer/gel mixture for 10 minutes at 50℃, mix vortexing the tube every 2-3 min during the incubation.
   (c) Once gel fragment has completely dissolved into the buffer add 1x weight of gel as volume of isopropanol and mix by vortexing.
   (d) Place mixture into *QIAquick column*.
   (e) Spin in TODO: bench-top centrifuge at 13,000rpm for 1 minute to bind DNA to column matrix and discard flow-through.
   (f) Add 750μl of PE buffer to column.
   (g) Spin at 13,000rpm for 1 minute to wash column and discard flow-through.
   (h) Spin again at 13,000rpm for 1 minute.
   (i) Place column unit into new 1.5μl tube and add 50μl of ddH_2O to matrix.

(j) Spin in TODO: bench-top centrifuge at 13,000rpm for 1 minute, collecting DNA suspended in ddH_2O in 1.5ml tube.

(k) Measure DNA concentration on TODO: nanodrop.

2.5 Protein Electrophoresis

This assay allows us to separate proteins by molecular weight. This allows us to obtain an indication of whether certain proteins of known weights are present.

1. Day 1

   (a) Grow cells overnight for 12 hours in 5ml of media.

2. Day 2

   (a) Centrifuge each culture for 1 minute at 13,000rpm 1.5ml tube to form a pellet of cells and discard supernatant.

   (b) Repeat 2 more times so total volume of each culture used is 4.5ml.

   (c) Resuspend each pellet in 600μl of ddH_2O.

   (d) Add 200μl of 4x LDS + loading buffer solution.

   (e) Boil for 5-10 minutes at 99℃ to lyse the cells.

   (f) Centrifuge for 10 minutes at 13,000rpm.

   (g) Place 20μl of each sample into well of (TODO which protein gel?) along with a well containing (TODO: which ladder?)

   (h) Run gel at 150V for 1 hour.

   (i) Remove gel from casing and wash with ddH_2O for 5 minutes.

   (j) Repeat previous step 2 more times with fresh ddH_2O.

   (k) Submerge gel in (TODO: which dye?) and gently shake for 1 hour.

   (l) Wash with ddH_2O for 1 hour.

   (m) Replace ddH_2O and leave overnight.

3. Day 3

    (a) Take picture of gel on (TODO: gel viewer)

2.6 Capacity Monitor Assays

2.6.1 3 Hour Exponential Phase Assay

These assays tested the growth rate, circuit output and monitor output over 3 hours of growth in defined media. The protocol for this was as follows:

1. Day 1

    (a) Inoculate defined media plus appropriate antibiotics with corresponding cells.

    (b) Grow up overnight

2. Day 2

    (a) Dilute grow ups into fresh media at a dilution of 100:1

    (b) Grow up for 3 hours to move cells into exponential phase.

    (c) Dilute into fresh media in 96 well plate. Each well contains 180μl of media with appropriate antibiotics and induction chemicals as detailed below. 20μl of cells grown into exponential phase were added to each well as detailed below

        i. 6 wells of media with no antibiotic or induction chemical to which was added chassis cell without capacity monitor or synthetic circuit.

        ii. 6 wells of media with no antibiotic or induction chemical to which was addedof chassis cell with capacity monitor and without synthetic circuit.

        iii. For each circuit being tested:

            A. 6 wells of media with antibiotic and no induction chemical to which was added of chassis cell with capacity monitor and synthetic circuit.

            B. 6 wells of media with antibiotic and induction chemical to which was added of chassis cell with capacity monitor and synthetic circuit.

(d) Measure OD, green fluorescence and red fluorescence TODO: (wavelengths) using Omega Fluorometer:

i. Measurements taken every 10 minutes of OD, GFP fluorescence and RFP fluorescence.

ii. Shake plate between readings

2.7 Cell Strains

Three strains of *E. coli* were used in this project:

DH10B were obtained from the Ellis Lab Culture Collection.

MG1655 were obtained from the Ellis Lab Culture Collection.

TransforMax™EC100D™pir-116 Electrocompetent E. coli cells were purchased from Cambio and were already electrocompetent.

2.8 Transformations

2.8.1 Electrocompetency

All cells were made electrocompetent through the following protocol:

Day One

1. Inoculate a 5 ml LB culture with a single colony of the E. coli strain and incubate O/N at 37℃.
2. Incubate a conical flask containing 500 ml of LB at 37℃ O/N.
3. Store 500 ml sterile H2O in the fridge O/N

Day Two

1. Use the O/N culture to inoculate the 37℃ LB broth 1:100 and incubate shaking at 37℃.
2. Get plenty of ice and pre-chill a sterile 20% (v/v) glycerol stock and the sterile H2O. Label microtubes and store in the -80℃ freezer. Pre-chill a rotor to 4℃.

3. When the OD600 of the culture reaches 0.5, transfer to 50 ml Falcon tubes (ensure that there is no more than 40 ml/tube) and chill on ice for 30 minutes.
4. Centrifuge the tubes in the rotor pre-chilled to 4℃ at 4000 rpm for 15 minutes.
5. Discard the supernatant and, on ice, re-suspend the cells in the equivalent volume of pre-chilled water.
6. Centrifuge as before.
7. Discard the supernatant, on ice re-suspend cells in pre-chilled 20% glycerol (volume is not important but ideally just enough to re-suspend the cells e.g. 2ml/tube) and pool all of the cells into one of the 50 ml Falcon tubes.
8. Centrifuge as before.
9. Discard the supernatant and, on ice, re-suspend the cells in 3 ml pre-chilled 20% glycerol.
10. Transfer the cells into the pre-chilled microtubes in 50 μl aliquots and store immediately at -80℃.

2.8.2 Electroporation

All transformations were done by electroporation into cells prepared by the above electrocompetency protocol as follows:

1. Add 0.5μl of prepared plasmid into a cold *BioRad E.Coli Pulse Cuvette™* from freezer.
2. Add 40μl electrocompetent cells into same cuvette.
3. Pulse using *BioRad MicroPulser™* on 'Bacteria' setting.
4. Immediately add 300μl of LB media to cuvette and mix with cells.
5. Remove mix of cells and media from cuvette and place into 1.5μl tube and shake for 45mins at 37℃.
6. Spread 40μl on plate with appropriate antibiotic and leave to grow overnight at 37℃.
7. Pick colonies the next day and grow overnight at 37℃.

2.9 Antibiotics

All antibiotics used in this project were used at standard concentrations for *E. coli*:

Chloramphenicol was used at a concentration of 34 μg/ml in all media and agar plates.

Kanamycin was used at a concentration of 100 μg/ml in all media and agar plates.

Ampicillin was used at a concentration of 100 μg/ml in all media and agar plates.

2.10 Growth Media

Two types of growth media were used in this project:

LB Media

M9 Media

2.11 CRIM Genomic Insertion

All genomic insertions were done using the CRIM system[?] into the *E.Coli* λ site using CRIM plasmid pAH63 and helper plasmid pINT-ts.

2.12 Oligonucleotides

2.13 DNA Synthesis

2.13.1 sfGFP

2.13.2 Slow Codons gBlock

2.14 Data Analysis

Data was analysed using *Microsoft Excel* and variables were calculated as follows:

1. OD: obtained by taking raw OD600 data for relevant cells and subtracting the average raw OD600 readings for wells containing only the media the cells were being grown in.
2. GFP: obtained by taking raw GFP data for relevant cells and subtracting the average raw GFP readings for wells containing cells without any fluorescent protein production at equivalent OD.
3. GFP/OD: obtained by dividing the GFP value by the OD value for a well.
4. mCherry: obtained by taking raw mCherry data for relevant cells and subtracting the average raw mCherry readings for wells containing cells without any fluorescent protein production at equivalent OD.
5. mCherry/OD: obtained by dividing the mCherry value by the OD value for a well.
6. Growth Rate: obtained as an estimate of average growth rate over time interval of length τ hours using the following equation:

$$\gamma = \frac{1}{\tau} \log_e \left(\frac{OD_{t+\tau}}{OD_t} \right)$$

   where OD_t is the OD value at time t.
7. GFP Production Rate: obtained as an estimate of average GFP production rate over the time interval of length τ hours using the following equation:

$$\text{GFP production rate} = \frac{\gamma(GFP_{t+\tau} - e^{-\gamma\tau} GFP_t)}{1 - e^{-\gamma\tau}}$$

   where γ is the growth rate (obtained as shown above) and GFP_t is the GFP value at time t.
8. mCherry Production Rate: obtained as an estimate of average mCherry production rate over the time interval of length τ hours using the following equation:

$$\text{mCherry production rate} = \frac{\gamma(mCherry_{t+\tau} - e^{-\gamma\tau} mCherry_t)}{1 - e^{-\gamma\tau}}$$

   where γ is the growth rate (obtained as shown above) and $mCherry_t$ is the mCherry value at time t.

2.15 Construction

Section 3.4 outlines the design of the monitor variants. Constructing these variants required a multi-stage approach. Initially all versions of the monitor were created in a pSB1AK3 backbone with each of the degradation tags. Once these variants were created, they were moved into different plasmid backbones for insertion into the cell (either directly on plasmid-based systems or into a CRIM plasmid ready for genomic integration.

2.15.1 Monitor device construction

Initial Device Construction

We designed an initial construct which we would then be able to create variants of (different degradation tags) and place into different insertion systems (plasmid systems or genomic insertion). In order to obtain a version of sfGFP that had been codon optimised for *E. coli* we used the DNA2.0 codon optimisation software and subsequently ordered synthesised DNA from them. The physical DNA we received consisted of promoter J23100 driving the transcription of codon optimised sfGFP with LVA degradation tag, driven by a strong RBS (as predicted by the Salis RBS calculator (CITE)) but without terminator (see Figure 2.1).

This was delivered in a DNA2.0 proprietary backbone. By ordering this synthesised DNA from DNA2.0 we were able to obtain a complete version of one of our test circuits (excluding terminator). In order to obtain the rest of the variants we first moved the insert into pSB1AK3 containing terminator B1002 to prepare it for being placed into other plasmids for either being inserted into the cell in the plasmid-based system or as a genomic insertion using the following steps:

1. Digested DNA2.0 plasmid containing monitor with EcoRI and SpeI.
2. Gel purification of insert fragment (849 bp fragment).
3. Digest pSB1AK3 containing terminator B1002 with EcoRI and XbaI.
4. Gel purification of pSB1AK3 containing B1002.
5. Ligate monitor with pSB1AK3 containing B1002.

6. Transform into DH10B onto kanamycin agar plates.

7. Grow up overnight in 5ml LB media (select colonies by green fluorescence).

8. Miniprep and sequence insert.

Once this 'master construct' was in pSB1AK3, PCR was used with 1) BioBrick VF primer and 2) the corresponding primer from Table 2.5) to change the degradation tag (or remove any degradation tag) and add an SpeI restriction site for simple insertion back into pSB1AK3.

Oligo	Description	Sequence
RS002	BioBrick VF primer	TGCCACCTGACGTCTAAGAA
RS003	No Tag	ATATACTAGTATCATTACTTATACAGCTCGTCCATACCG
RS004	AAV Tag	ATATACTAGTATCATTAAACCGCCGCCGCGTAATTCTCATCATTTGCAGC
RS005	DAS Tag	ATATACTAGTATCATTAGCTCGCGTCCGCGTAATTCTCATCATTTGCAGC
RS006	LAA Tag	ATATACTAGTATCATTACGCCGCCAGCGCGTAATTCTCATCATTTGCAGC

Table 2.5: Primers for changing monitor degradation tags

We used PCR to amplify the fragments in standard conditions and followed the following steps to obtain each variant of the monitor, in terms of degradation tag, in pSB1AK3:

1. PCR pSB1AK3 containing monitor with RS002 and one of RS003 (No tag), RS004 (AAV Tag), RS005 (DAS Tag) or RS006 (LAA Tag).

2. PCR purification of PCR products.

3. Digestion of purified PCR product with EcoRI and SpeI.

4. PCR purification of digestion product.

5. Digest pSB1AK3 containing terminator B1002 with EcoRI and XbaI.

6. Gel purification of pSB1AK3 containing B1002.

7. Ligate monitor variants with pSB1AK3 containing B1002.

8. Transform into DH10B onto kanamycin agar plates.

9. Grow up overnight in 5ml LB media (select colonies by green fluorescence).

10. Miniprep and confirm insert sequence by sequencing.

The next step was to move the monitors into different different plasmids ready to be inserted into the cell. As mentioned above three different plasmid-based systems and one genomic integration were being tested. For the plasmid-based systems all that was required was to

perform a digestion and ligation followed by a transformation. For the genomic integration it was necessary to introduce the monitor into a CRIM plasmid CITE which would subsequently be integrated into the genome of the cell at the λ-site with the aid of a 'helper plasmid'. Therefore for each insertion type it was necessary to insert the monitor into a plasmid backbone. This was done as follows:

1. Digest monitor variants in pSB1AK3 with EcoRI and PstI.
2. Gel purification of insert fragment (849 bp fragment).
3. Digest plasmids in Table 2.6 with EcoRI and PstI.
4. Gel purification of plasmids.
5. Ligate each monitor variant (5x) with each plasmid backbones (4x).
6. Transform into DH10B onto kanamycin agar plates (20 transformations).
7. Grow up overnight in 5ml LB media (select colonies by green fluorescence).
8. Miniprep and sequence inserts.

Backbone	Description
pSB1K3	High copy number BioBrick plasmid with kanamycin resistance marker
pSB3K3	Medium copy number BioBrick plasmid with kanamycin resistance marker
pSB4K5	Low copy number BioBrick plasmid with kanamycin resistance marker
pAH63	CRIM plasmid for integrating into λ-site in *E. coli* genome with kanamycin resistance marker

Table 2.6: Backbone plasmids for inserting monitor into

After the above steps had been completed we had all the plasmids prepared necessary for inserting all of our monitors variants into the host cell.

Inserting into host cell

Plasmid-based insertion

Inserting the plasmid-based systems (monitor in pSB1K3, pSB3K3 or pSB4K5) into cells and verifying that we have the correct constructs is a relatively simple process and proceeds as follows:

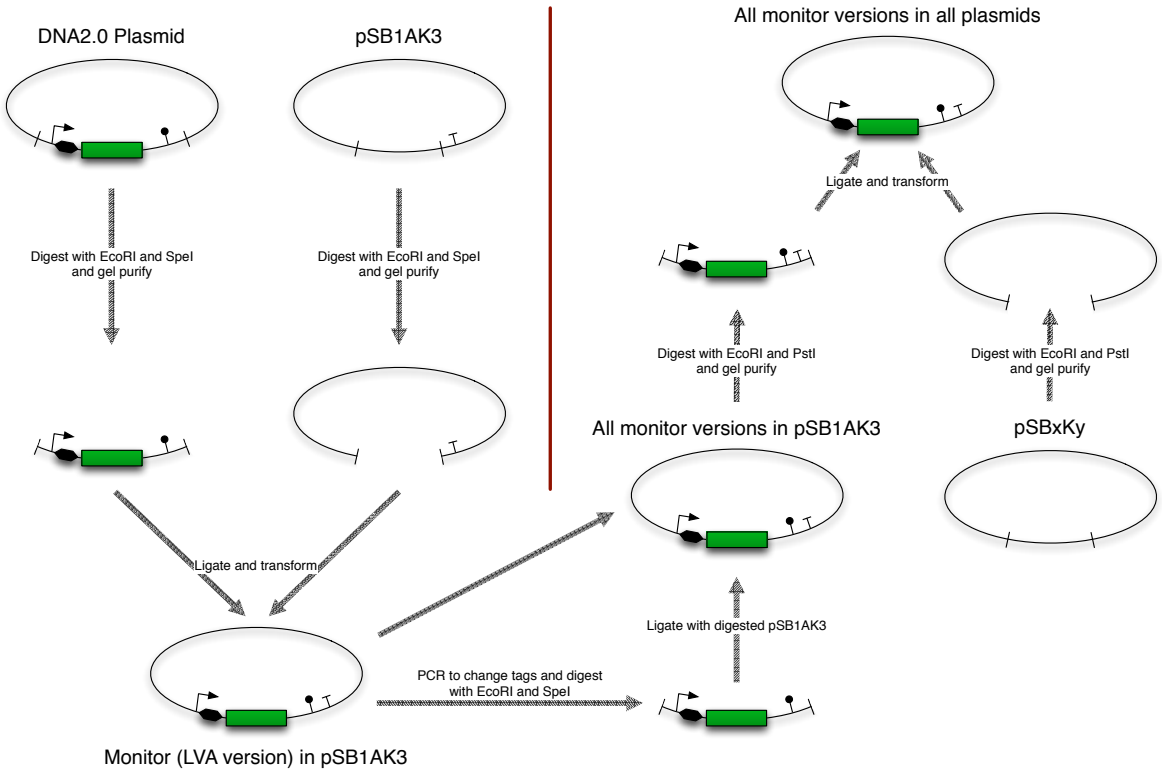


Figure 2.1: The final monitor design consists of promoter J23100 driving the transcription of a strong RBS (as predicted by the Salis RBS calculator CITE) and coding region for a fully codon optimised sfGFP without any degradation tag, followed by a fully synthetic B1002 terminator (not to scale).

1. Transform ligation product into DH10B using protocol outlined in section **??** (Transformations in Materials and Methods) and plate each transformation onto a separate kanamycin resistant plate.
2. Grow up 2 colonies from each plate overnight.
3. Miniprep overnight growth.
4. Digest with ??? and ??? to confirm correct backbone.

After the above steps were completed we have all versions of our monitor in all plasmid-based systems inserted into our host cell (DH10B).

Genomic insertion

To insert the monitor into the genome we used the CRIM system (CITE). This system involves two separate plasmids, one of which contains the monitor and will be inserted into the genome, the other being a 'helper plasmid' that facilitates the genomic integration.

The CRIM system works by placing the circuit you wish to insert into the genome into the CRIM plasmid corresponding to the integration site. CRIM plasmids have the γ replication origin of R6K, which requires the trans-acting Π protein (encoded by pir) for replication. This means that these plasmids can only be maintained in cells which have a pir^+ genotype. In order that we can maintain our CRIM plasmid with monitor inserted we needed to transform into pir^+ cells. For this we used TransforMax™EC100D™pir-116 Electrocompetent E. coli cells.

1. Transform pAH63 plasmids containing monitors into TransforMax™EC100D™pir-116 Electrocompetent E. coli cells.
2. Prep and test digest.
3. Transform pINT-ts helper plasmids into DH10B at 30℃.
4. Make DH10B containing pINT-ts electrocompetent using protocol outlined in Section **??**.
5. Integrate CRIM plasmids containing monitors into genome using protocol outlined in Section 2.11.

Chapter 3

Results: Constructing a Capacity Monitor

3.1 Motivation

As mentioned above, a key aim of the project was to gain an insight into the way in which a synthetic circuit interacts with its host cell through shared resources. Using currently available techniques, it is not currently possible to get a direct quantification of the amounts of these resources being used in native cellular processes, processes associated with the synthetic circuit and those being used by neither (i.e. *free resources*). Therefore, we require a proxy for these values.

3.1.1 Impact on Cell

There are a number of ways to infer the impact of a synthetic circuit on its host cell. 1) Growth rate is a commonly used indicator of the state of a cell whereby a decreased growth rate indicates a larger 'burden' being placed on the cell CITE. 2) We can investigate how the transcriptome, proteome or metabolome change in cells that contain synthetic DNA. The complex nature of biological systems means that the reactions of a cell to the presence of synthetic DNA will be multi-faceted and all of these techniques are useful and informative. However, none of them give a decent insight into the impact on the core means by which the cell and circuit interact - i.e. through shared resources. Changes in the availability of these resources have global

effects on the cell in terms of growth rate, shifts in the transcriptome, proteome, etc. We can envisage these impacts being a *network* of global physiological effects on the cell that are all interlinked with the availability of shared resources being the central node that interacts with the synthetic circuit (see Figure 3.1).

By creating a means of investigating the availability of these resources we will be able to gain a greater insight into this core interaction.

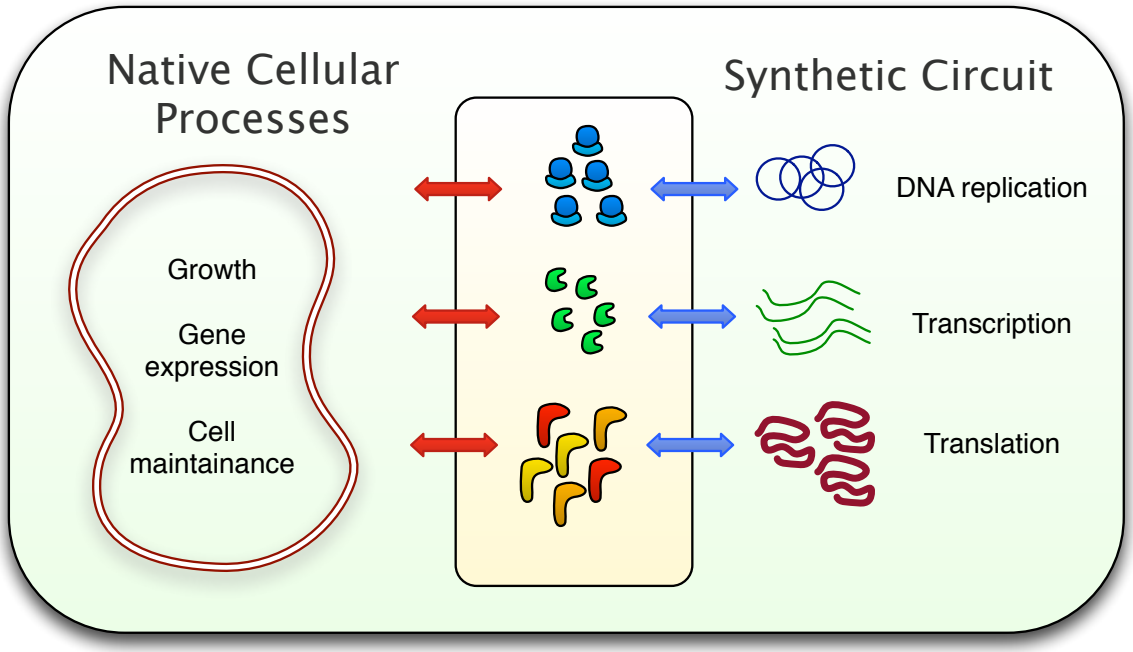


Figure 3.1: A synthetic gene circuit and its host cell interact in a number of ways. The key interface being investigated in this project is the *shared resource pool*. This is the set of common resources that are used by both the cell and circuit for cell maintenance and growth as well as for gene expression from the circuit. Key examples of *shared resources* are ribosomes, polymerases etc.

3.1.2 Impact on Additional Heterologous Gene Expression

Expression of a heterologous protein within a cell affects both the cell itself as well as the expression of other heterologous proteins within the cell. For example, we consider two expression units A and B that when placed alone into a chassis cell, these express protein at rates x and y respectively. When these units are placed simultaneously into the same chassis cell they will be competing for the same resources and expression machinery. Therefore the level of expression of units A and B would be reduced to $< x$ and $< y$ respectively (see Figure

3.2).

It is therefore important to understand how the output of two genes changes when they are co-expressed as compared to in isolation. A greater understanding of how these interactions occur may lead to improvements in the design of multi-gene synthetic systems and will be an increasingly important consideration as engineered biological systems become increasingly complex.

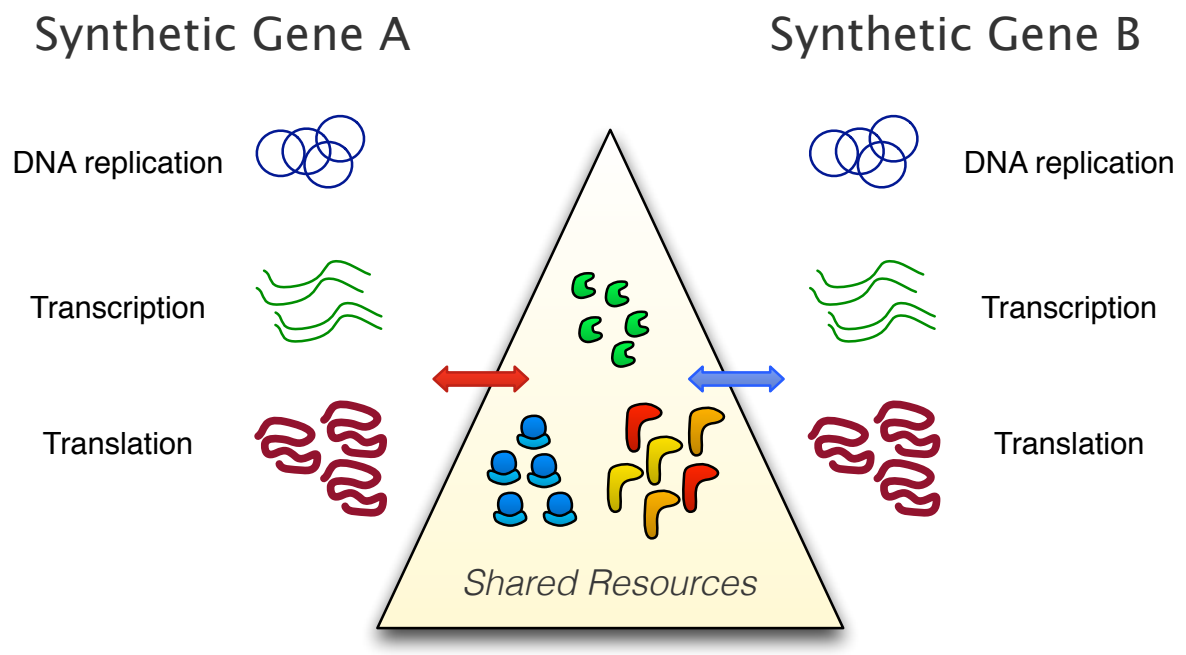


Figure 3.2: Expression of a synthetic gene circuit will also have interaction with other synthetic genes within the cell through the same shared resource pool. The two genes may be on different plasmids in the same cell, or may simply be different genes in the same circuit. Expression of gene A will be competing for the same shared resources as gene B.

3.1.3 Capacity Monitor

In order to gain an insight into the interactions detailed above it would make sense to investigate the interfaces between cell and circuit. As mentioned, the core interface is through the shared resource pool. We wanted to know whether we could measure changes in this shared resource pool by placing a small synthetic circuit into the cell that would allow us to quantify the amount of 'capacity' available in this pool. While we did not envisage this monitor being able to give direct quantifications of the number of individual resources, we wanted to get an insight into the availability of resources required for gene expression. We can look at gene expression

as a single step (and therefore as a function of all machinery and building blocks), and once we have confirmation that we are able to observe changes in this machinery we will aim to separate the processes that constitute gene expression - i.e. transcription and translation (see Figure 3.3). While it may be desirable to gain a greater insight into the availability of individual resources that affect the transcriptional and translational rates we believe quantifying the rates of these steps is sufficient for the scope of this project. Obtaining quantification of resources at a resolution where each individual resource is quantified separately would require a significant amount of work beyond what is achievable within the timeframe of this project, and we believe the insight into shared resources just from looking at transcriptional and translational resources will be enough to gain a better understanding of cell-circuit interactions. This will allow us to understand which processes are most important and, if necessary, try to identify individual resource availabilities in further work.

This capacity monitor can be considered analogous to the system monitor found on many computers, which informs the user of the availability of certain resources available to programs (such as RAM or CPU).

3.2 Requirements

Once the core functionality of the capacity monitor had been defined, it was necessary to create some specifications for its design. By creating a list of specifications it was possible to clearly delineate the features we required from a *capacity monitor*. These were as follows:

3.2.1 Allow quantification of capacity in *E. coli* cells

This is clearly a key feature of the capacity monitor. By being able to quantify the resource capacity within a cell it is possible make comparisons between different circuits or circuit variants. It was also necessary to ensure that the capacity monitor did not interact with the cell via any mechanisms other than shared resources such as toxicity or cross-talk.

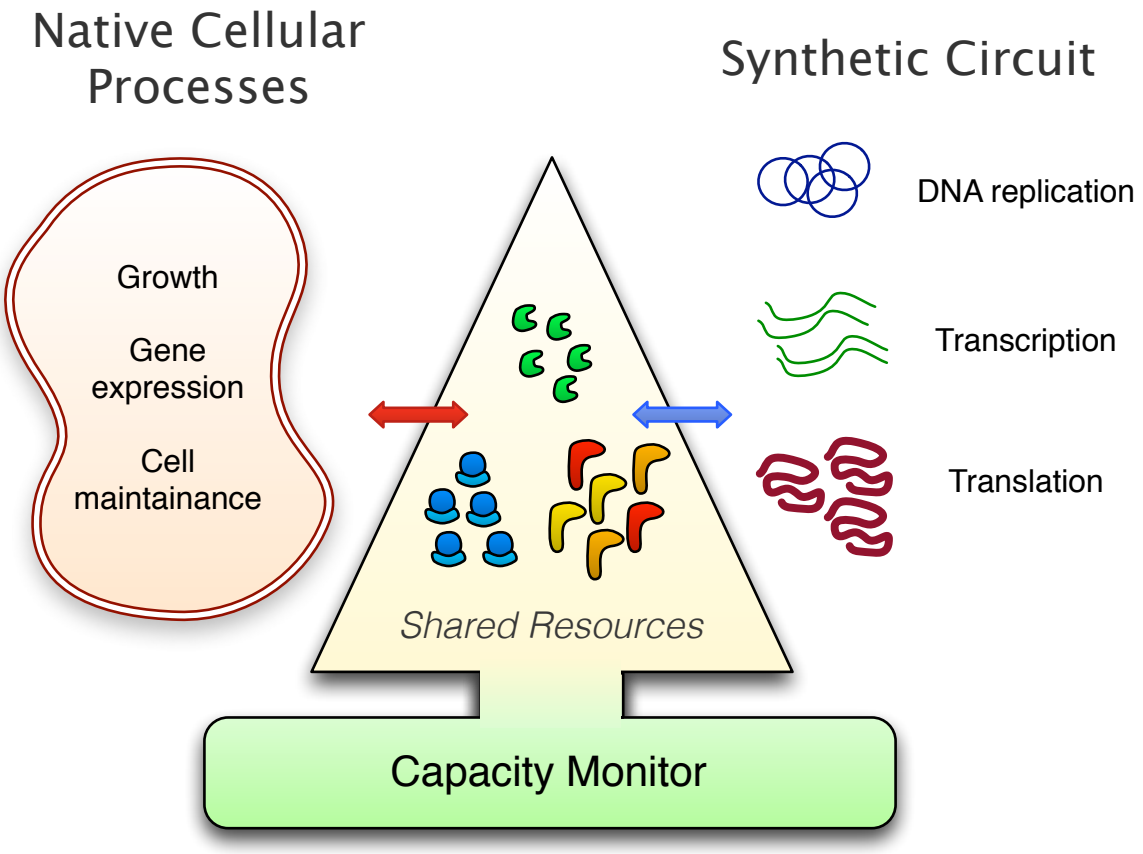


Figure 3.3: Capacity monitor acts as a proxy for the availability of shared resources. An increase (or decrease) in the output of the monitor translates as an increase (or decrease) in availability of shared resources.

3.2.2 Interact with Shared Resources

It was necessary to ensure that any non-resource-derived interactions between the cell and the circuit were avoided or minimised to a level where their effect was negligible. An important consideration was the choice of proteins used so that they were not toxic when expressed in *E. coli*. In order to do this, it was useful to choose proteins that are widely used and well understood in the literature not to cause a toxic affect on a host cell.

In addition, we wanted to minimise any cross-talk where by the regulatory mechanisms of the cell and circuit overlap. This was done, where possible, by using synthetic elements that do not occur naturally in *E. coli* or, ideally, in any natural organism. This was achieved through different approaches:

1. Designing sequences *de novo* such that they are entirely synthetic and not derived from any naturally occurring sequence.

2. Creating mutations of sequences that exist in nature such that they maintain their functionality while having different sequences from the natural origin. This can be done either through random mutant library creation or by using techniques such as codon optimisation to rationally redesign sequences.

3.2.3 Easily quantifiable output

As mentioned above, the monitor requires a quantifiable output. This output should be quantifiable at a population and single-cell level and be measurable at regular intervals over a period of time. Using a fluorescent protein would enable the easy and rapid quantification of protein levels within cells[?].

Use of fluorescent proteins as reporter proteins is a very well established technique and many variants of these proteins have been made with a range of excitation and emission wavelengths as well as folding speeds and stabilities[?]. Therefore we were able to select our reporter protein from a large set of potentials with a range of characteristics to ensure the specifications for our capacity monitor are able to be met.

Many synthetic biology labs have the ability to measure green fluorescence and green fluorescent protein (GFP) is commonly used as a reporter protein in synthetic systems CITE. While using GFP as the reporter protein in our capacity monitor poses some issues in terms of compatibility (see below), is it highly beneficial in a number of ways.

GFP is highly characterised and there exist a number of variants with desirable characteristics. Super folder GFP (sfGFP) is a rapidly folding protein[?] and therefore allows us to minimise the delay between the protein being translated and becoming measurable via fluorescence readings. In addition it is highly stable with a degradation rate in excess of 24 hours CITE, which means that in combination with the 5 degradation tags we have a wide range of degradation rates ($<$ 30 mins to $>$ 24 hours). These are key benefits that we consider outweigh the negatives associated with using GFP as the output of our monitor.

3.2.4 Maximal interoperability with other synthetic circuits

By maximising the number of synthetic circuits we are able to use in conjunction with our capacity monitor we will clearly increase its effectiveness and utility.

Synthetic circuits are introduced into cells either on a plasmid-based system or are integrated into the genome CITE. Plasmid-based systems can have compatibility issues whereby different plasmids cannot co-exist in the same cell if they have the same mechanism for replication CITE. Therefore, if the capacity monitor was plasmid-based, we would experience issues when trying to monitor the burden imposed by synthetic circuits that use the same replication mechanism as that of the capacity monitor. If we were able to place the capacity monitor on the genome we would remove compatibility issues with both plasmid- and genome-based synthetic circuits.

Compatibility between selection markers also presents an issue when trying to maximise interoperability. If the capacity monitor and synthetic circuit use the same selection system then it will not be possible for them to reliably co-exist in the same cell. Since maintaining a plasmid within a cell requires a selection marker and most systems for genomic integration also require a selection marker, avoiding the use of a selection marker in our capacity monitor was not necessarily feasible.

In order to measure the output of our monitor we need a quantifiable reporter protein. This reporter would not be able to be used in any synthetic circuit we are testing as we would not be able to differentiate between protein produced by the capacity monitor and that produced by the synthetic circuit . In selecting our reporter we had a number of considerations such as how well characterised is the protein and how easily we could quantify it on a population and single-cell level.

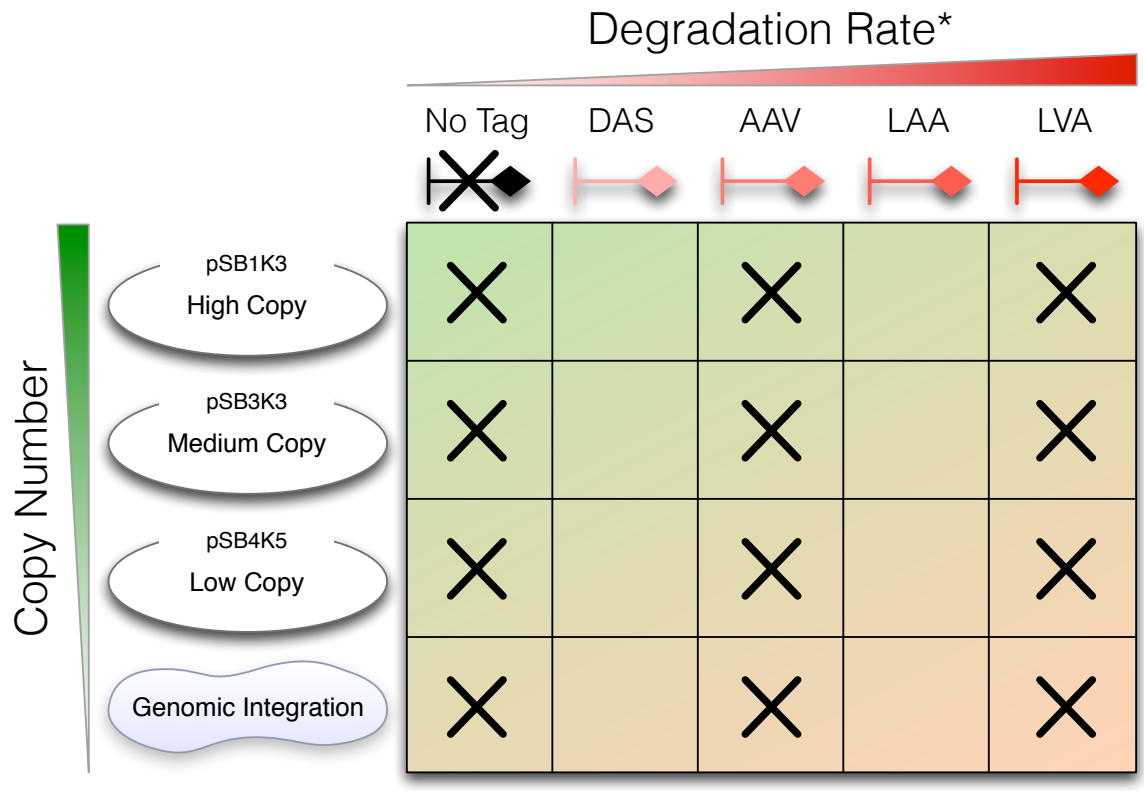
3.2.5 Minimal burden on cell

We wanted our capacity monitor to place a minimal burden on the cell by using minimal resources. Much of this thesis concentrates on understanding the ways in which different control points in a circuit affect the burden it places on a cell and therefore it is unlikely that the system we design to be a capacity monitor will place the minimal possible burden on the cell. However, there are some strategies we can use to decrease it such as minimising the amount of protein

that needs to be produced. In order to minimise the amount of protein produced we can attempt to maintain the capacity monitor at as low a copy number as possible. It will be important to balance this requirement with the need to produce enough protein to be measurable for most levels of burden placed on the cell. We also used *DNA2.0* codon-optimisation algorithms to optimise the DNA sequence for efficient expression in *E. coli* and remove any slow codons that may cause ribosomal traffic jams and unnecessary burden.

3.3 Implementation

After taking into consideration the above specifications for our monitor we decided upon the following implementation of our monitor. We would have a constitutively expressed GFP variant driven by a relatively strong RBS. In order to optimise our monitor we required a number of variants that had both different copy numbers and degradation rates. To do this we placed the circuit on three different plasmid backbones as well as a genomic integration, and used a range of 4 degradation tags as well as a version with no tag. See Figure 3.4 for an outline of the combinations. The degradation tags used are discussed in further detail in Section 3.4.6 and the integration systems are discussed in Section 3.4.1.



*predicted

Figure 3.4: This array shows all of the constructs we built, where each box is a separate construct. An X in a box denotes that we tested this combination. Overall we constructed 20 variants of potential capacity monitor systems and tested 12.

3.4 Device Design

3.4.1 Copy Number

From the requirements outlined above, it was noted that a genomic insertion would provide the highest level of compatibility with other synthetic gene circuits as there would have been no issues with cross-compatibility of plasmid origins. In addition, a genomic insertion was predicted to produce the lowest amount of protein and use up the least of the cell's resources thereby impacting on the cell the least amount.

It was decided that implementations of both plasmid-based systems and genome insertion-based systems should be tested in order to compare how they behaved with regards to a

number of factors including:

1. Signal strength - Are the GFP fluorescence levels sufficiently high to detect?
2. Signal to noise ration - How do different copy number implementations affect not only the signal but the amount of variance/noise in the signal outputs.
3. Growth Rate - Does the implementation of the monitor system impact the cell in terms of its ability to grow?

In order to vary the copy number of our monitor we placed the circuit onto 3 different plasmid backbones (see Table 3.1 for more details) and used the CRIM system for genomic integration[**?**]. The CRIM system allows a circuit to be placed into a number of different phage sites in the *E. coli* genome. Attempts were made to insert monitor systems into both λ and ϕ80 sites (see Figure 3.6 for a guide to the locations of the different sites in the genome) in order to maximise chance of successful integration. Monitor variants were successfully integraded into the λ-site before the ϕ80-site and therefore the λ was subsequently used throughout the project.

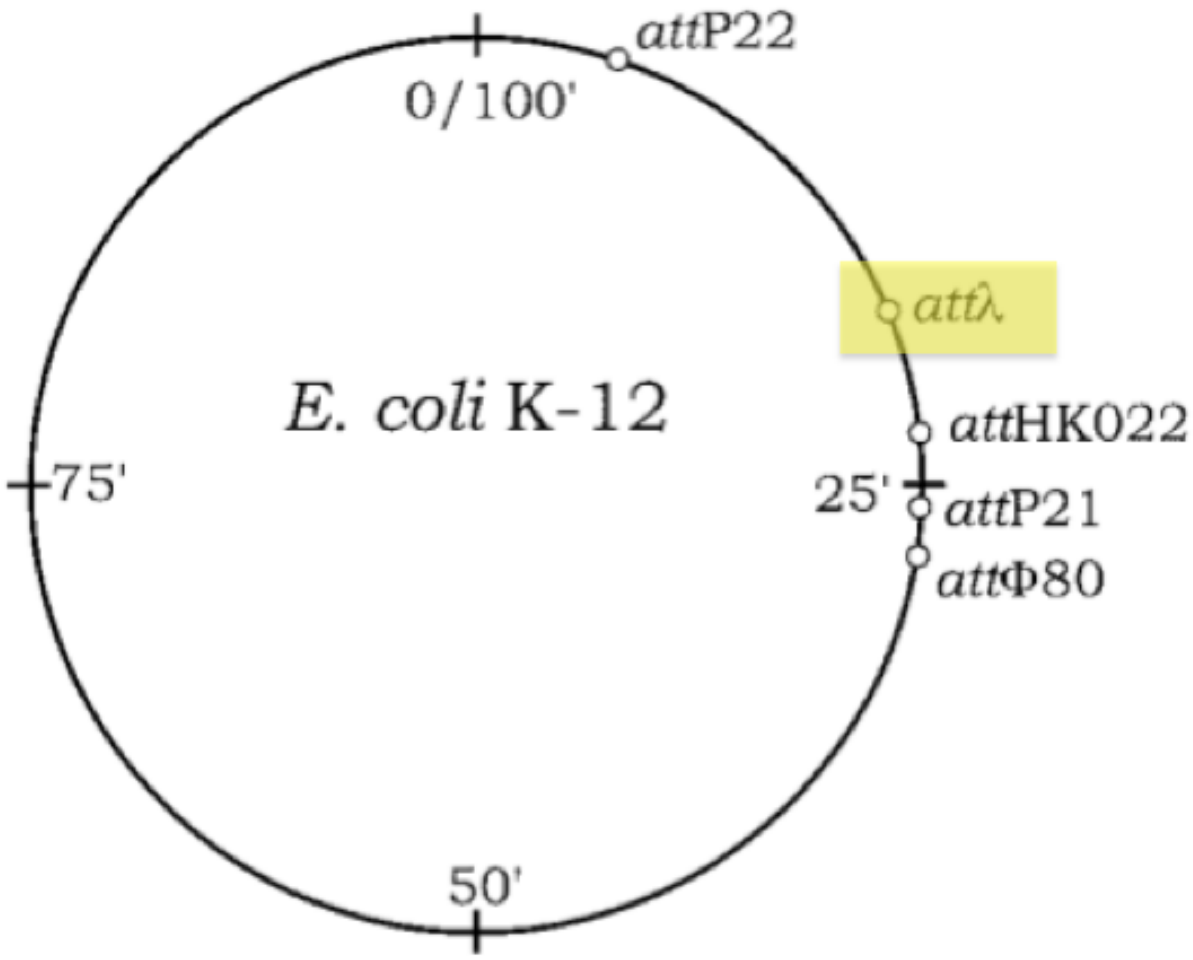


Figure 3.5: Locations on *E. coli* genome where CRIM system inserts constructs into. λ-site used in this project is highlighted in yellow. Figure adapted from **??**.

The CRIM system integrates a kanamycin resistance marker into the genome along with the desired circuit, and in order to ensure consistency we used 3 plasmid backbones with kanamycin resistance for testing the circuit on plasmids. The plasmids we used are widely-used backbones from the Registry of Standard Biological Parts (see Table 3.1.

Plasmid Backbone	Origin	Estimated Copy Number
pSB1K3	pMB1	100-300[?]
pSB3K3	p15A	10-12[?]
pSB4K5	pSC101	5[?]

Table 3.1: Plasmid backbones used for capacity monitor candidates. These are all from the Registry of Standard Biological Parts.

3.4.2 Promoter - J23100

It was decided that a constitutive promoter would be the optimal choice for this project. By avoiding the requirement for regulation at the promoter, the rate of transcription is be decoupled from the activity of any specific regulatory proteins. Instead the rate is a function of the availability of the global resources associated with transcription, such as polymerases and sigma factors.

The promoter selected was J23100, and was chosen from a library of combinatorial variants of a promoter sequence that is a fully synthetic design based on consensus sequences for -35 and -10 promoter regions and has no homologs in bacteria. This library was created by the Berkeley 2006 iGEM team CITE and consists of a single consensus sequence (J23119) as well as 19 additional synthetic sequences. From this library J23100 is reported to have the highest level of expression, apart from the consensus version CITE. Since it is a synthetic sequence, it would be expected that cross-talk interactions with the host *E. coli* cells are highly unlikely.

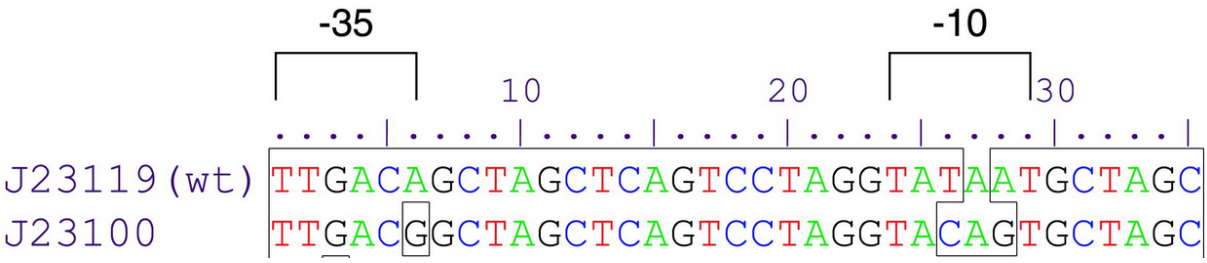


Figure 3.6: Sequence alignment of J23100 against the wild-type J23119 sequence. Figure adapted from Registry of Standard Biological Parts page for J23100. The two main resources this promoter interacts with are the σ^{70} (sigma factor) which binds to the -35 sequence, and RNA polymerase which binds to the -10 sequence.

3.4.3 RBS - Synthetic Design

The RBS sequence for the monitor was designed using the Salis RBS Calculator[?]. The forward design program was used with no upstream sequence (since the J23100 promoter ends at its -1 base) and with a downstream sequence consisting of the initial 40 bases of our codon optimised sfGFP sequence (see below). The calculator was set to 'maximise' the RBS strength and gave a predicted strength of 244950.7 au. Since this sequence was generated by the RBS calculator, it is a fully synthetic sequence and is not known to occur naturally.

TACTAGAGAAATCAAATTAAGGAGGTAAGATA

3.4.4 CDS - optimised sfGFP

We decided to use GFP and specifically the *Superfolder GFP* variant (*sfGFP*) for the reasons given before. *Superfolder GFP* is a protein that was designed by *Pdelacq et al.* in 2006 (CITE PMID 16369541). This protein is not known to act as an enzyme for any reactions within *E. coli* and its functionality is not known to be affected by any native cellular processes. It was codon optimised by DNA2.0 for efficient *E. coli* expression.

3.4.5 Terminator - B1002

This terminator is artificially designed and does not exist natively in the genome of any organism that we are aware of. It has been characterised as having a particularly high termination efficiency (90%[?]) which ensures that there will be minimal 'run-through' of polymerases as the gene is transcribed.

3.4.6 Degradation Tags - SsrA

We used a set of degradation tags from the SsrA family. These tags are derived from the natural LAA variant which is a sequence of 11 amino acids that are naturally added to the C-terminal of an amino acids chain when translation is interrupted. This sequence of 11 amino acids targets proteins for degradation by ClpXP and ClpAP (http://www.ncbi.nlm.nih.gov/pubmed/9573050). The variants of this tag differ in their final 3 amino acids where the name of the tag is the abbreviated name of the final 3 amino acids. These degradation tags are not fully orthogonal from the host *E. coli* cell since they use some of the same machinery as the cell does naturally to degrade proteins. In CITE it was seen that competition for this degradation machinery can have a number of adverse affects on the cell such as growth retardation. Tags were considered as part of the design because they may have assisted in getting dynamic readings of protein expression.

3.5 Testing

Once all of the variants of our monitor had been constructed and inserted into the host DH10B cells either via plasmid based systems or CRIM integration, we had to test them to see which best fitted the requirements mentioned above. The key factors we were interested in were:

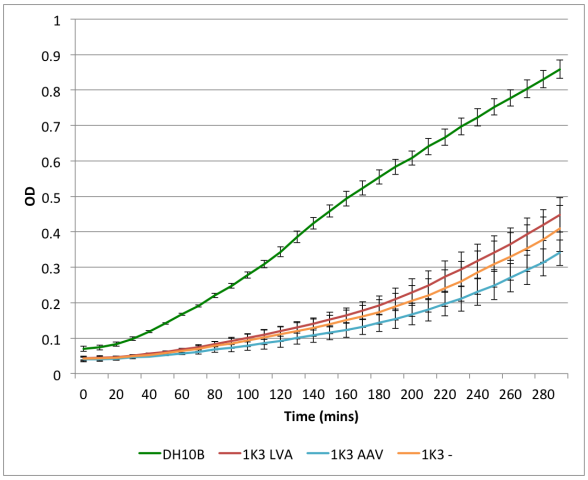
- Whether we were able to detect GFP at sufficiently high levels and whether we would be able to reliably and accurately measure changes in fluorescence levels.
- What the impact on growth rate was of all the versions of the monitor. We wanted to make sure we were impacting the cell as little as possible and growth rate is a reasonable indicator of this.
- What the strengths of the degradation rates were for each of the tags and whether they would be constant enough to allow us to make accurate calculations about the production rates of GFP given only fluorescence measurements.

We designed a simple protocol to answer these important questions. The constructs we decided to test were the variants with either a) no degradation tag, b) AAV degradation tag or c) LVA degradation tag across all expression systems (see Figure 3.4).

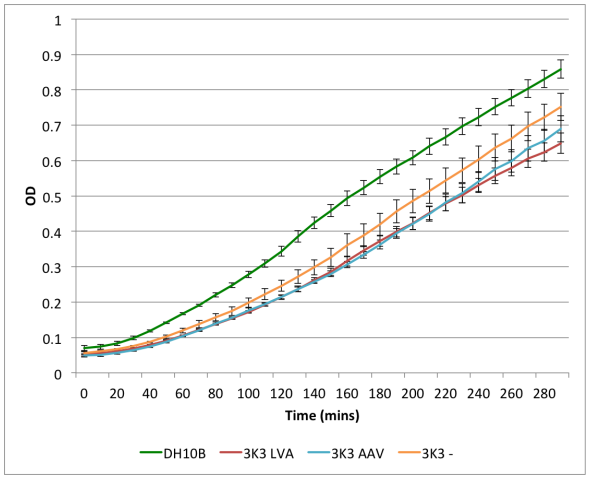
3.6 Results

3.6.1 Growth Rates

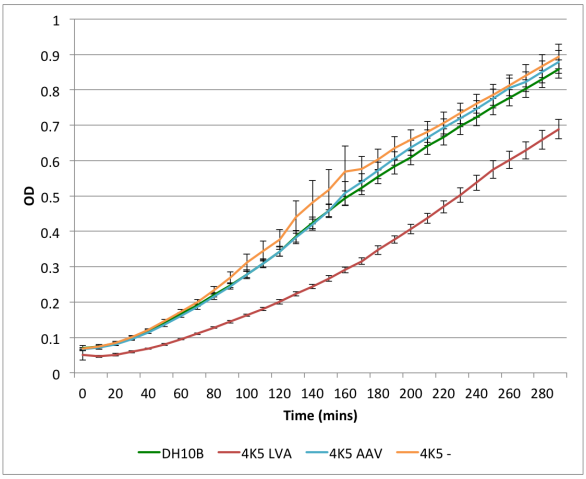
Figure 3.7 shows how the cells containing different variants of the monitor grew in LB media over a 5 hour timeframe. It is clear that generally as copy number is increased (from genomic, to pSB4K5, to pSB3K3, to pSB1K3) the cells are growing at a decreasing optical density (OD). In Figure 3.7c it can be seen that, with the exception of the LVA tagged variant, the growth of cells containing low-copy number versions of the monitor have very similar growth profiles to DH10B cells without any monitor device. Medium-copy plasmid based systems (3K3), and 4K5 with an LVA tag, grow at slightly lower OD compared to DH10B cells. High-copy plasmid systems (1K3) grow at significantly lower OD compared to DH10B. This indicates that both medium- and high-copy variants of the monitor are having significant retarding affects on the cell behaviour, decreasing growth rate.



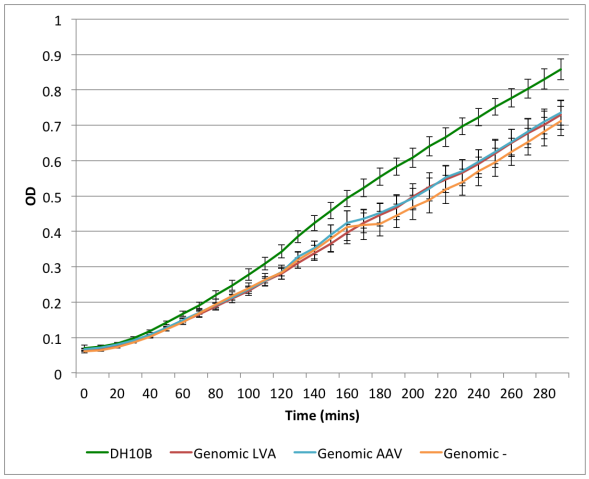
(a) Growth Curves for High-copy candidates (pSB1K3)



(b) Growth Curves for Medium-copy candidates (pSB3K3)



(c) Growth Curves for Low-copy candidates (pSB4K5)

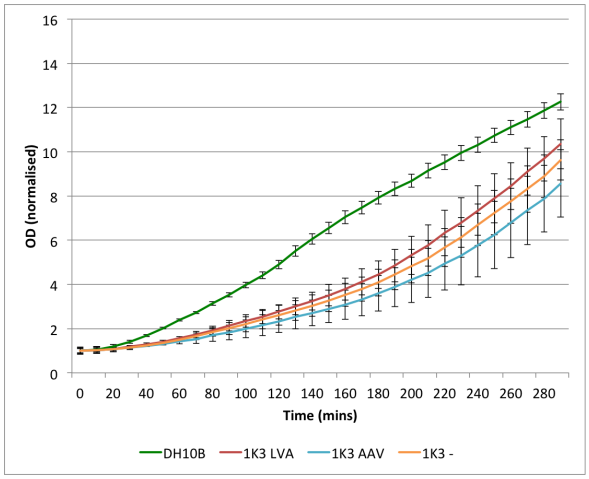


(d) Growth Curves for genomic candidates

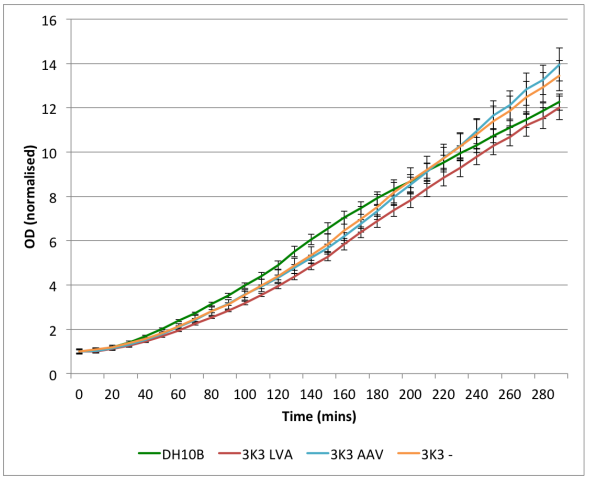
Figure 3.7: Growth curves for all monitor candidates grown for 5 hours in M9 media supplemented with 0.4% Fructose with readings taken at 10 minute intervals. Measurements are done on plate reader and averages over 6 repeats with error bars showing standard deviation. DH10B cells are included in all graphs to indicate how growth curves compare to cells without any monitor circuit. Normalisation is preformed by dividing each OD reading by the initial reading for the corresponding curve at time = 0.

In addition to looking at basic OD levels, the normalised ODs were also calculated. These values are obtained by normalising the OD values to the initial value at $t = 0$. Figure 3.8 shows the curves for normalised OD. These show that the decrease in OD for pSB3K3 candidates (medium-copy) and the pSB4K5 (low-copy) with LVA candidate are mainly due to the lower initial ODs for these candidate strains. In addition, the amount of decreased OD relative to DH10B for pSB1K3 candidates (high-copy) is lower when ODs are normalised, again because the cultures were initialised at slightly lower ODs.

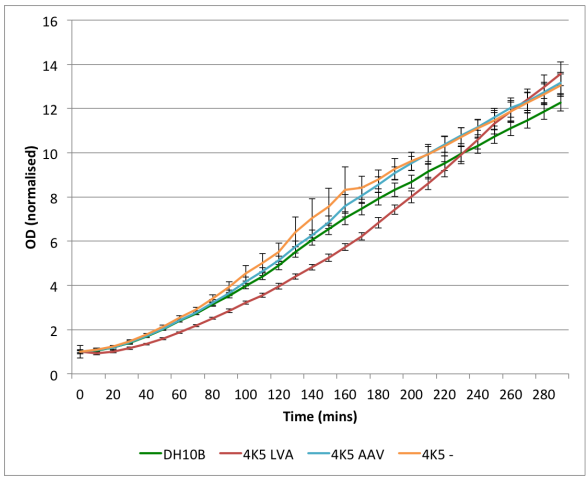
Growth rates for these different candidate strains were calculated over and can be seen in Figure **??**. This shows that high-copy candidates have a decreased growth rate of 5%-8%, medium-copy candidates have a decreased growth rate of 0%-2% and low-copy and genomic integrations have negligible decrease in growth rate.



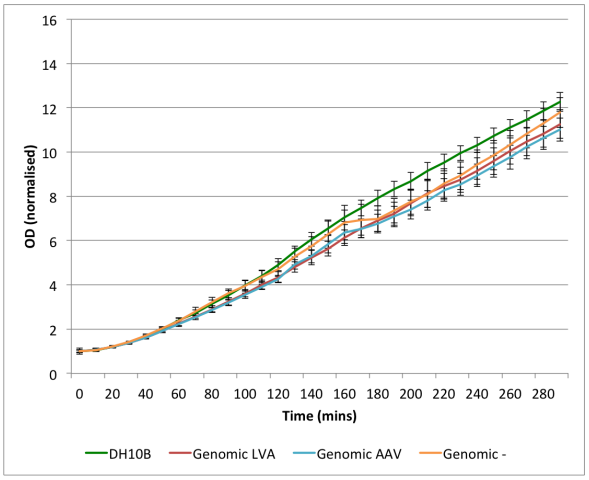
(a) Growth Curves for High-copy candidates (pSB1K3)



(b) Growth Curves for High-copy candidates (pSB3K3)



(c) Growth Curves for High-copy candidates (pSB4K5)



(d) Growth Curves for genomic candidates

Figure 3.8: Growth curves for all monitor candidates grown for 5 hours in M9 media supplemented with 0.4% Fructose with readings taken at 10 minute intervals. Measurements are done on plate reader and averages over 6 repeats with error bars showing standard deviation. DH10B cells are included in all graphs to indicate how growth curves compare to cells without any monitor circuit. Normalisation is performed by dividing each OD reading by the initial reading for the corresponding curve at time = 0.

Growth rates for these different candidate strains were calculated using the formula in Equation **??** over a 40 minute period from 60 minutes to 100 minutes and are shown in Table 3.2. This shows that high-copy candidates have a decreased growth rate of 5%-8%, medium-copy can-

didates have a decreased growth rate of 0%-2% and low-copy and genomic integrations have negligible decrease in growth rate.

$$\text{growth rate} = \frac{1}{\tau} \log_2 \frac{\text{OD}_{t+\tau}}{\text{OD}_t} \tag{3.1}$$

where τ is the time interval between readings.

Candidate	LVA	AAV	No Tag
pSB1K3	5.2%	7.8%	6.8%
pSB3K3	2.1%	0%	0.6%
pSB4K5	0%	0%	0.1%
Genomic	0%	0%	0%

Table 3.2: Decreases in growth rate and maximal OD relative to DH10B for all monitor candidates tested. Estimates made using ODs at 60 and 100 minutes

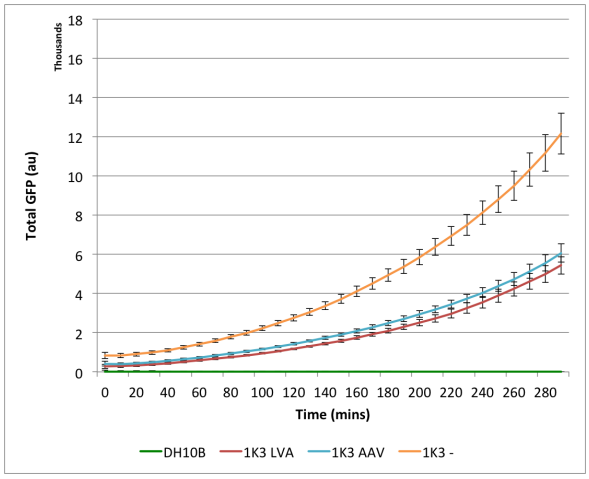
It is clear from these results that the high-copy and medium-copy candidates have retarding effects in terms of the growth rate of cells. Low-copy and genomic candidates have minimal affect on growth rate.

In summary, both high- and medium-copy candidates appear to have major drawbacks when compared to the low-copy and genomic candidates. This is due to the negative impact on growth rate and maximum OD. When considering one of the key requirements for a monitor is that there should be minimal burden placed on the cell (see Section 3.2.5) the medium- and high-copy candidates do not fulfil this.

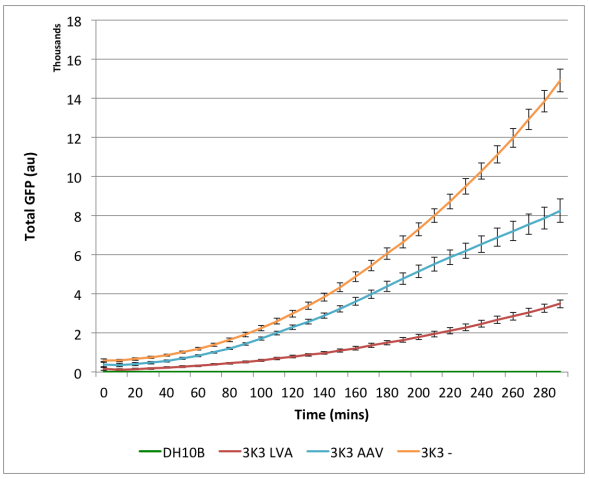
3.6.2 GFP Production

A crucial factor in creating a suitable monitor is that GFP levels are able to be detected at sufficient levels for accurate estimations of productions rates to be made. Figure 3.9 shows the total GFP amounts for each monitor candidate. It can be clearly seen that for each copy-number, the total amount of GFP decreases with the increase in predicted degradation rate: LVA is predicted to degrade proteins faster than AAV, which is predicted to be faster than no

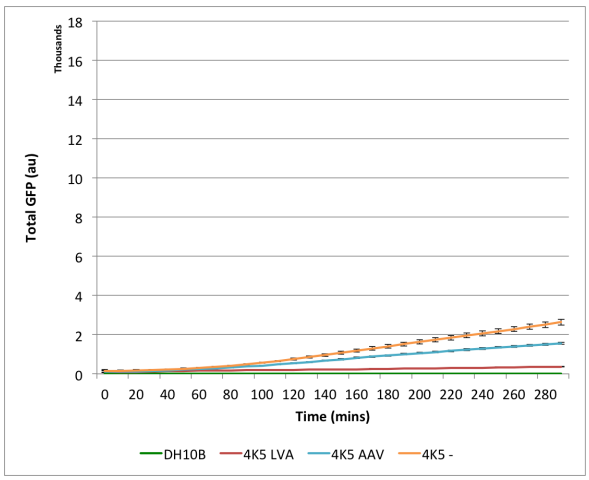
tag (-). An interesting result is that the total GFP accumulation is higher for untagged medium copy candidates than for untagged high-copy candidates.



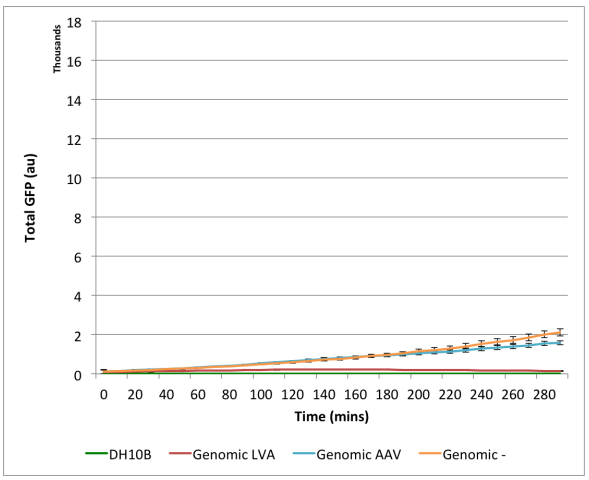
(a) Total GFP for High-copy Candidates (pSB1K3)



(b) Total GFP for Medium-copy Candidates (pSB3K3)



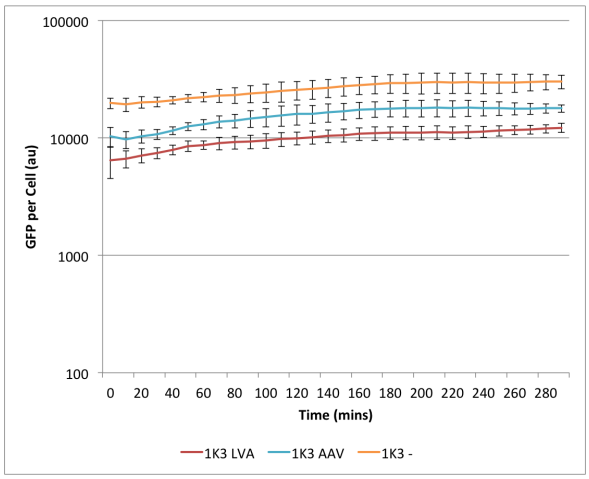
(c) Total GFP for Low-copy Candidates (pSB4K5)



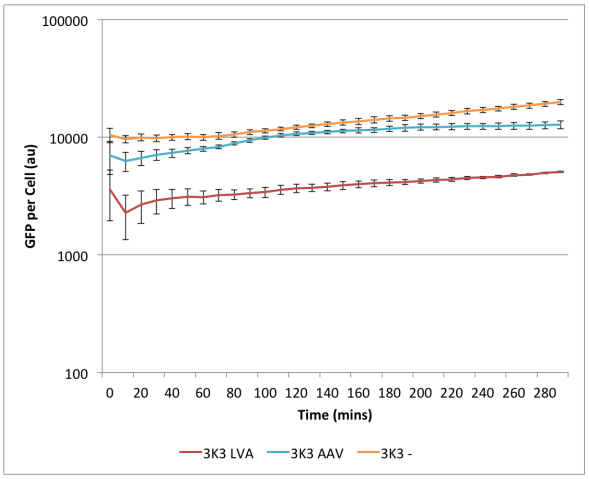
(d) Total GFP for Genomic Candidates

Figure 3.9: Total GFP for all monitor candidates grown for 5 hours in M9 media supplemented with 0.4% Fructose with readings taken at 10 minute intervals. Measurements are done on plate reader and averages over 6 repeats with error bars showing standard deviation. DH10B cells are included in all graphs to indicate how growth curves compare to cells without any monitor circuit.

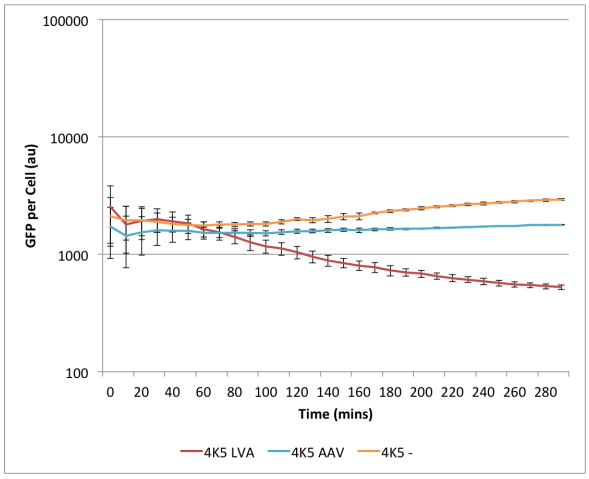
We have seen above that the growth rates and ODs over time can be very different for these different variants. Therefore it is important to look not only at the total GFP, but also to look at the amount of GFP per cell (estimated by dividing GFP fluorescence by OD). We can see the results of this calculation for all candidates in Figure 3.10. By normalising the GFP fluorescence reading we see that the GFP per cell decreases as copy number decreases. Similar to total GFP, the GFP per cell decreases as the predicted strength of the degradation tag increases.



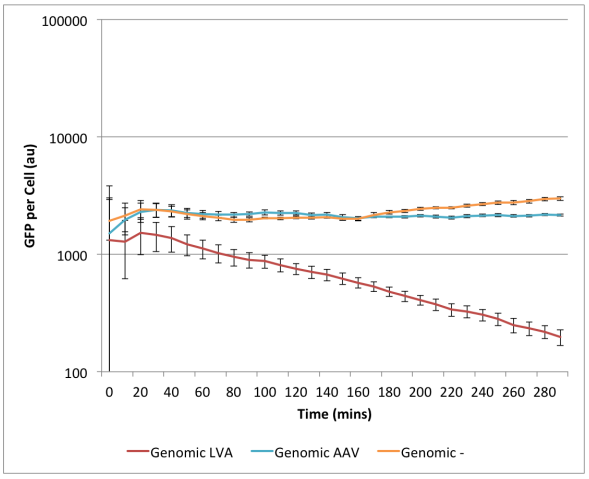
(a) GFP per Cell for High-copy Candidates (pSB1K3)



(b) GFP per Cell for High-copy Candidates (pSB3K3)



(c) GFP per Cell for High-copy Candidates (pSB4K5)



(d) GFP per Cell for Genomic Candidates

Figure 3.10: Total GFP for all monitor candidates grown for 5 hours in M9 media supplemented with 0.4% Fructose with readings taken at 10 minute intervals. Measurements are done on plate reader and averages over 6 repeats with error bars showing standard deviation. DH10B cells are included in all graphs to indicate how growth curves compare to cells without any monitor circuit. Calculated by dividing total GFP by OD. All graphs are shown to the same logarithmic scale to allow ease of comparison. It can be clearly seen that the amount of GFP per cell decreases for each degradation tag and that GFP per cell decreases as the predicted strength of the degradation tag increases.

Using GFP and OD measurements it is possible to calculate GFP production rates for the monitor candidates that do not have degradation tags. We use the Equation 3.2 as a simple ODE model for the rate of change of GFP.

$$\frac{dGFP}{dt} = \alpha - \delta GFP \tag{3.2}$$

where GFP is the amount of GFP per cell, α is the production rate of GFP and δ is the decay rate of GFP (degradation rate + dilution or growth rate). Since the half-life of sfGFP, the variant of GFP we are using in this project, is so long ($>$ 24 hours CITE) the rate of decay is approximated as being the rate of dilution (growth rate of cells). Using these approximations and by taking integrals of Equation 3.2 over a time period τ, it is possible to use Equation 3.3 to calculate the protein production rate α for GFP.

$$\alpha = \frac{\log_2\left(\frac{OD_{t+\tau}}{OD_t}\right)\left(GFP_{t+\tau} - GFP_t \exp\left(\log_2\left(\frac{OD_t}{OD_{t+\tau}}\right)\right)\right)}{\tau\left(1 - \exp\left(\log_2\left(\frac{OD_t}{OD_{t+\tau}}\right)\right)\right)} \tag{3.3}$$

Figure 3.11 shows the GFP production rate for monitor candidates without degradation tags. The high-copy candidate clearly shows the highest production rate of GFP protein, and the rate of protein production decreases alongside the copy number with the genomic insertion having the lowest GFP production rate as would be expected. This is further revealed in Table 3.3.

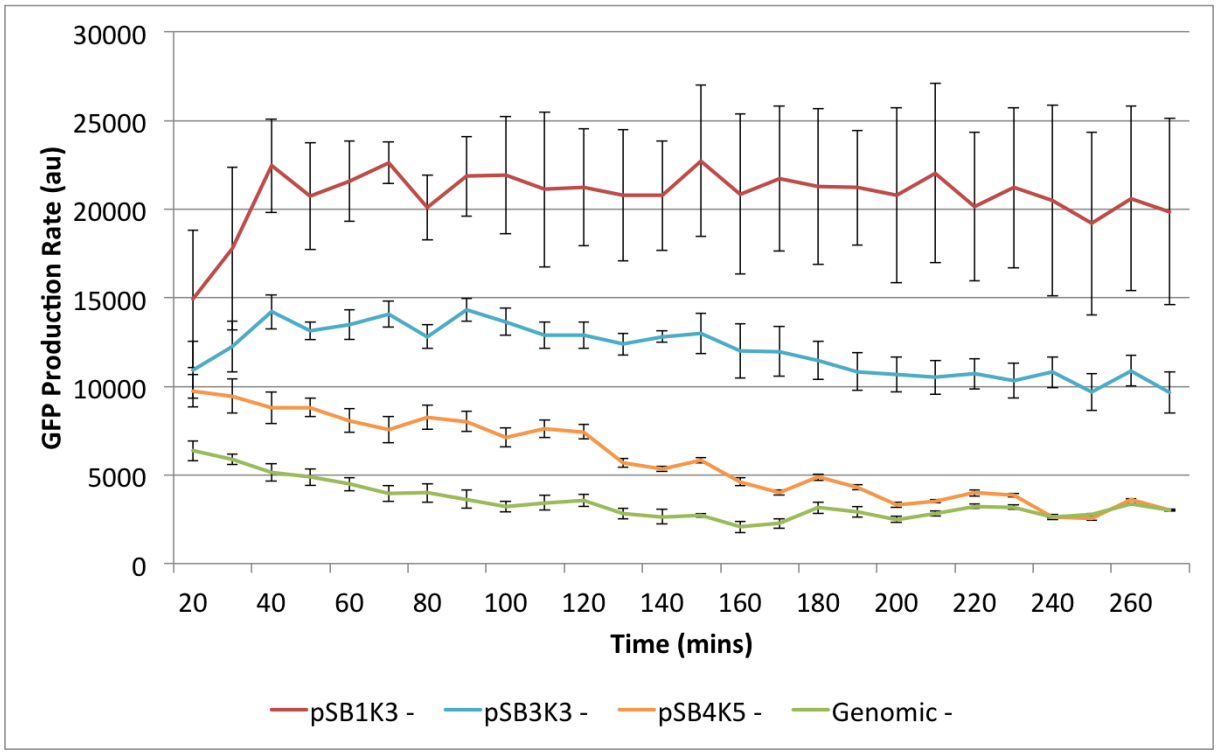


Figure 3.11: Estimated GFP production rates for all candidates with no degradation tag. Estimates made over 10 minute intervals using GFP fluorescence and OD readings and Equation 3.3. All readings made on plate reader and are averages of 6 repeats where errors bars indicate standard deviation.

Copy Number	GFP Production rate (α)
High copy number (pSB1K3)	22882
Medium copy number (pSB3K3)	13814
Low copy number (pSB4K5)	8067
Genomic integration	4482

Table 3.3: Estimated GFP production rates for monitor at each copy number at 60 minutes from initial reading.

In order to accurately estimate production rates for tagged proteins it is important to have accurate values for the degradation rates of the proteins as these values, added to the growth rate, give the protein decay rate δ. Therefore it is crucial to verify the degradation rates of proteins tagged with these degradation tags. These rates are reported in the literature for certain[**?**], however it is necessary to also characterise them in the specific conditions being used in these experiments.

It is possible to calculate the degradation rates of the proteins with different tags in our experiments by utilising the fact that protein production rates across all versions of the monitor at the same copy number are expected to be highly similar. This assumption is made based on the fact that the only differences in the tagged variants are the tags themselves and that factors influencing protein production rate such as copy number, promoter, RBS and codon usage remain constant.

Equation 3.2 cannot be solved analytically to obtain the degradation rate as a function of the other variables and parameters. Therefore Wolfram Alpha CITE was used to apply numerical methods to Equation 3.4 in order to calculate the degradation rates, given GFP per cell values, growth rates and protein production rates.

$$\frac{dGFP}{dt} = \alpha - \delta GFP \tag{3.4}$$

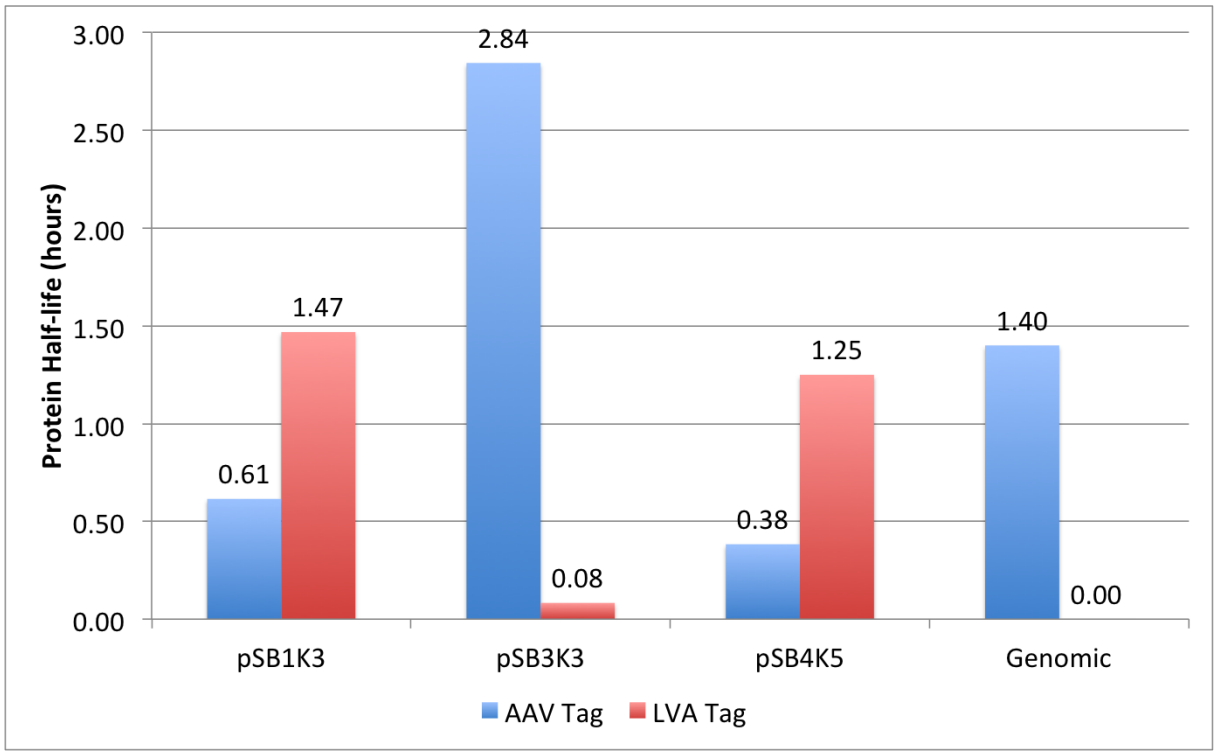


Figure 3.12: Estimated half lives of proteins with degradation tags from monitor candidate devices at various copy numbers. Estimates obtained by using Wolfram Alpha[] to solve Equation 3.2 where δ is sum of degradation and dilution rates.

The results of these calculations are shown graphically in Figure 3.12. The calculated half-lives for the tags at different copy numbers show that there is very little constancy between different copy numbers and from the reported figures CITE. This indicates that the degradation tags should not be included in the final design since estimations of protein production rates for degradation tagged proteins require accurate degradation rates. In addition, it can be seen that monitor candidates at lower copy numbers with no degradation tag show lower amounts of variation than medium- (pSB3K3) and high-copy (pSB1K3) candidates (see Figure 3.3).

3.7 Final Design

It can be clearly seen from the data that GFP fluorescence can be detected at all copy numbers, including genomic integration. As mentioned above, there are a number of benefits to having a monitor system that has been inserted into the genome including:

- No origin of replication incompatibility with any other plasmid-based circuits.
- Lower copy number means there will be a lower impact on the host cell in terms of growth

rate (see Figure 3.10d) and we would also expect a lower impact on shared resources in the cell as there will be lower rates of DNA replication, transcription and translation.

- Maintaining a genetic circuit is known to be much easier when it is integrated into the genome CITE. This means that the monitor will be maintained for a much larger number of generations and will be much more likely to be maintained in cells that are experiencing stress.
- No need for continuous selection with an antibiotic.

There did not appear to be any direct or indirect benefits from using a plasmid-based system over a genomic integration that would mitigate these clear benefits in any way. Therefore it was decided that the final monitor device would be integrated into the *E. coli* genome at the λ-site.

The data clearly show that there is a large inconsistency in degradation rates corresponding to individual tags. This inability to reliably and consistently have an accurate estimate of protein degradation rate means that it becomes very difficult to estimate protein production rates. By using sfGFP without a tag it is possible to accurately approximate the decay rate of GFP per cell as being growth rate, a variable we can accurately and reliably calculate directly and dynamically. The growth rates of *E. coli* mean that changes in GFP production can be measured even when a degradation tag is no present.

Considering all factors there appeared to be no direct or indirect benefit to using a monitor version with a degradation tag, and therefore it was decided that the final monitor version would not use degradation tags.

Our final monitor design was therefore a genomic integration into the λ-site of DH10B. The integrated construct was a linearised CRIM plasmid pAH63 CITE containing synthetic promoter J23100 driving the transcription of a strong synthetic RBS (as predicted by the Salis RBS calculator CITE) and coding region for a DNA2.0 codon optimised sfGFP without any degradation tag, followed by a fully synthetic B1002 terminator (Figure 3.14)

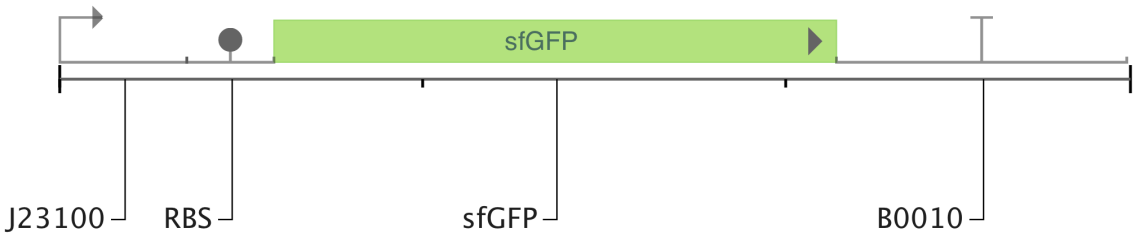


Figure 3.13: The final monitor design consists of promoter J23100 driving the transcription of a strong RBS (as predicted by the Salis RBS calculator CITE) and coding region for a fully codon optimised sfGFP without any degradation tag, followed by a fully synthetic B1002 terminator (not to scale).

It should be noted that whilst this design was able to fulfil most of the requirements outlined in Section 3.2, the final design has some drawbacks that should be acknowledged.

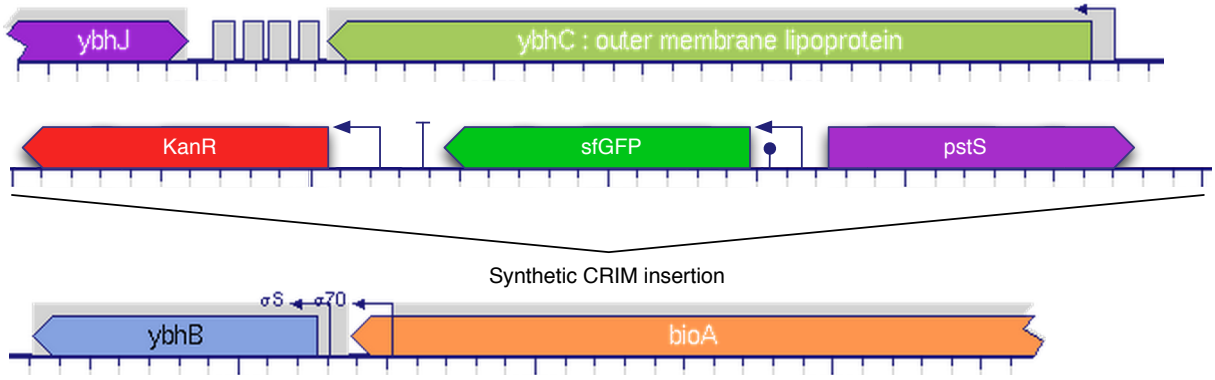


Figure 3.14: Local context of capacity monitor within *E. coli* genome (not to scale). CRIM plasmid including kanamycin resistance marker and pstS site are integrated into the λ-site in the genome between ybhB and ybhC at approximately 860000 bases into the genome of DH10B. When integrated into MG1655, the location is approximately 806000 bases into genome. Figure modified from EcoCYC[]

3.7.1 Design Implications

Clearly the use of GFP as the output protein of the monitor means the monitor will be incompatible with synthetic circuits that use a fluorescent protein that has overlap in the emission or excitation spectrum of GFP. Many synthetic circuits employ GFP (or a derivative of it) as a reporter protein and the use of GFP in the project renders these incompatible. However, many of the benefits that cause GFP to be so widely used, such as extensive characterisation, rapid folding and infrastructure compatibility, mean that the benefits of using it in the monitor design outweigh the disadvantages.

In addition, when integrating the monitor plasmid into the genome using the CRIM system, the entire CRIM plasmid is integrated. This means that features that are crucial for maintenance of the plasmid such as origin of replication and an antibiotic resistance marker are also included into the genome. The origin of replication in the CRIM plasmid is a *pir* conditional origin and only functions in the presence of the *pir* protein. In the *E. coli* strains where the monitor is implemented this protein is not present and the origin is non-functional. The kanamycin resistance marker is still present and functional in the genome. This means that the monitor is incompatible with synthetic circuits that use a kanamycin resistance marker. While kanamycin is a popular antibiotic used in many circuits, many alternatives are available.

CRIM is not the only methodology for integrating into the genome of *E. coli* CITE. After a survey of the alternative methods it was clear that CRIM offered the best set of advantages (such as speed and ease of integration) compared to disadvantages (such as full plasmid integration) of all the options.

Chapter 4

Results: Testing and Verifying Function of the Capacity Monitor

4.1 Designing Test Circuits

In order to test whether the monitor was able to detect a decrease in the resources available in the cell a test circuit was designed. This test circuit was designed to produce mRNA and protein under the presence of an inducer. In addition to the test circuit, suitable controls were designed to confirm that protein production was causing the burden as opposed to the presence of the inducer. In order to design a suitable test circuit a number of specifications were outlined.

4.1.1 Compatibility with Monitor

Clearly it is important to ensure that any test circuit is compatible with the monitor. As mentioned in Section 3.7.1 there are a couple of constrains that the implementation of the monitor places on circuits it can be used with. Circuits must:

1. Not contain GFP or a close derivative of the GFP protein.
2. Not use kanamycin as a selective marker.

These are relatively easy constraints to work with and are taken into consideration in the design of the test circuit.

4.1.2 Interact Through Shared Resources

This test circuit is designed to test whether the monitor is able to detect a decrease in shared resources. In order to do this it is important to ensure that the interactions between the test circuit, the monitor and the host cell occur solely through the shared resource pool. This means other interactions such as cross-talk and toxicity must not be present between the systems. In Chapter 3 we describe how they are avoided between the monitor and the host cell, and will discuss in this chapter how they are avoided in the test circuit.

4.1.3 Inducible Circuit - The AraBAD Promoter Unit

Making the test circuit inducible meant that we could control whether or not protein was being produced from it through the addition of a small inducer chemical. This is important as it was necessary to make comparisons between cells containing the same DNA where one set was producing protein from the circuit and the other was not. By making this comparison one can be sure that any phenotypic differences between the two cases is due to the presence of the inducer chemical.

It was necessary to confirm that the only impact the inducer chemical was having on the cells was to cause additional protein production from the synthetic circuit, and not having any direct interaction with any native cellular functions or behaviours. This could be confirmed by growing similar strains without the inducible circuit, but including the plasmid backbone, and untransformed cells with the capacity monitor and comparing their behaviour with and without the induction chemical.

A widely used inducible promoter system in synthetic biology, and especially *E. coli*, is the AraBAD promoter. This promoter unit consists of a constitutive promoter on the reverse strand (with built-in RBS sequence) driving the expression of protein AraC, followed by a P_{BAD} promoter on the forward strand (see Figure 4.1)[].

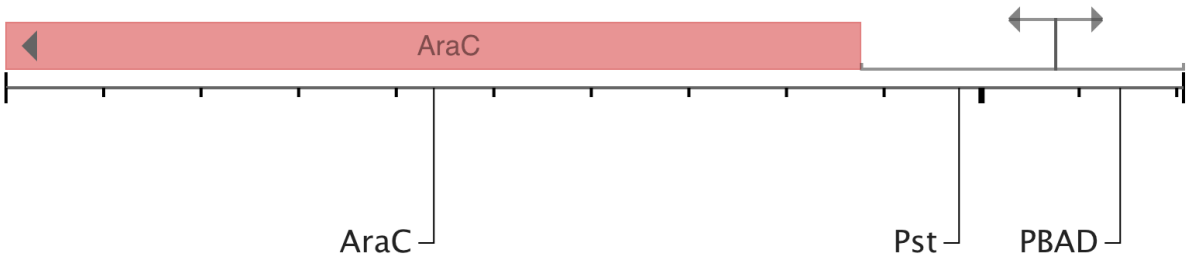


Figure 4.1: The design of the AraBAD promoter unit consists of a constitutive promoter on the reverse strand (with built-in RBS sequence) driving the expression of protein AraC, followed by a P_{BAD} promoter on the forward strand.

The AraC protein acts as both a repressor and activator of the P_{BAD} promoter dependent on the presence of arabinose. When arabinose is not present, AraC forms a dimer which binds to the *araO*$_2$ and *araI*$_1$ sites in the P_{BAD} promoter and causes it to fold over, preventing the binding of RNA polymerase. When arabinose is present, it causes a conformal shift in the AraC dimer which causes it to bind to the *araI*$_1$ and *araI*$_2$ sites and actively recruits RNA polymerase to the promoter[].

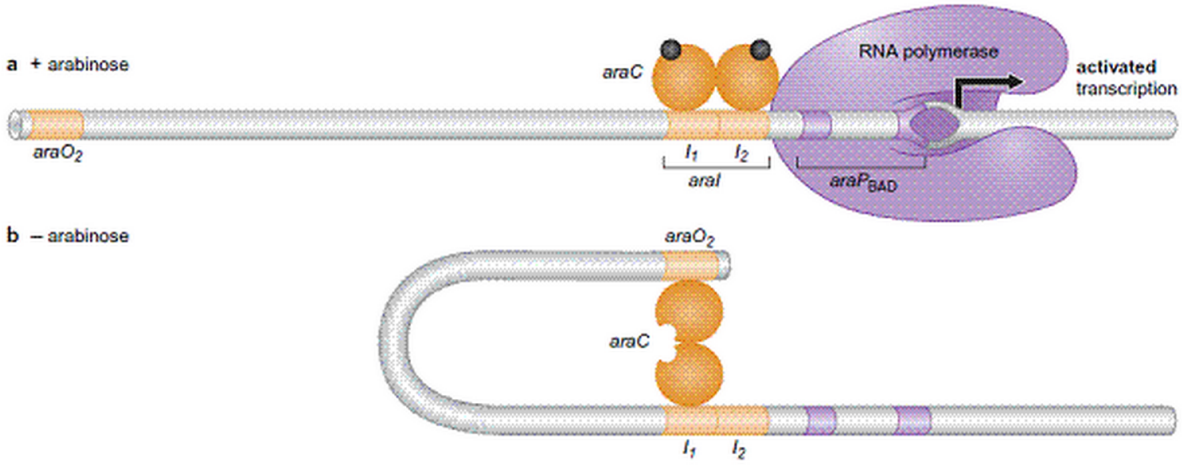


Figure 4.2: When arabinose is not present, AraC forms a dimer which binds to the *araO*$_2$ and *araI*$_1$ sites in the P_{BAD} promoter and causes it to fold over, preventing the binding of RNA polymerase. When arabinose is present, it causes a conformal shift in the AraC dimer which causes it to bind to the *araI*$_1$ and *araI*$_2$ sites and actively recruits RNA polymerase to the promoter. Figure adapted from[].

The cell strain being used as the main platform for this project is *DH10B* which has had the *ara* operons deleted[]. This means that any potential for cross-talk between the host cell or monitor and the AraC gene and P_{BAD} promoter has been removed. In addition, as a native *E. coli* protein, AraC is not toxic to *E. coli* satisfying the key requirement that interactions between

the test circuit and the cell and monitor are exclusively through the shared resource pool.

4.1.4 High Burden on Shared Resources

In order to increase the probability that our test circuit is going to place sufficient burden on the shared resources we need to design a circuit that is likely to place high levels of burden on these resources. Two separate circuits were tested in order to maximise the likelihood that any decreases in monitor output observed are due to the usage of cellular resources in the expression of protein rather than the proteins being expressed. The two circuits tested both contained the P_{BAD} mentioned above expressing operons within a pSB1C3 backbone and are shown in Figure 4.3.

The first operon is the Lux operon from *Vibrio fischeri* which contains the Lux C, D, A, B, E genes. This operon controlled by P_{BAD} was designed by the Cambridge 2010 iGEM team and has a total operon length of 6407 bp. The combination of a strong promoter (P_{BAD}) and the fact it is expressing 5 proteins means a lot of resources should be used in expressing it.

The second operon is the Red Firefly Luciferase from the Japanese Firely *L. Cruciata* which contains two proteins that are 927 bp and 1644 bp long. This operon controlled by P_{BAD} was also designed by the Cambridge 2010 iGEM team.

Both of these operons are entirely heterologous to *E. coli* and therefore should not have cross-talk with the native cellular metabolism. They have both also been codon optimised by DNA2.0 and are placed in a pSB1C3 backbone plasmid[].

4.1.5 Suitable Controls

It was necessary to have a control that would show that changes in capacity monitor output were due to protein expression. The control is the pSB1C3 backbone plasmid backbone with no AraBAD promoter unit or operon, instead replaced by some non-functional DNA. The non-functional DNA is an 81 bp sequence containing 7 zinc finger binding sites that has no functionality within the contact of the chassis cell or plasmid backbone (BioBrick K323039).

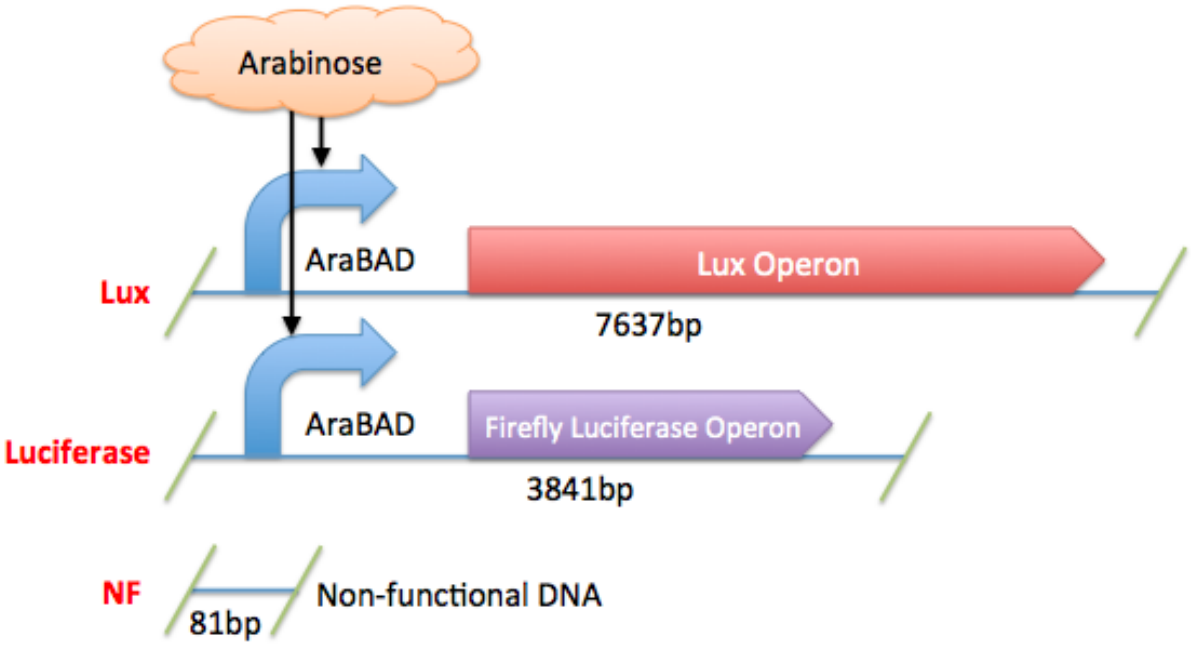


Figure 4.3: Three circuits used to initially test capacity monitor. All inserts transformed into cell with pSB1C3 chloramphenicol resistant plasmid backbone. *Lux* is Lux operon under control of AraBAD. *Luciferase* is Firefly Luciferase operon under control of AraBAD. NF is non-functional DNA. AraBAD promoter is induced by arabinose (black arrows indicating activtion). Lengths of inserts shown under respective diagrams. All circuits are transformed in a BioBrick pSB1C3 backbone.

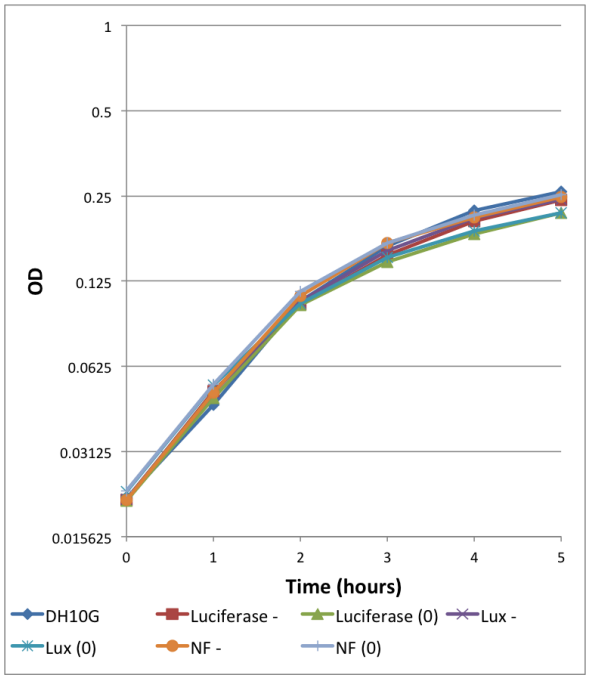
4.2 Results

4.2.1 Initial Results

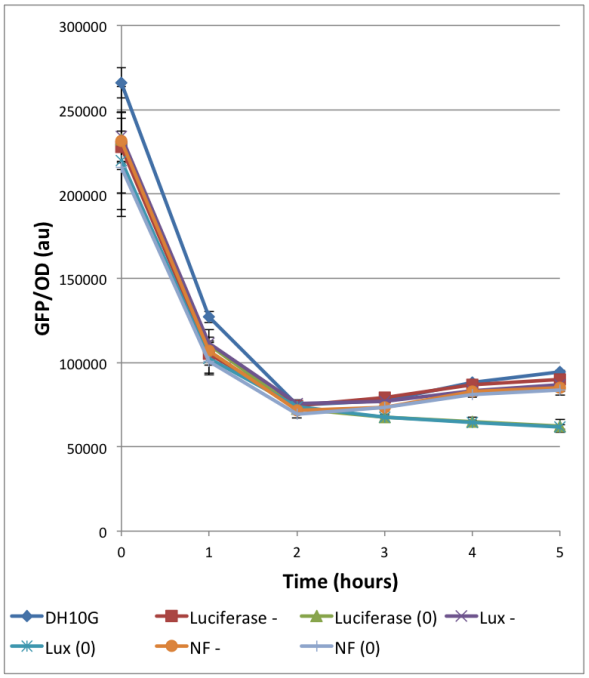
Figure 4.4 shows growth curves and estimates of capacity monitor activity (GFP production rate) over time. Each plasmid was inserted in *E. coli* and grown in LB with and without arabinose added at t=0 and we can clearly observe that when arabinose is added to cells with the inducible promoter (Lux and Luciferase) there is a drop of approximately 50% in capacity monitor activity after 2 hours while cells lacking this additional expression (NF) maintain a higher level of capacity monitor activity in line with control (DH10G cells only containing the capacity monitor). This is a proof of principle that the capacity monitor is able to detect the expression of a level of additional protein by GFP production per cell decreasing.

Once the fact that a decrease in output from the capacity monitor under protein production had been confirmed attention was directed to what it was that the capacity monitor was actually detecting. The drop in GFP production rate only occurs after 1-2 hours after induction. In LB the shift in growth rate at approximately 2 hours occurs as a result of a change in carbon

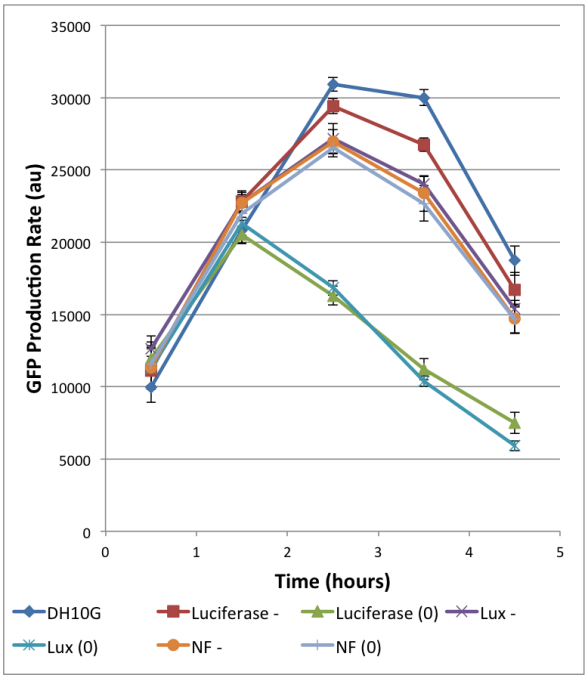
source[?]. The initial carbon source used by the cell is maltose, which is a relatively favourable carbon source for the cell and allows for higher growth rates. After approximately 2 hours this carbon source runs out and the cell has to produce high levels of additional protein so it can use a 'large group of other carbon sources including D-mannose, melibiose [...] glycerol, and lactate'[?]. These carbon sources are less favourable for *E. coli* and limit its growth rate, which is seen in figure 4.4a where the gradient of the graph falls at approximately 2 hours. In order to utilise these alternative carbon sources the cell needs to reconfigure its proteome which requires a large number of ribosomes due to the large number of proteins being translated. Figure 4.4c shows clearly that the drop in output rate from the capacity monitor occurs after 1-2 hours and not at the time of induction.



(a) OD 600 Growth Curves



(b) GFP per Cell



(c) GFP Production Rate

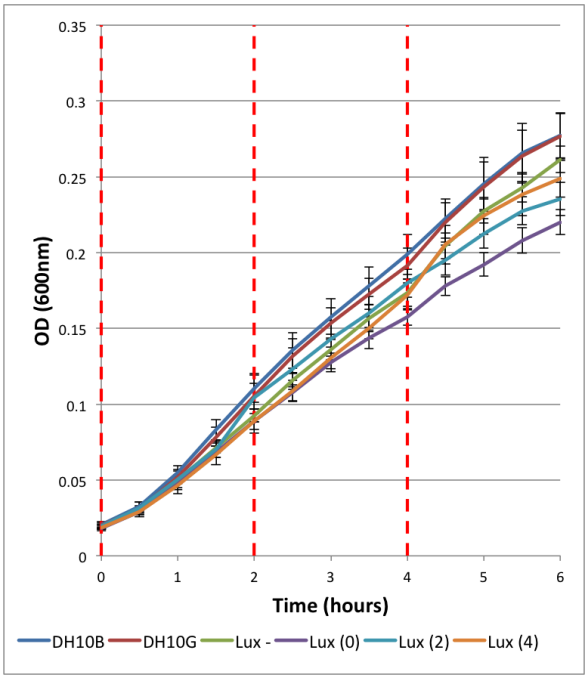
Figure 4.4: Growth, GFP/OD and estimated GFP production rates in LB with arabinose induction at t=0 for cells with no plasmid (**DH10G**) or plasmids with following inserts: Luciferase (K325219) both without arabinose induction (**Luciferase -**) and induction at t=0 (**Luciferase (0)**); Lux (K325909) both without arabinose induction (**Lux -**) and induction at t=0 (**Lux (0)**); and Non-functional DNA (K323039) both without arabinose induction (**NF -**) and induction at t=0 (**NF (0)**). Readings of OD 600 and GFP taken every hour and estimations of production rate made using standard equation between readings an hour apart (a) OD, (b) GFP divided by OD, and (c) Estimated GFP production rate. Error bars show standard deviations over 3 repeats.

4.2.2 Changing Induction Time

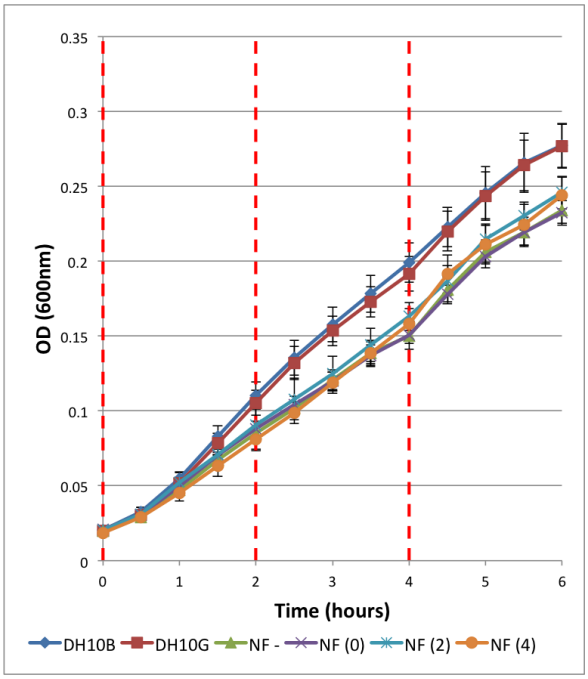
To identify if the drop in monitor output rate was due to a lag caused by transcription or due to a shift in the growth rate, arabinose was added to the media at different time points. If the timing of the drop in GFP production was due to an expression delay we would expect the timing of this drop to shift with induction time. If it was linked to a shift in growth rate we would expect the decrease to occur consistently at the same time. More experiments were performed (not shown) with different induction times and it was observed that there was no significant difference in the behaviour between Lux and Luciferase device-containing cells and therefore in further experiments only Lux device cells were used.

In the next investigation, arabinose was added to the cells in LB media at 0, 2 and 4 hours after the start of the experiment, meaning induction occurs before, approximately at the time of and after the shift in growth rate. The results (Figure 4.5) show that Lux-expressing cells induced at 0 and 2 hours drop in monitor GFP production at the same time, just after the shift in growth rate. Cells induced at 4 hours maintain the same GFP production rate as uninduced cells for approximately half an hour after induction after which GFP production drops in line with cells that had been induced earlier. This means that before the shift in growth rate cells do not see a significant burden due to expression of extra protein from the AraBAD promoter. However, 'at the shift' cells producing extra protein see a decrease in GFP production from the capacity monitor by 30-40% within 45 minutes, indicating that the change in carbon source is causing the burden to manifest.

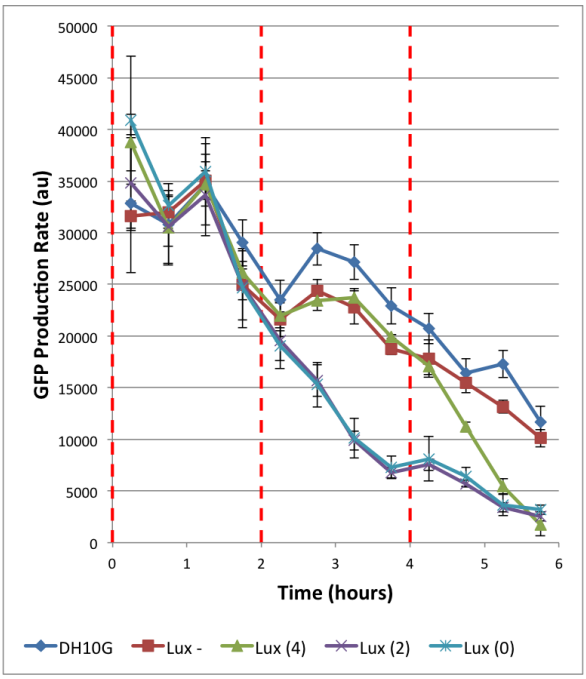
From 0-2 hours the carbon source being used is maltose, which inhibits AraBAD activity, but after this point all of the maltose is used up and the AraBAD unit is not longer repressed. Any induction between 0-2 hours only causes an decrease in GFP production after 2 hours because the induction is negated by the presence of maltose, but any induction after 2 hours causes a decrease in GFP production right away.



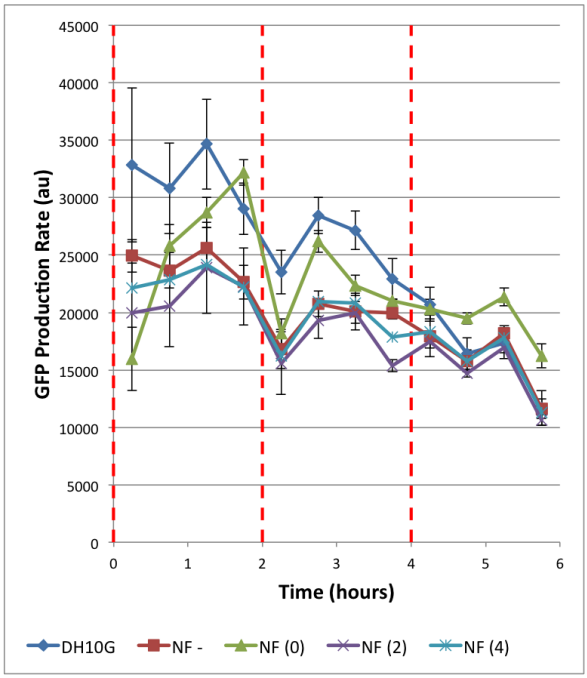
(a) Growth Curve - Lux Device



(b) Growth Curve - NF Device



(c) GFP Production - Lux Device

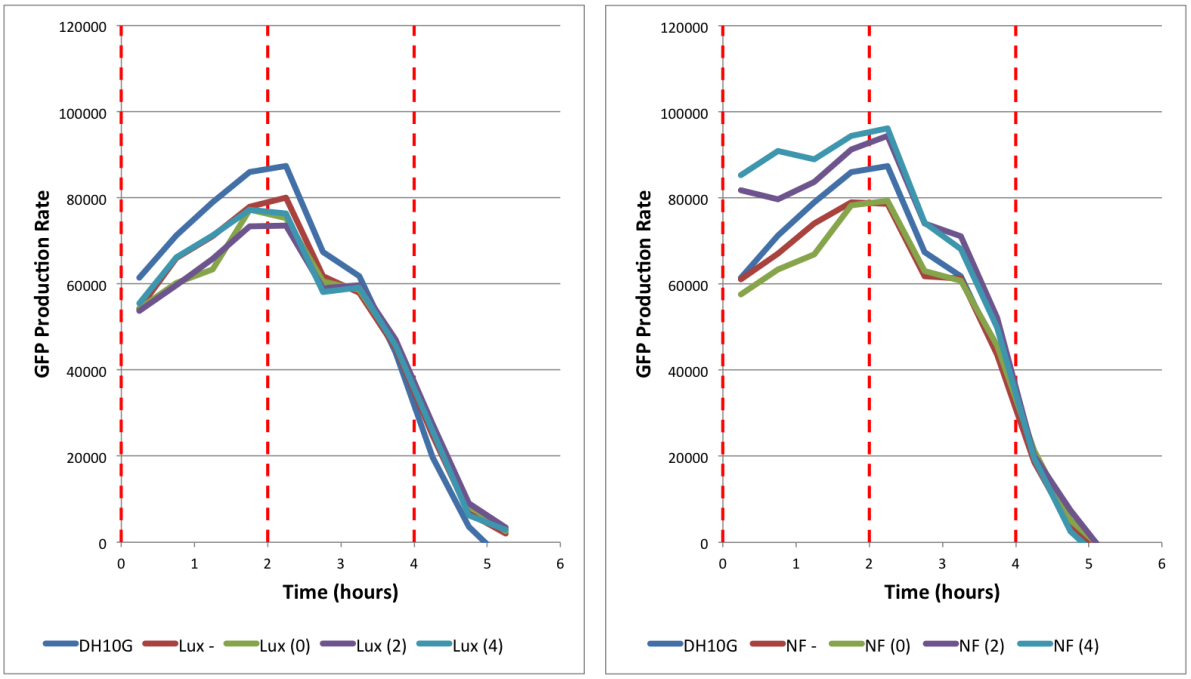


(d) GFP Production - NF Device

Figure 4.5: Growth and estimated GFP production rates in LB with arabinose induction at t=0,2,4 for cells containing plasmids with AraBAD controlled Lux operon - graphs (a) and (c) - and non-functional DNA - graphs (b) and (d). Readings of OD 600 and GFP taken every hour and estimations of production rate made using standard equation between readings an hour apart. (a), (b) OD readings shown with DH10B cells as a negative control for cells containing Lux operon and non-functional (NF) DNA respectively, and (c), (d) estimated GFP production rate for cells containing Lux operon and non-functional (NF) DNA respectively. '-' suffix indicates no induction and numbers in brackets indicated time of induction. Red lines indicate times of arabinose induction. Error bars show standard deviations over 3 repeats.

4.2.3 Using Alternative Media and Carbon Sources

In order to understand fully the impact of different carbon sources used by the cell it is necessary to use a defined media where carbon sources are controlled. LB is an undefined media with a large, undefined mixture of carbon sources[**?**]. M9 is a widely used defined media which can be supplemented by various carbon sources and amino acids[]. The impact of the induced circuit on the cell was first tested in supplemented M9 media with 0.4% glucose, a highly favourable carbon source for *E. coli* growth. In this media we see no significant drop in growth rate or GFP production for cells induced at any time point (see figure 4.6). Since there is a single carbon source there is less likelihood of a shift in the proteome during a growth curve and the culture will simply run out of carbon source and no longer be able to grow. Figure **??** shows that cells containing the Lux circuit show no drop in GFP production, even after induction. This is due to the fact that glucose acts as a repressor of AraBAD and means that no protein is produced from the Lux circuit.



(a) Lux Device

(b) NF Device

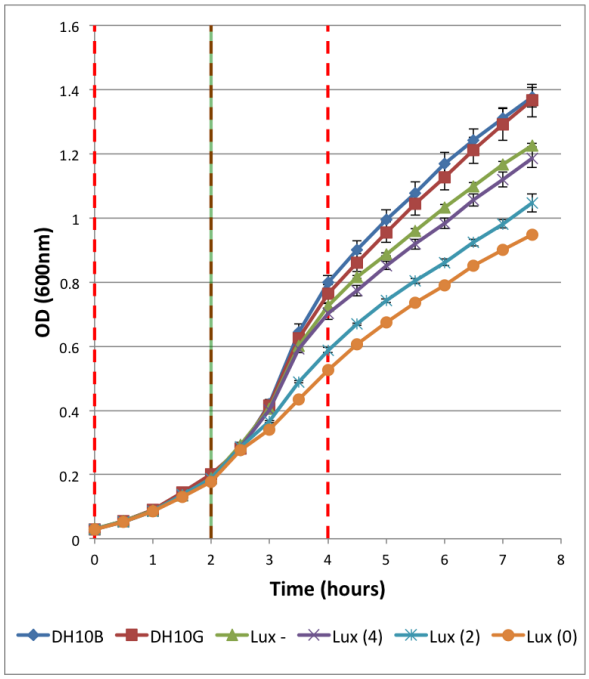
Figure 4.6: Estimated GFP production rates in supplemented M9 + 0.4% glucose with arabinose induction at t=0,2,4 for cells containing plasmids with AraBAD controlled Lux operon - graph (a)) - and non-functional DNA - graph (b). '-' suffix indicates no induction and numbers in brackets indicated time of induction. Red lines indicate times of arabinose induction. DH10G are DH10B cells containing capacity monitor.

In order to separate 1) impact of carbon source shift and 2) extra protein production from the circuit we design an additional experiment. In this experiment we observe the GFP production rate of cells grown in M9 with glycerol for 6 hours as well as cells grown in M9 with glucose and then shifted into M9 with glycerol after 2 hours.

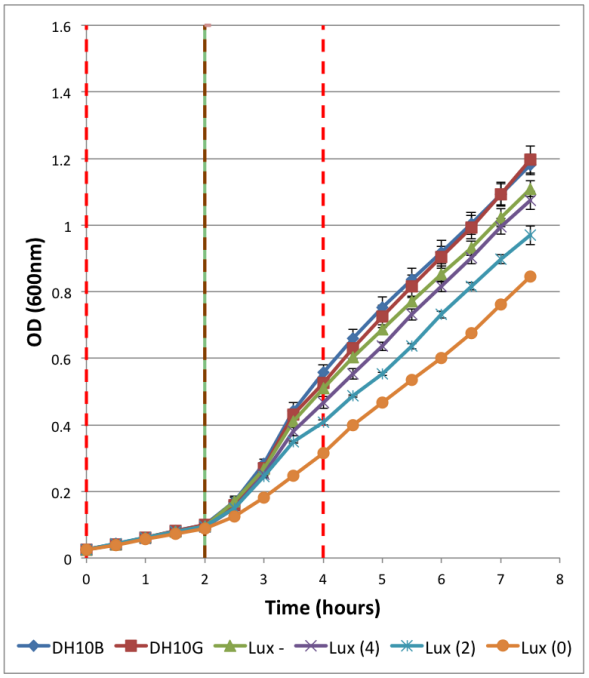
For cells grown in M9 with glycerol a drop in GFP production rate is observed approximately half an hour after induction for each of the induction times (see figure 4.7d). This drop is to 30-40% of the GFP production rate of untransformed DH10G cells. Since there is no shift in carbon source and the cells have been diluted into fresh media the cells do not need to produce extra protein to cope with alternative carbon sources and so we expect no burden effects due to this factor. This indicates that this burden is solely from the burden due to heterologous protein production.

Alongside this experiment we also tested how the cells behaved when initially grown in supplemented M9 with glucose (at 0.04%) and diluted into media containing glycerol as the carbon source (glucose runs out at approximately 2 hours when provided at 0.04%, at which time the cells were diluted into supplemented M9 media with 0.4% glycerol). In this situation, cells that were induced at 0 hours showed no drop in GFP per cell until after the shift in carbon source (see figure 4.7c). GFP production rates for these cells drop to approximately 20% relative to untransformed DH10G cells which is a greater decrease than for cells that have not experienced this diauxic shift.

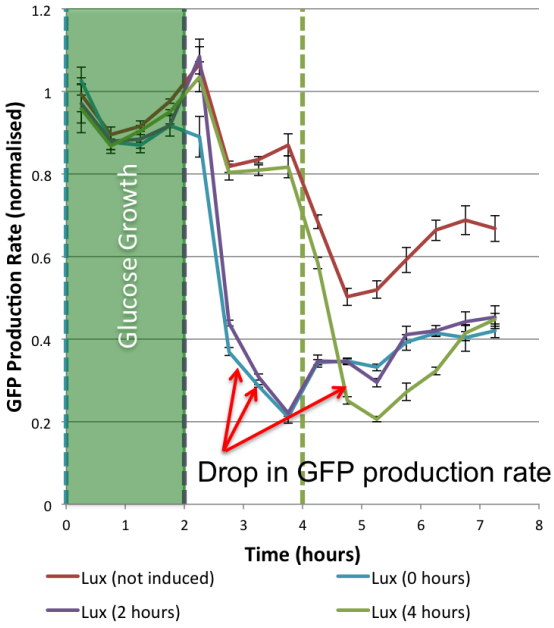
This shows that cells that have not required a shift in their proteome to cope with a shift in carbon source still display a drop in capacity monitor output when heterologous protein production is induced, therefore the monitor is certainly able to detect the increase in burden on shared resources required for this production. A shift in carbon sources requires the cell to produce additional proteins to survive, and this experiment shows that our monitor is able to detect this extra burden caused by a compounding of this increase in cellular protein production with the burden from heterologous protein production.



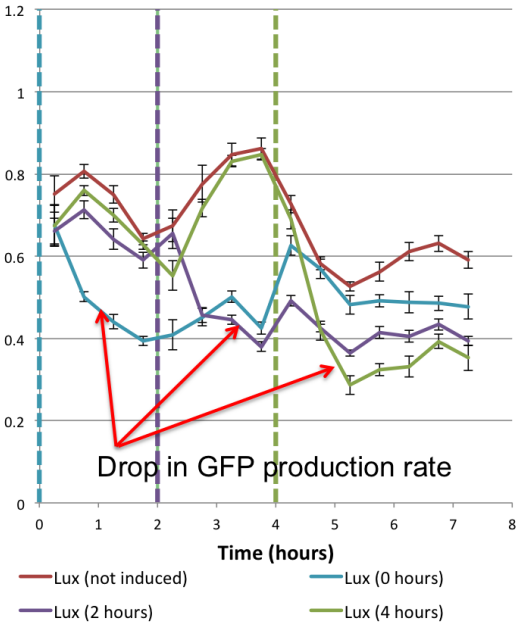
(a) Growth Curve - Glucose -> Glycerol



(b) Growth Curve - Glycerol -> Glycerol



(c) GFP Production - Glucose -> Glycerol



(d) GFP Production - Glycerol -> Glycerol

Figure 4.7: Growth and GFP production rates in supplemented M9 with arabinose induction at t=0,2,4 and dilution at t=2 (but not 4 hours) for cells containing plasmids with AraBAD controlled Lux operon. In (a) and (c) cells were initially grown in M9 + 0.04% glucose and diluted into M9 + 0.4% glycerol at 2 hours and in (b) and (d) cells were initally grown in M9 + 0.4% glycerol and then diluted into fresh M9 + 0.4% glycerol. (a), (b) OD readings with DH10B cells as a negative control, and (c), (d) estimated GFP production rate, normalised against DH10G grown in same conditions. '-' suffix indicates no induction and numbers in brackets indicated time of induction. DH10G are DH10B cells containing capacity monitor. Red lines indicate times of arabinose induction. Green line indicates time of dilution.

As we have seen, there is a difference between how the capacity monitor responds to the expression of protein from a synthetic circuit in different carbon sources. In glucose we do not observe a significant drop in monitor output upon the induction of the Lux device, however in glycerol there is a significant drop in monitor output upon induction. In addition to the presence of this 'burden' another difference we see between cells grown in the different carbon sources is the rate of decline of GFP production from the capacity monitor in cells with no additional synthetic circuit. Figure 4.8 shows that after having approximately the same GFP production rate in the second half hour of growth, for cells grown in glycerol this declines to approximately 75% relative to cells grown in glucose at t=1.5-2 hours. After dilution into fresh media both sets of cells recover to the same GFP output levels at t=3-3.5 hours after which they appear to decline again at approximately the same rate. Since dilution into new media allows the recovery of GFP output, we hypothesise that it is the depletion of resources which causes this decrease. It may be that it is the carbon source itself which is running out, however it may also be the case that cells using glycerol as their carbon source may use some nutrients at a higher rate. In order to identify which resource this might be, rather than diluting into fresh media we could add different nutrients, such as carbon source or amino acids, to the cells to see which induces recovery of GFP output. Additional experiments (not shown) indicated that fructose was a carbon source that could be used without affecting the induction of the AraBAD promoter unit.

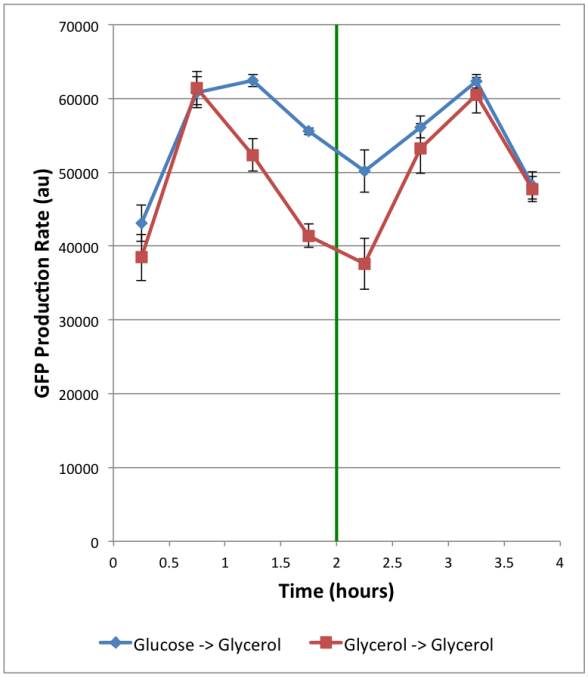


Figure 4.8: GFP production rates in supplemented M9 with dilution at t=2 (with carbon sources before/after: glucose/glycerol and glycerol/glycerol) for cells containing only capacity monitor and no additional synthetic circuits. Green line indicates time of dilution.

4.3 Conclusion

From the results shown in this chapter we can see that our monitor is able to fulfil all of the key requirements outlined in Section 3.2. We are able to use the GFP production rate (as estimated using GFP fluorescence and OD 600 time series readings) as a capacity monitor to estimate the amount of shared resources available within the cell. These quantifications allow us to make dynamic estimates of the capacity in the cell an observe a decrease in cellular capacity within half an hour after the induction of a synthetic gene.

This has been achieved by placing a constitutively expressed, fully synthetic gene into the genome of *E. coli* DH10B at the λ-site. By placing the monitor into the genome we have ensured that it is compatible with any synthetic circuit that 1) does not express GFP and 2) does not use kanamycin resistance for selection. The use of sfGFP as our monitor output means it is highly portable and can be used by any researcher with access to GFP fluorescence quantification equipment. We have also shown that by placing the monitor device into the genome we have minimised the impact on the cell, specifically by not changing the growth rate relative to DH10B cells without the monitor device.

Through a series of experiments we have been able to establish that the phenomenon being quantified by the monitor device is the use of shared resources by a synthetic circuit. We have also been able to observe the additional resources being used by the chassis cell in order to cope with a change in carbon source. Both timing and media choice have been important in these experiments as some of the carbon sources used have caused interactions with the synthetic circuit that mean the induction does not occur at the correct time. In order to avoid these interactions, in subsequence experiments fructose was used as the carbon source in defined M9 media.

Chapter 5

Results: Effects of Gene Expression Control Points on Cellular Burden

5.1 Introduction

After constructing our capacity monitor and confirming that we were able to detect the production of heterologous protein using it we decided to use it to understand how different design strategies could affect the amount of burden placed on a cell.

We designed a combinatorial library where we altered a number of gene expression control features (plasmid backbone, promoter, RBS and codon usage) in order to gain an understanding how altering these impacted on circuit output and the burden it induces.

5.2 Protein Expression Control Points

There are a number of points in a genetic circuit that can be altered in order to change the amount of protein produced. Commonly used control points are detailed below and shown in Figure 5.1.

5.2.1 Copy Number

The number of copies of the circuit that exist in a cell. This can be altered by changing the way the circuit is inserted into the cell. Generally a genomic insertion will have the lowest copy number as it is inserted into a single point in the genome of the host cell. This copy number can change dependent on the location on the genome and the growth rate of cells CITE. Different plasmid-based insertion systems will have different copy numbers that will be dependent on the origin of replication used CITE (or the amount of an inducer for inducible copy number plasmids CITE) as well as growth rate CITE. Changing the copy number can have a profound affect on the amount of protein production as well as the level of impact the synthetic circuit has on a cell. The *Registry of Standard Parts* advises:

> "Oftentimes, it pays to use a different set of plasmid backbones for operating or running your BioBrick device or system than you use for assembly. For example, some BioBrick devices and systems consume too many resources when operated on a high copy plasmid backbone and significantly impact cell growth. In those cases, you might want to switch to a low or medium copy plasmid backbone."[]

This quote applies not only to copy number, but is also indicative of the precise problems this project is looking to understand.

5.2.2 Promoter

The promoter used in expressing a gene has a number of important characteristics. All bacterial promoters require a sigma factor to be present in order to initiate the transcription of RNA molecules. Different sigma factors are present under different conditions, such as σ^{70} which is present in exponential growth CITE or σ^{54} which controls nitrogen regulated genes. In addition, many promoters are either inducible or repressible under the presence of a protein, or protein-substrate complex and are very useful in regulating the amount of transcription that occurs from a promoter. Finally there is the maximal output of the promoter, which is a measure of the amount of transcription that occurs from that promoter under maximal induction, minimal repression or under constitutive expression (dependent on the nature of the promoter). Recent techniques have allowed the creation of promoter libraries CITE that have varying strengths, which can be particularly useful if one wants to maintain the same regulatory characteristics of

a promoter whilst varying its maximal output CITE.

5.2.3 Ribosome Binding Site (RBS)

The strength of an RBS dictates the rate at which translation initiation on an mRNA occurs. The strength is a function of the availability of the RBS to be bound by a ribosome, the binding affinity between ribosomes and an mRNA, and the rate at which elongation is initialised. Numerous tools exist for both the forward and reverse engineering of RBS sequences with respect to strength CITE.

5.2.4 Codon Usage

The selection of codons used to encode a protein can play an important part in the efficiency and rate at which proteins are produced from mRNA transcripts. Of the 20 natural amino acids, 18 can be coded for by at least two different codons, and 10 of those have 3 or more potential codons. The codon used to encode for a specific amino acid affects the rate at which amino acids are added to a protein and therefore how rapidly ribosomes move along the transcript. It has been shown that the choice of DNA sequence chosen to encode a given amino acid sequence can greatly impact protein production rates[] as well as cause other phenomena such as ribosomal stalling[].

5.3 Optimisation of Gene Expression

There exist a number of tools for ‘optimising’ gene expression. These tools include the DNA2.0 codon optimisation algorithm that designs the sequence of a coding region to maximise protein production, the Sails RBS calculator allows researchers to forward and reverse engineer RBS sequences with respect to the rate at which translational initiation occurs. However, these are generally ‘maximising’ approaches whereby they are designed to maximise the expression of a gene. For example, the DNA2.0 technique is designed to provide a DNA sequence that maximises the amount of protein produced per transcript[]. The Salis RBS calculator allows users to select a specific desired RBS strength to have an RBS designed to, or alternatively to ‘Maximize’ the RBS strength for increased protein production[].

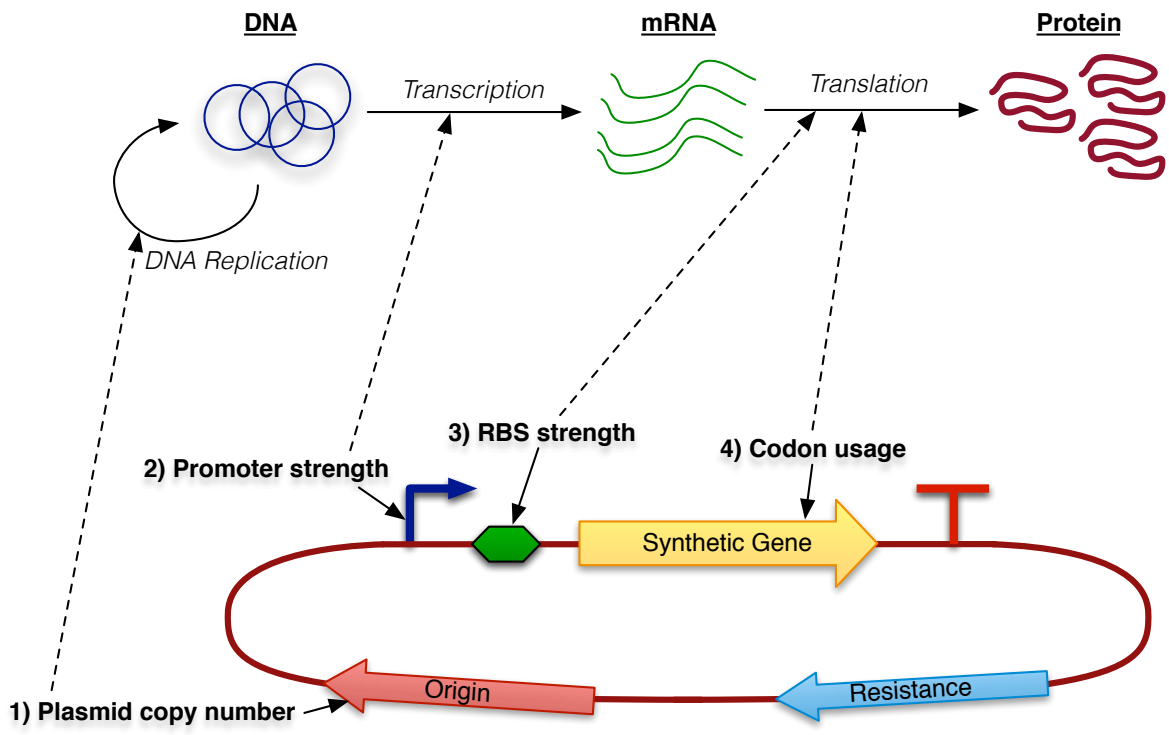


Figure 5.1: Control points for gene expression are shown, as well as how they affect different processes in the central dogma (dashed lines) of protein production from DNA.

It is well known that over expression of heterologous protein can have deleterious effects on a host cell[]. 'Optimisation' of the control points by simply by attempting to maximise protein output can often lead to a slower growth rate, and subsequently a decrease rate of production of the desired product. Usually the researcher must then use a trial and error approach to change some other control points that have not been 'optimised' to improve the circuit so that growth is less impacted and output is recovered.

In this project we are also addressing other types of optimisation, such as investigating how we can minimise the impact a synthetic circuit has on shared resources for a given output. This is important for a number of reasons. It is known that interactions between a synthetic circuit and its host cell can have adverse effects on the circuit behaviour, causing it to 'break' in a number of ways such as causing cell death[], circuit evolution so that function is diminished[]. By minimising the burden placed on the cellular resources we can increase the likelihood of circuits behaving as designed. In addition, if we design two parts of a circuit and characterise them in isolation, we know that expression from one circuit can reduce the resources available to the second and can affect its behaviour, again potentially 'breaking' the circuit. Minimising the resource usage from both circuits will increase the likelihood that they both behave as predicted

when used together and prevent evolution from deleting the parts.

Since the shared resources we are talking about in this project are of many different types, it is also important to understand how heterologous gene expression impacts these different resources (or at least groups of resources). Optimising circuits to minimise the impact on shared resources is a highly non-trivial problem since lowering the amount of one resource that a circuit uses may cause the amount used of a second resource to increase. Understanding the balance between different types of resources and how they affect the behaviour of the cell and of additional synthetic circuits is crucial.

5.3.1 Key Metrics

In order to understand how changes in the control points we have mentioned impact upon these different resources we need to make sure we clarify the key metrics we are both interested in and able to measure. There are a number of quantities that we are interested in with regards to the circuit behaviour, such as:

1. DNA copy number (DNA replication rate).
2. RNA per cell/RNA production rate (transcription rate).
3. Protein per cell/protein production rate (translation rate).

These metrics then have a knock on effect on global cellular factors such as growth rate, contents of the transcriptome or contents of the proteome, as well as many others. These interactions with global cellular factors occur predominantly through the shared resource pool. This section of the project takes a look at how altering these control points affects the cell and other synthetic circuits by utilising the monitor we described in Chapter 3.

From the sort of biological scheme outlined in figure 5.1 we might simply assume that changing copy number affects the amount of DNA polymerase usage, changing promoter strength affects RNA polymerase usage and RBS and codon usage affect the amount of ribosomes being used alone. The actual picture is, as we see both in the literature CITE and from the results in this project, much more complex. These additional complexities manifest themselves in many forms, including relatively obvious phenomena such as copy number and promoter strength affecting the ribosomal usage, all the way to less intuitive interactions such as certain differ-

ent combinations of RBS and promoter giving the same protein production rate whilst altering ribosomal usage.

In addition to these circuit metrics, we are interested in a number of cellular metrics. These allow us to look at how the impact of a synthetic circuit on shared resources translates into changes in native cellular functions and behaviours. The metrics we will be looking at in terms of the cell are:

1. Growth Rate.
2. Total RNA amount per cell.
3. Total protein per cell.

5.4 Test Circuit Design - Specifications

In order to test the impact of the control points mentioned above on both protein output and the impact on cellular resources we designed a combinatorial library of test circuits with a number of different plasmid backbones, promoters, RBSs and codon usages.

As with our monitor we had a number of specifications which needed to be met for a useful library of test constructs.

5.4.1 Compatiblity with Monitor

Clearly our test circuits need to be compatible with the monitor we have designed and implemented. This requires taking into consideration a number of points such as:

- Monitor uses GFP as output, making this an inappropriate reporter from the test circuit.
- Monitor is integrated into the genome along with a kanamycin resistance marker, therefore our test circuit cannot use the kanamycin resistance marker.
- Monitor is inserted into the λ-site of the genome meaning that the expression system for the test devices cannot be a genomic integration into this site.

Aside from these limitations, the design of our monitor system means we have a large amount of freedom in terms of how we implement the circuits.

5.4.2 High Levels of Burden Caused at Maximal Expression

In order to understand how a synthetic circuit impacts a host cell across a wide range of expressions we want to ensure that at the highest levels of expression our circuit is using large amount of cellular resources. This can be done by ensuring that the 'highest strength' variants of each of the control points we are investigating are 'maximised' using the techniques described in Section 5.3. We will do this by implementing the following requirements:

- The highest copy number plasmid we use should be a 'high copy number' with a reported abundance in the order of magnitude of hundreds of copies per cell.
- An inducible promoter should be used such that at its highest levels of induction induces high rates of transcription. For example pBAD is known to have high levels of transcription initiation at maximal induction levels CITE.
- The Salis RBS calculator[**?**] will be used to design our strongest RBS to have a strength of 100,000 au, which is close to the theoretical maximum.
- We will use a long protein (over 2kb) into which we can introduce both slow codons and anti Shine-Dalgarno CITE sequences. Both of these motifs have been shown to reduce translation rates CITE and it is not within the scope of this project to differentiate between the effects of each (indeed, this is covered elsewhere in the literature), all we require is the rate at which ribosomes move along the transcript to be decreased.

5.4.3 Easily Quantifiable Output

As with the monitor, we want to be able to easily and rapidly quantify the rate of protein output from our test circuits. As mentioned above, fluorescent proteins are a commonly used method for quantifying protein levels. As we see above, we require a circuit that will cause high levels of burden at maximal expression levels and therefore want to use a large protein. In order to quantify the amount of this large protein we can fuse a fluorescent protein to it using a small peptide linker sequence.

5.4.4 Minimal non-resource interaction with cell

We require the interaction with the cell and monitor to occur through the shared resource pool. Therefore we must minimise the amount of non-resource-based interaction that occurs between the cell and test circuit. These undesirable interactions can be in the form of regulatory mechanisms, metabolic mechanisms or toxic effects.

5.4.5 Simple Construction of Library

This investigation requires the creation of a large combinatorial library. In order to efficiently create this library we must design our circuits it in a way that allows us to simply and quickly create all of the library members. There are a number of different construction techniques currently available, ranging from restriction based ones to homology based recombinations. We must be pragmatic in our use of these techniques and ensure we are using the best techniques for each part of the build process.

5.5 Test Circuit Design - Implementation

Taking the factors mentioned above into consideration we designed a test circuit that consisted of a single fusion protein expressed from an arabinose inducible promoter unit.

5.5.1 Plasmid Backbones

We chose two plasmid backbones to use in this project that differed in copy number. In order to be compatible with the monitor, we chose to use antibiotic resistance markers for Chloramphenicol and Ampicillin. The plasmid backbones used were:

- **pSB1C3** - High copy number BioBrick plasmid backbone with chloramphenicol resistance and pMB1 origin of replication.
- **pSB3C5** - Medium copy number BioBrick plasmid backbone with chloramphenicol resistance and p15A origin of replication.

5.5.2 Promoter

We decided to use an arabinose inducible promoter unit *AraBAD* as in Chapter 3. More details of this promoter can be found in Section 4.1.3.

In addition to being strong and inducible, another key advantage of using the AraBAD promoter unit was that there exist a characterised library of variants of the P_{BAD} promoter CITE. This allowed a P_{BAD} variant that had increased output to be quickly and easily identified. This meant it was possible to use two versions of P_{BAD} that had different activity levels. The promoter we used from the library at DTU Denamark 2011 iGEM page was version 3[]. The sequence differences between these two variants can be see in Figure 5.2.

It must be noted that whilst the wild-type version of P_{BAD} is a strong promoter when compared to other native *E. coli* promoters, it is weaker than the P_{BAD} variant being used and therefore will be referred to as the *'weak promoter'* throughout the rest of this chapter.

Figure 5.2: Alignment of P_{BAD} Versions - Red bases indicate differences between weak (wild-type) P_{BAD} sequence and strong P_{BAD} sequence. The green and blue boxes indicate the -35 and -10 regions respectively. There is one base mutation in the -35 region and three in the -10 region with no mutations outside of these regions.

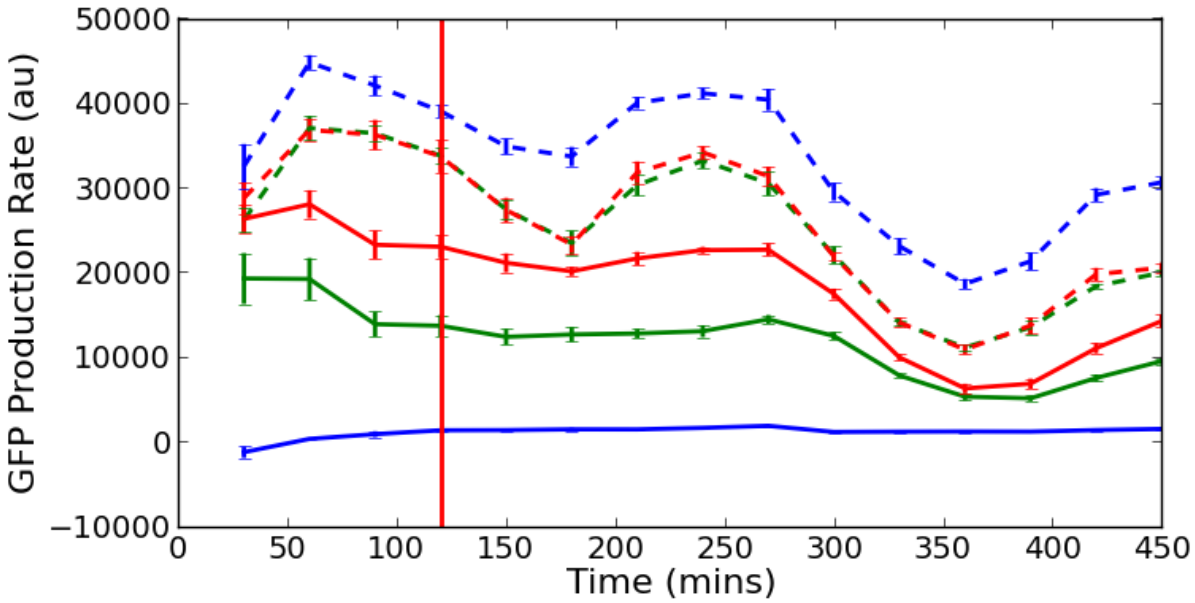
Characterising the variants

As mentioned at the end of the previous chapter, all experiments for this chapter are performed in M9 media supplemented with 0.4% fructose as this carbon source avoid any unintended repression of the AraBAD promoter unit due to carbon source. The cell strains used were, unless otherwise stated, DH10B (an industrially relevant *E. coli* strain and DH10G (DH10B cells with the monitor device inserted into the λ-site in the genome).

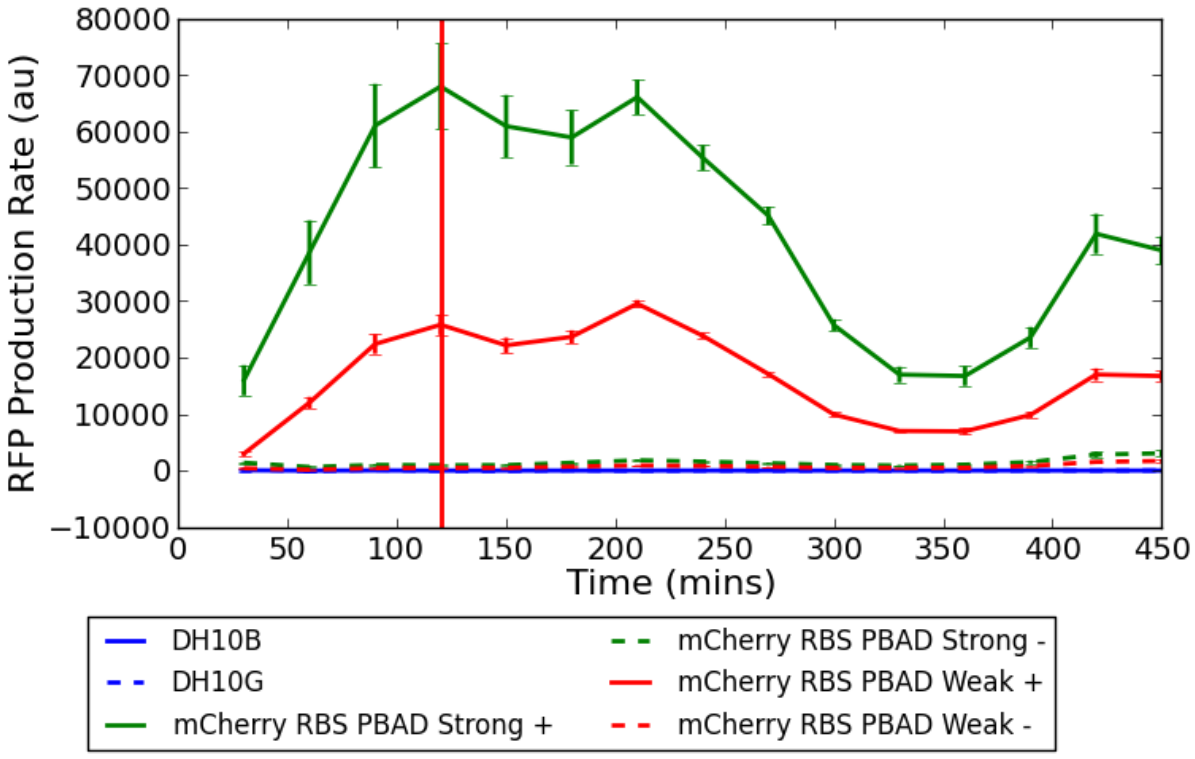
Before using these promoters, their activity levels were characterised by using them to express

mCherry and comparing the protein production rates. This was done by designing a RBS sequence using the Salis RBS calculator CITE (with predicted strength of 20,000au) in a pSB1C3 backbone transformed into DH10B. These two strains were grown for 6 hours in M9 media supplemented with 0.4% fructose along with DH10B and DH10G as controls. Measurements of OD, GFP fluorescence and red fluorescence were taken at 30 minute intervals. These measurements were used to estimate protein production rates and infer promoter strengths using the methodology described in Section 2.14.

Relative activity rates were calculated by taking the relative RFP production rate for each time interval between 2 hours and 8 hours and taking the mean of these relative strengths. Standard deviations were calculated by compounding the standard deviations for each time point with the standard deviation in the mean over time. The low standard deviations seen in Figure 5.4 show that these relative promoter strengths are maintained over the whole growth curve.



(a) Monitor output as represented by GFP production rate



(b) Promoter activity as represented by mCherry production rate

Figure 5.3: P_{BAD} Characterisation. Vertical red lines indicate dilution into fresh media 2 hours after induction a) monitor output as represented by GFP production rate b) promoter activity as represented by mCherry production rate

Figure 5.4 shows the calculated relative mCherry production rates for the two promoter variants.

It can be seen that when induced, the activity of the strong version of P_{BAD} is 2.42 times as

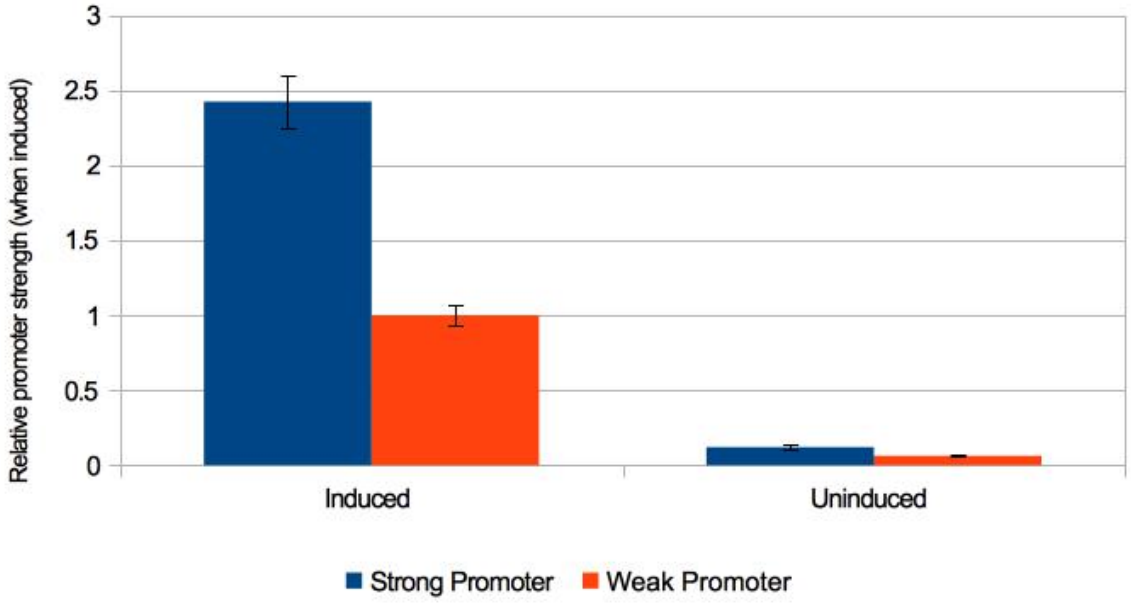


Figure 5.4: Relative promoter strengths of P_{BAD} variants. Average of relative strengths across all time points between 2 hour and 8 hour measurements as shown in Figure 5.3. Error bars show standard deviations.

Name	DTU iGEM Version	ON Activity (relative)	OFF Activity (relative)
Strong P_{BAD}	3	2.42 ± 0.17	0.12 ± 0.02
Weak P_{BAD}	wt	1 ± 0.04	0.06 ± 0.01

Table 5.1: Characterisation data for P_{BAD} variants, including activity in ON (full induction) and OFF (no induction) states. $\pm$ indicates the standard deviation from the mean value.

strong as the weak (wt) version (see Table 5.1 for details).

5.5.3 RBS

The Salis RBS calculator CITE was used to design RBS sequences with a range of strengths. The strengths designed were as shown in table 5.2. Since the RBS calculator is not fully accurate in its predictions, it is usually recommended that 2 sequence designs are ordered for each strength. Since it was only necessary to create a library of RBSs with strengths from weak to strong rather than with specific strengths, it was not required that the RBS sequences designed correlated directly to their predicted values. The primary requirement was that the RBSs used had a range of strengths that include a very high RBS strength. In order to do this, 6 potential RBSs were designed and run through preliminary screening to obtain a small library of 3 RBS sequences with sufficiently different strengths.

We are not able to get accurate estimates for the relative RBS strengths since changing the RBS strength has a number of additional affect on shared resources (which is exactly what this project is looking to address). Therefore we whose the 3 RBS variants that gave a wide range of outputs and behaviours. The results for the RBS characterisation can be seen in Section 5.7.7 where there is a full characterisation of the impact of RBS strength on circuit output, growth rate and monitor output.

Name	Predicted Strength
Strong RBS	100,000
Medium RBS	20,000
Weak RBS	5,000

Table 5.2: RBS strength as predicted by the Salis RBS calculator

5.5.4 Coding Region

The protein we chose to use is VioB, which is the second protein in the Violacein pathway and catalyses a reaction that dimerises *indole-3-pyruvic acid imine (IPA imine)* into *IPA imine dimer*. From a search of the *E. coli Metabolome Database (ECMDB)* and the *Kyoto Encyclopedia of Genes and Genomes (KEGG)* we have ascertained that neither the substrate nor the product of this reaction are naturally present in *E. coli*. This means that this enzyme should not interact with the native metabolism of the host cell, i.e. it is orthogonal. VioB is a bacterial protein from *Chromobacterium violaceum*, which along with *E. coli* is a gram-negative bacteria and is not known to cause any toxic effects on *E. coli*.

VioB is 2994 bp long (998 amino acid protein) and has been codon optimised by DNA2.0 for the 2009 Cambridge iGEM team. This codon optimised version forms the ‘fast codon’ version of the coding region and the ‘slow codon’ version has all of the argenine, isoleucine, leucine and proline codons between 2772 bp and 2952 bp replaced with ‘slow’ versions (AGG, ATA, CTA and CCC respectively) as well as anti Shine-DalgarDalgarnono sequences[**?**] (see Figure 5.5).

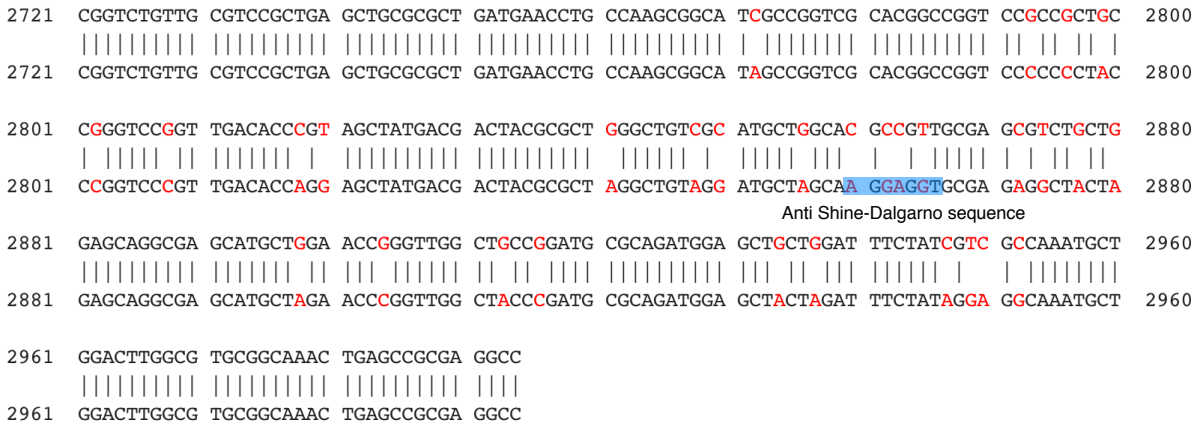


Figure 5.5: Comparison of the sequence of slow and fast codon versions of VioB between 2721 bp and 2994 bp. Top sequence is codon optimised VioB and bottom sequence is VioB with slow codons and anti Shine-Dalgarno sequence inserted. Blue highlighted region indicated anti Shine-Dalgarno sequence and red bases indicate mismatches between the two sequences.

In order to easily quantify this protein, it has been tagged with mCherry protein. This is a fluorescent protein that has an emission wavelength in the red spectrum and does not overlap significantly with the GFP spectrum, meaning it can be accurately quantified when in the same system as GFP proteins (the monitor). The amino acid sequence of the linker is show below and is suggested as a suitable linker for bifunctional fusion proteins by Arai et al.[?].

AEAAAKEAAAKEAAAKA

5.6 Final Design

In total, the library of constructs included combinations of:

1. 2 plasmid backbones
2. 3 RBS sequences
3. 2 promoter sequences
4. 2 codon sequences

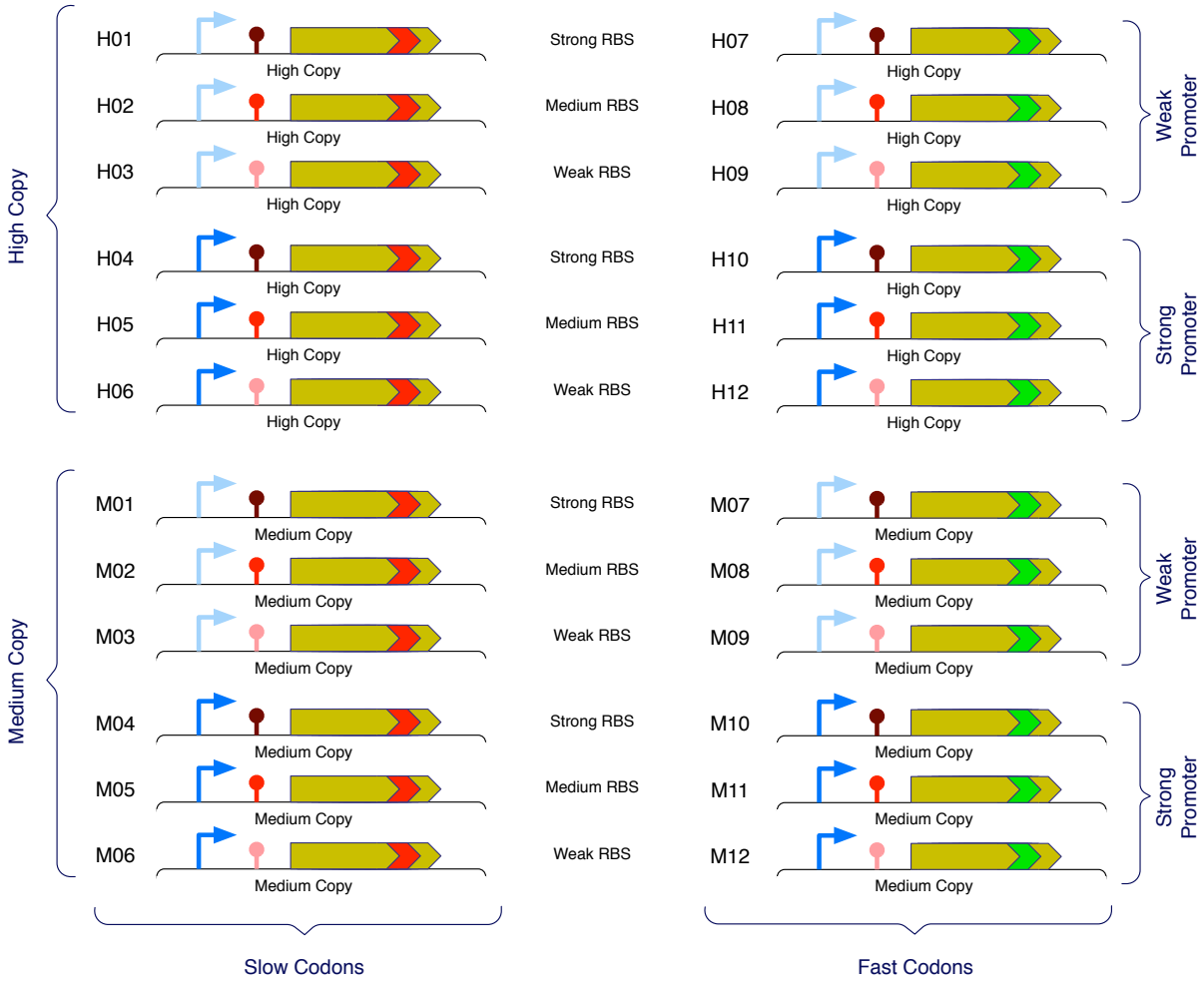


Figure 5.6: Library of construct being used in this chapter consists of all combinations of medium or high copy backbone (as indicated by text), RBS strength (dark red, medium red and light red represent strong, medium and weak RBS respectively), promoter strength (dark and light blue represent strong and weak promoter respectively) and codon speed (green arrow and red arrow in CDS indicate fast and slow codons respectively).

This is a total of 24 constructs that were characterised both induced and uninduced, and the results of these characterisations are shown through the remainder of this chapter. All the combinations and construct names are shown in Table **??**.

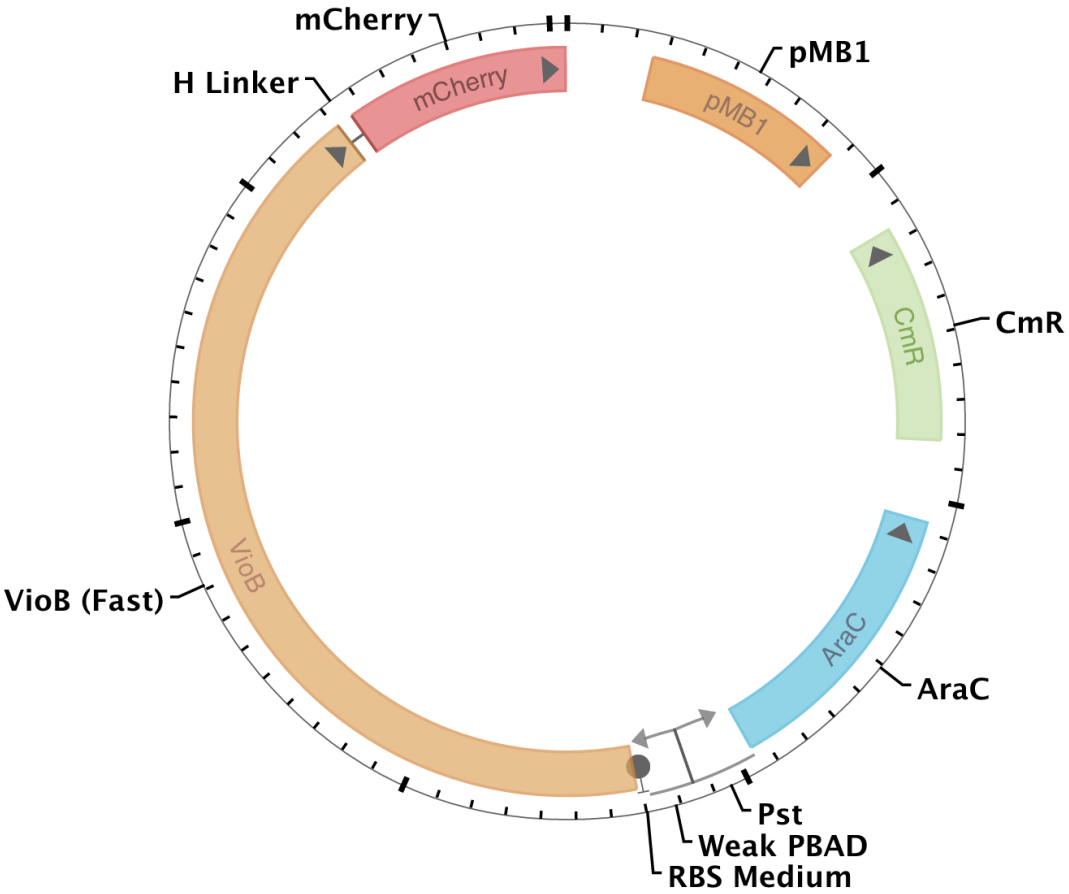


Figure 5.7: Plasmid map showing reference construct with pMB1 origin and CmR resistance marker in the pSB1C3 backbone. AraC, Pst and Weak PBAD form the AraBAD promoter unit and the VioB (Fast) is the codon optimised version of VioB tagged to mCherry.

In order to make reliable comparisons a reference construct was chosen that could be compared against for changes in each of the control points. This construct is referred to throughout this results chapter as the reference construct (see Figure 5.7 for the plasmid map) and has the following properties:

1. Weak (wild-type) P_{BAD} promoter.
2. Medium RBS
3. High-copy plasmid pSB1C3
4. Fast codon design

5.7 Reference Construct Characterisation

The full characterisation of the reference construct is shown here, where the full data on the following variables is shown over a time series:

1. Total OD
2. Growth Rate
3. Total monitor protein
4. Monitor protein per cell
5. Monitor protein production rate
6. Total circuit protein
7. Circuit protein per cell
8. Circuit protein production rate

Using this information we motivate how we move towards a snapshot of key metrics to represent the circuit characterisation in the rest of this thesis.

5.7.1 OD and Growth Rate

Untransformed DH10G and both induced and uninduced DH10B cells transformed with reference construct H07 were grown over a period of 3 hours in M9 media supplemented with 0.4 % fructose. OD readings were taken every 10 minutes and growth rates were estimated over a 1 hour moving window using Equation 3.1.

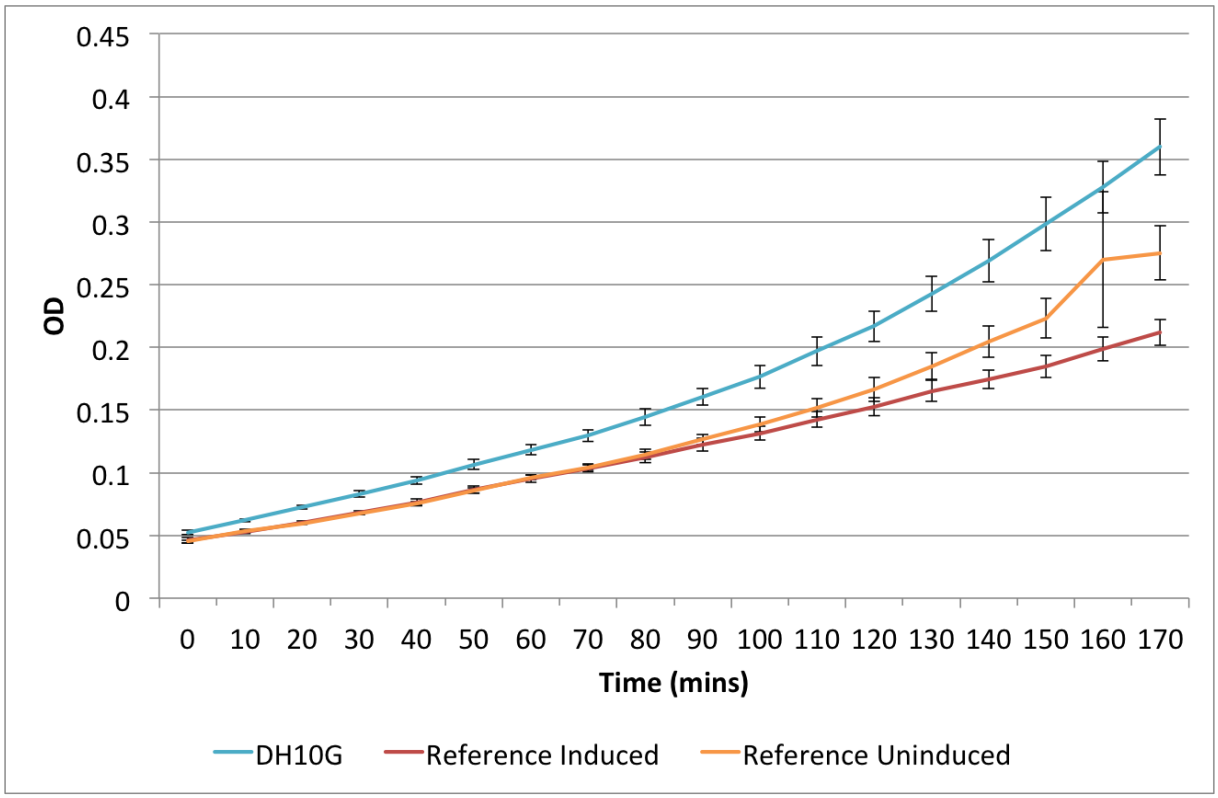
Figure 5.8a shows the growth curves for DH10G with no additional circuit as well as DH10G cells with the H07 reference circuit both induced and uninduced. It can clearly be seen that empty DH10G cells are at a higher OD at all time points, including the initial time point. The induced and uninduced cells containing the circuit start at very similar ODs but diverge as the cells grow, indicating that the induced cells grow at a slower rate compared than those that are uninduced. At 160 minutes after induction there is a jump in the average OD of uninduced cells that does not conform to the trajectory of the curve up until this point. There is also an increase in the standard deviation and upon closer inspection of the underlying data it is apparent that

this is due to a single sample having a high OD reading, most likely an error in the reading. Due to the differences in starting OD this graph does not show clearly the difference in the growth rate of empty DH10G cells compared to uninduced cells containing the reference construct.

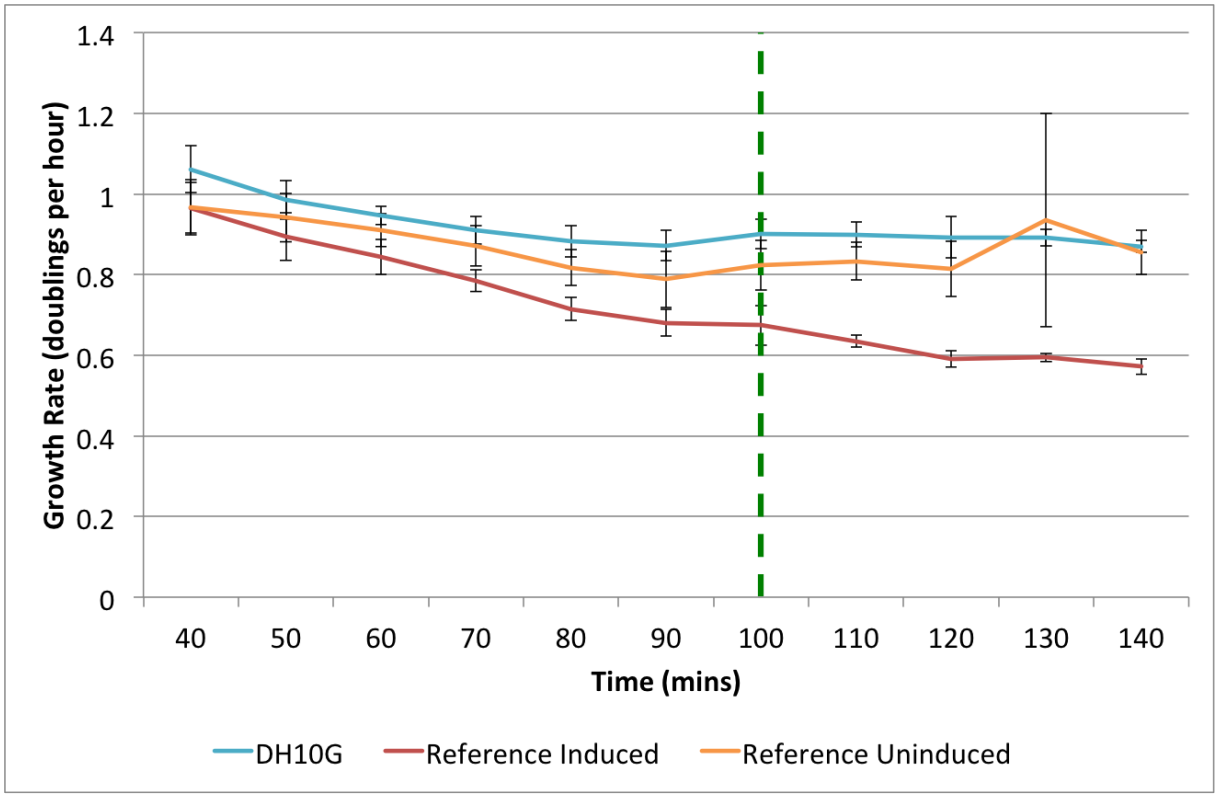
Figure 5.8b shows the calculated growth rates over a 60 minute window (30 minutes either side of the corresponding time point shown on the x-axis). It can be seen that DH10G cells without a circuit grow at the highest rate, confirming what Figure 5.8a indicates. This means that the maintenance of the plasmid containing the circuit as well as the expression of the antibiotic resistance marker (chloramphenicol resistance) causes a slight decrease in the growth rate. The increase in average growth rate for uninduced reference strain cells at 130 mins (average growth rate between 100 and 160 mins) is an artefact of the erroneous measurement of the sample at 160 mins mentioned above.

It can be clearly seen that upon induction the growth rate of induced reference strains cells decreases relative to the uninduced ones. This means that the extra production of protein is causing a decrease in the growth rate of the induced cells.

The dashed vertical green line in Figure 5.8b indicates the mid-point of the interval at which a snapshot of growth rate is taken (between 100 and 160 minutes). This can be seen in Figure 5.11a, where growth rate is shown alongside a snapshot of the monitor output taken over the same time interval. A snapshot of circuit output from the same time interval is shown in Figure 5.11b.



(a) OD



(b) Growth Rates

Figure 5.8: OD and Growth Rate Comparison for Reference Construct. DH10G and both induced and uninduced DH10B cells transformed with reference construct grown in M9 + 0.4% fructose for 3 hours continually in 96-well plate with 200 µl volume per well. Readings taken every 10 minutes and rates estimated over a 1 hour moving window. a) OD levels b) growth rate calculated over 60 minute window, dashed green line indicates time of growth rate, monitor output and circuit output snapshot.

5.7.2 Monitor Output

As in the previous section, DH10B, DH10G and both induced and uninduced DH10B cells transformed with reference construct H07 were grown over a period of 3 hours in M9 media supplemented with 0.4 % fructose. OD and GFP fluorescence readings were taken every 10 minutes and GFP production rates were estimated over a 1 hour moving window using Equation 3.3.

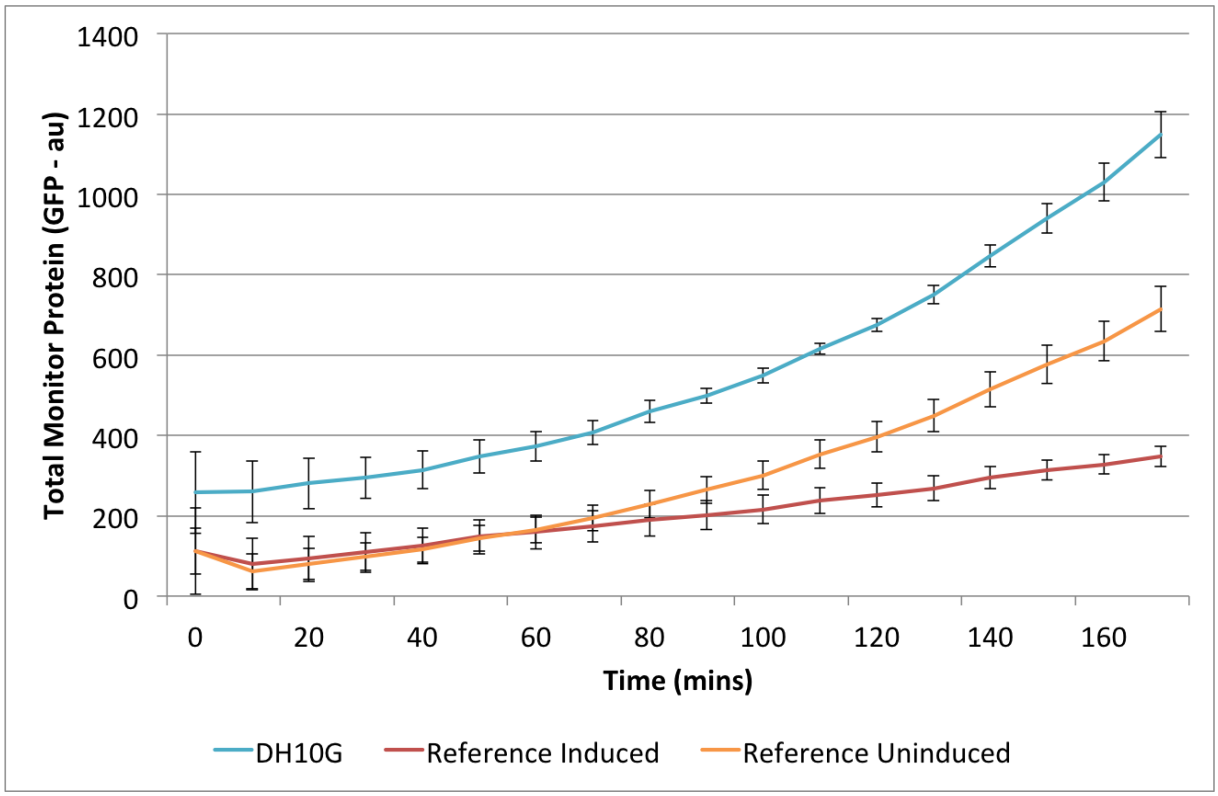
Figure 5.9a shows the average total amount of GFP fluorescence per culture. Empty DH10G cells clearly have the largest amount of total GFP and the distance between the DH10G curve and that of the uninduced reference construct strain increases, meaning that the total amount of GFP production is greatest for DH10G cells. Both the induced and uninduced reference strains have the same initial total GFP as they are dilutions from the same starting culture and have not had any time after being induced to develop distinct phenotypes. After approximately one hour the curves start to noticeably diverge and it can be seen that the total amount of GFP being produced by the uninduced reference strains is higher than the induced ones.

This graph does not contain enough information to understand the differences between the monitor output for the different strains. This is because the effect of the number of cells producing GFP (OD) is not taken into account. Figure 5.9b shows the calculated amount of GFP per cell (GFP fluorescence divided by OD). At the start of the growth the amount of variance is large, due to dividing by small ODs where small amounts of noise due to measurement errors can have a large impact. After approximately 90 minutes a clear trend emerges in terms of the amount of GFP per cell where DH10G cells have the most GFP per cell, uninduced cells containing the reference construct have approximately 80% of the GFP per cell and induced cells containing the reference construct have slightly greater than 50% of the GFP per cell compared to empty DH10G cells.

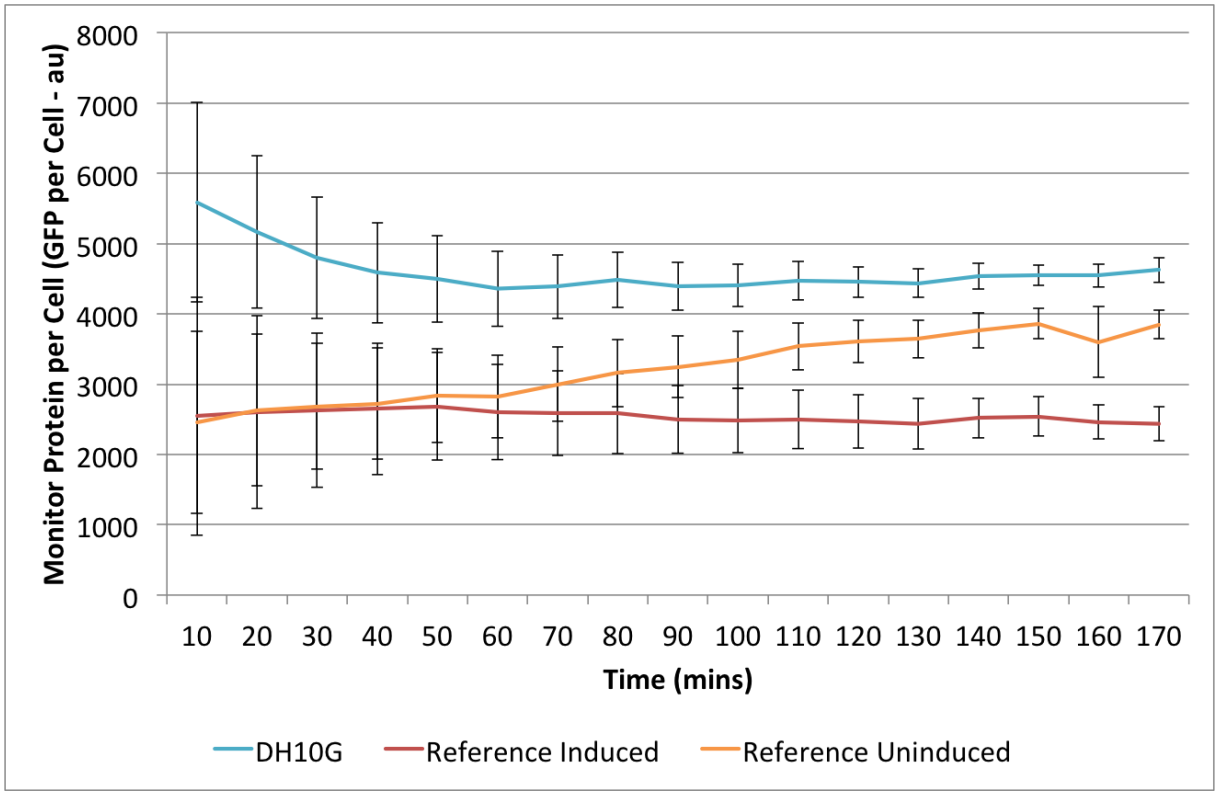
Figure 5.9b also does not contain enough information to tell what the monitor output is for the different strains as it does not take into account the dilution rate (cellular growth rate) of the GFP protein. Figure 5.9c shows the calculated GFP production rate per cell for the three different strains. These calculations were made using Equation 3.3. Again, only after approximately 90 minutes does a clear trend emerge. This is likely due to the time required to move from early log-phase growth into steady-state mid-log-phase growth where gene expression will occur at

a burden-inducing level.

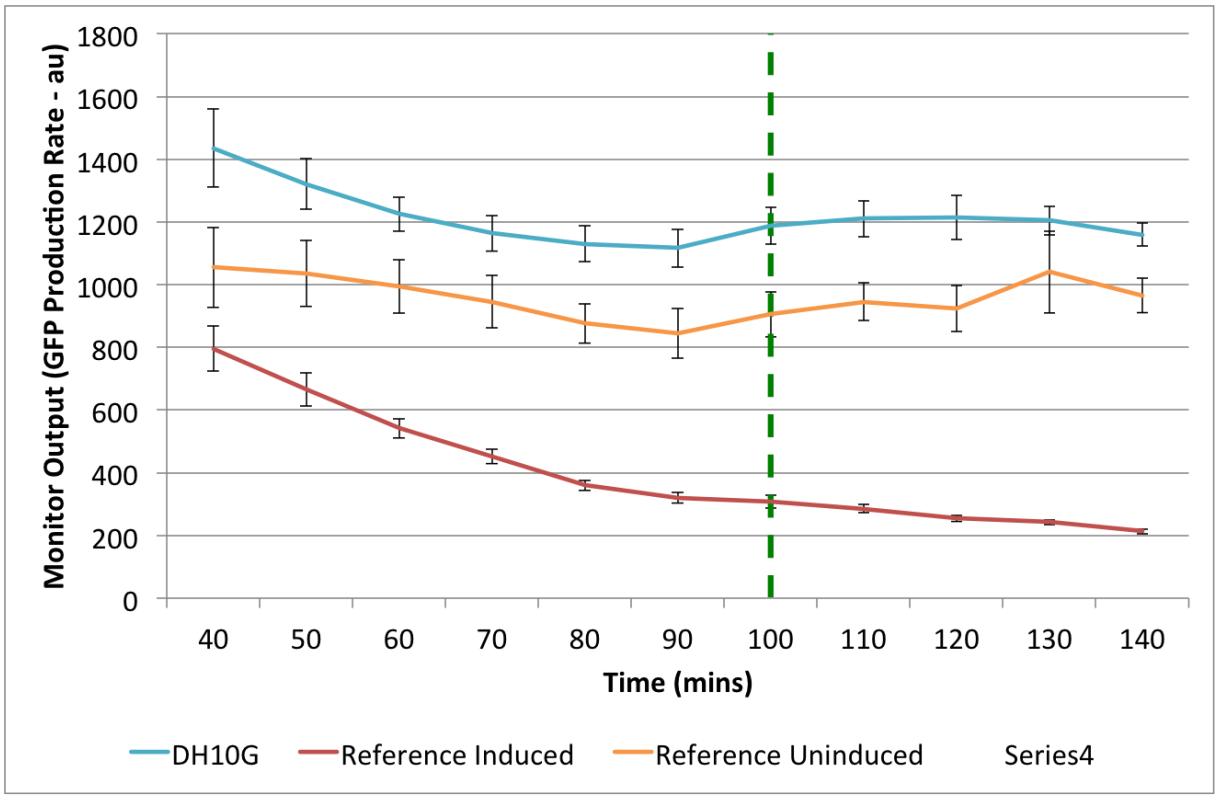
The dashed vertical green line in Figure 5.8b indicates the mid-point of the interval at which a snapshot of monitor output is taken (between 100 and 160 minutes). This can be seen in Figure 5.11a, where growth rate is shown alongside a snapshot of the growth rate taken over the same time interval. A snapshot of circuit output from the same time interval is shown in Figure 5.11b.



(a) Total Monitor Protein



(b) Monitor Protein per Cell



(c) Monitor Output

Figure 5.9: Monitor Protein Comparison for Reference Construct. DH10G and both induced and uninduced DH10B cells transformed with reference construct grown in M9 + 0.4% fructose for 3 hours continually in 96-well plate with 200 μl volume per well. Readings taken every 10 minutes and rates estimated over a 1 hour moving window. a) total culture GFP fluorescence levels are indicative of the total amount of monitor protein (GFP) in all cells b) amount of monitor protein per cell is calculated by dividing the GFP fluorescence levels by the OD c) monitor output (GFP production rate) calculated over 60 minute window, dashed green line indicates time of growth rate, monitor output and circuit output snapshot.

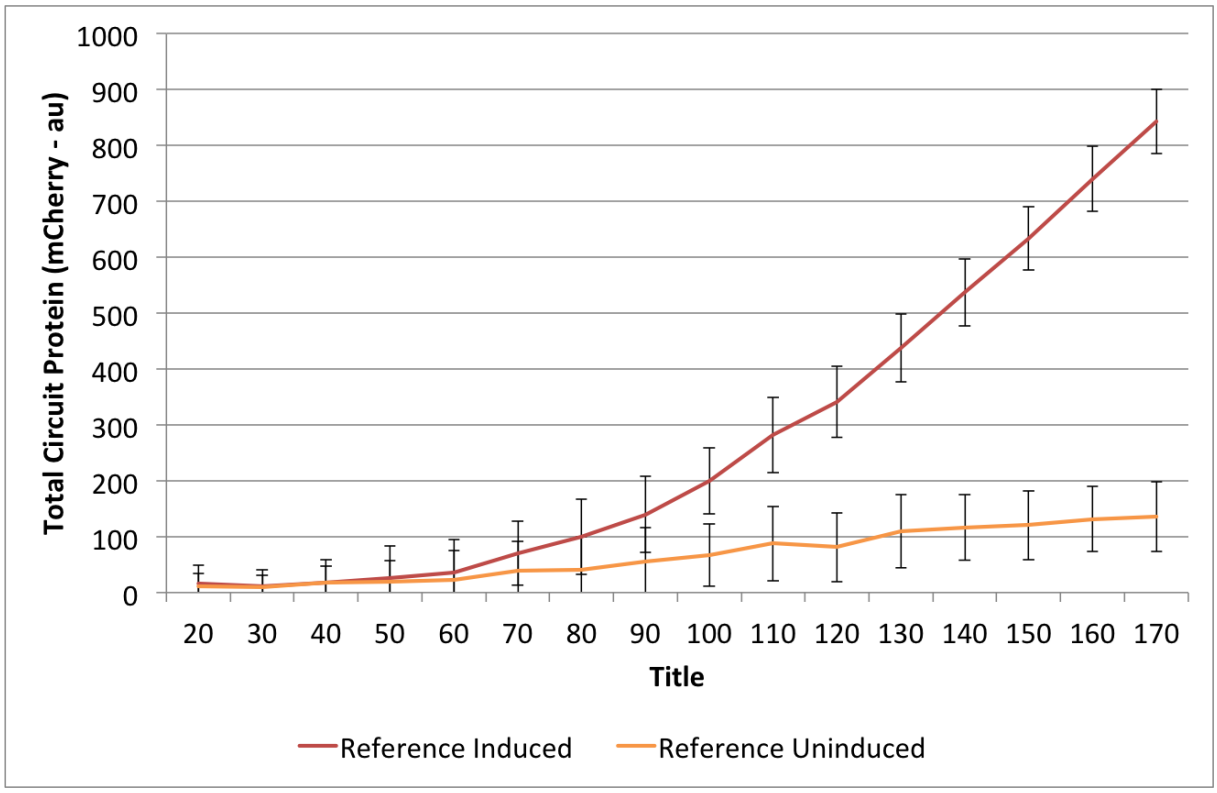
5.7.3 Circuit Output

As in the previous sections, DH10B, DH10G and both induced and uninduced DH10B cells transformed with reference construct H07 were grown over a period of 3 hours in M9 media supplemented with 0.4 % fructose. OD and mCherry fluorescence readings were taken every 10 minutes and mCherry production rates were estimated over a 1 hour moving window using Equation 3.3.

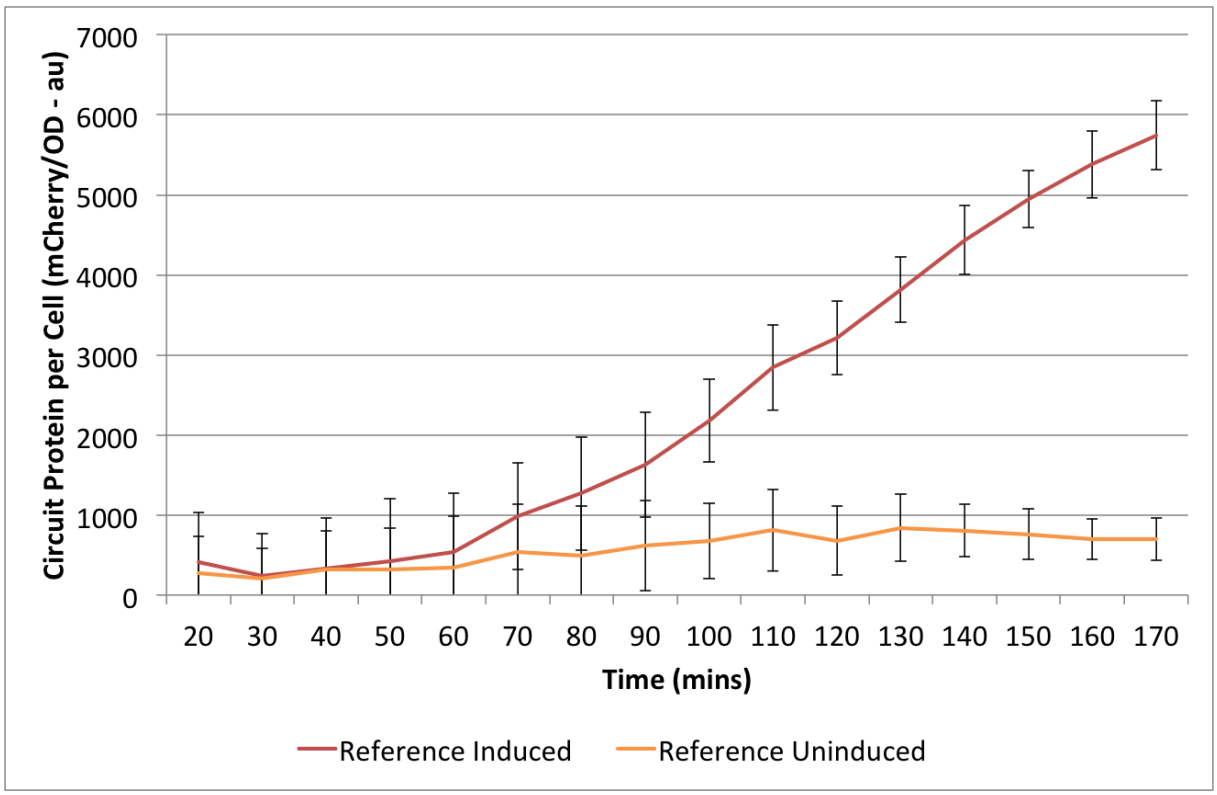
Figure 5.10a shows the total mCherry fluorescence for cells containing the reference construct, both induced and uninduced. Approximately an hour after induction the total amount of circuit protein for induced cells becomes noticeably higher than in uninduced cells. This corresponds with the time required for increased transcription and folding of mCherry to take place after induction. The fact there is a slight increase in the total amount of circuit protein over time even for uninduced cells indicates that there is leakiness from the P_{BAD} promoter, which can also be seen from the characterisation data for the promoter (see Figure 5.4).

The amount of circuit protein per cell (Figure 5.10b) follows very similar trajectories, though the rate of increase is slower due to the mCherry fluorescence being divided by an increasing OD.

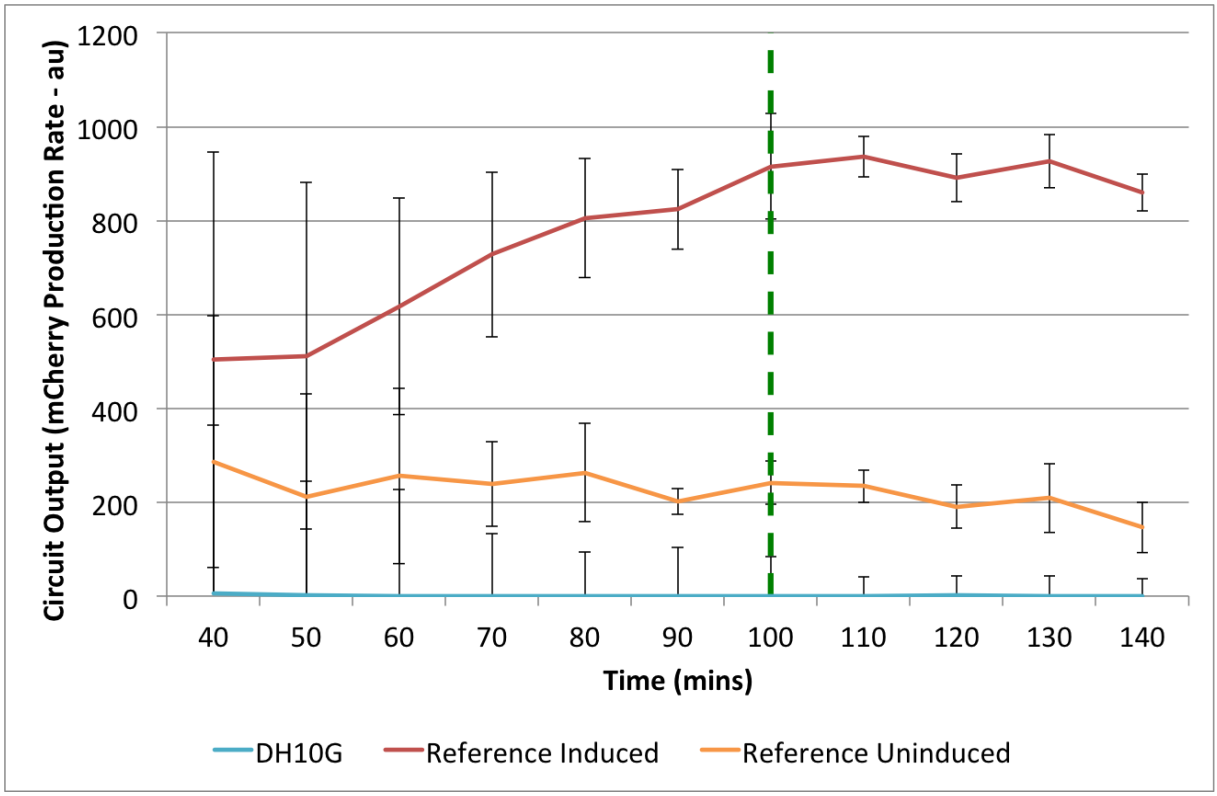
Calculated production rates for the circuit protein are shown in Figure 5.10c. This shows that, as expected, there is no mCherry fluorescence produced in the DH10G cells. The uninduced cells show a level of protein production, which corresponds with the data seen in the above figures and implies a leakiness from the P_{BAD} promoter. The circuit output (rate of production of circuit protein) for induced cells is approximately 5x higher than for uninduced cells. This suggests there is a greater level of leakage from the uninduced P_{BAD} promoter than would be predicted from the characterisation in Section 5.5.2. The reason for this is likely due to RBS-promoter interaction and is discussed further in Section 5.7.7.



(a) Total Circuit Protein



(b) Circuit Protein per Cell



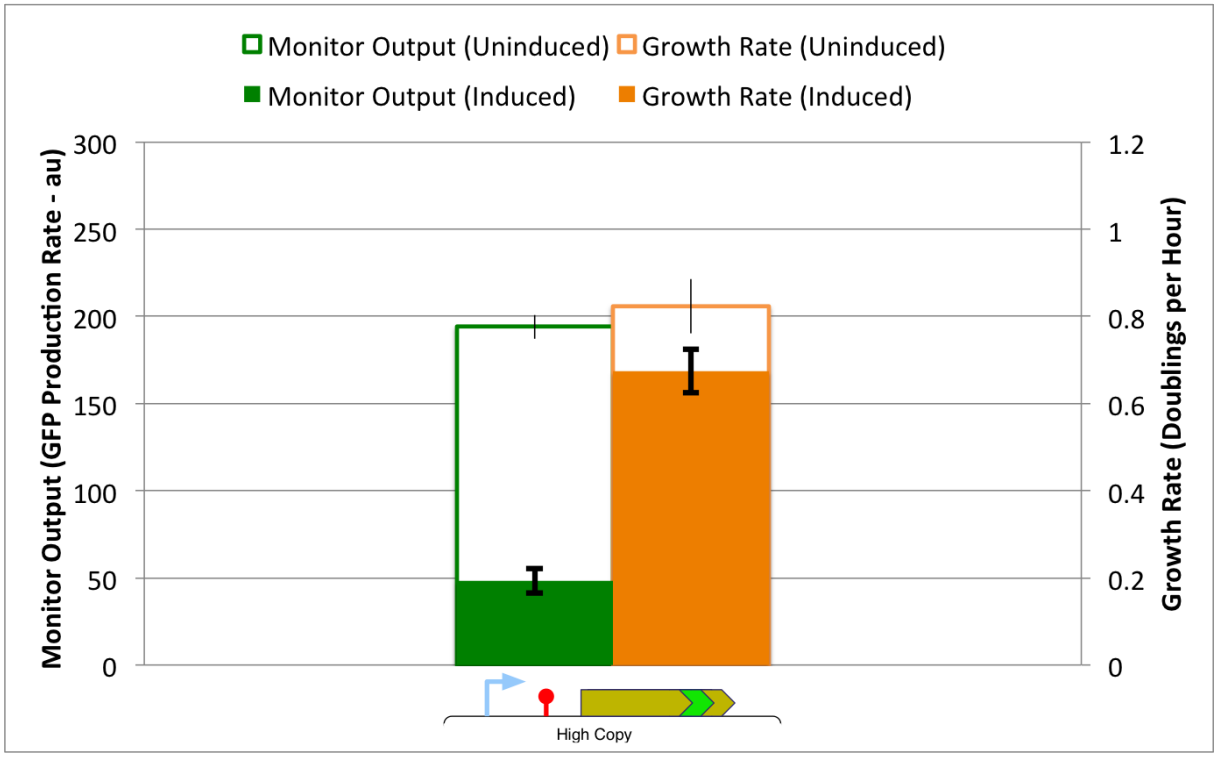
(c) Circuit Output

Figure 5.10: Circuit Protein Comparison for Reference Construct. DH10G and both induced and uninduced DH10B cells transformed with reference construct grown in M9 + 0.4% fructose for 3 hours continually in 96-well plate with 200 μl volume per well. Readings taken every 10 minutes and rates estimated over a 1 hour moving window. a) total culture mCherry fluorescence levels are indicative of the total amount of circuit protein (mCherry) in all cells b) amount of circuit protein per cell is calculated by dividing the mCherry fluorescence levels by the OD c) circuit output (mCherry production rate) calculated over 60 minute window, dashed green line indicates time of growth rate, monitor output and circuit output snapshot.

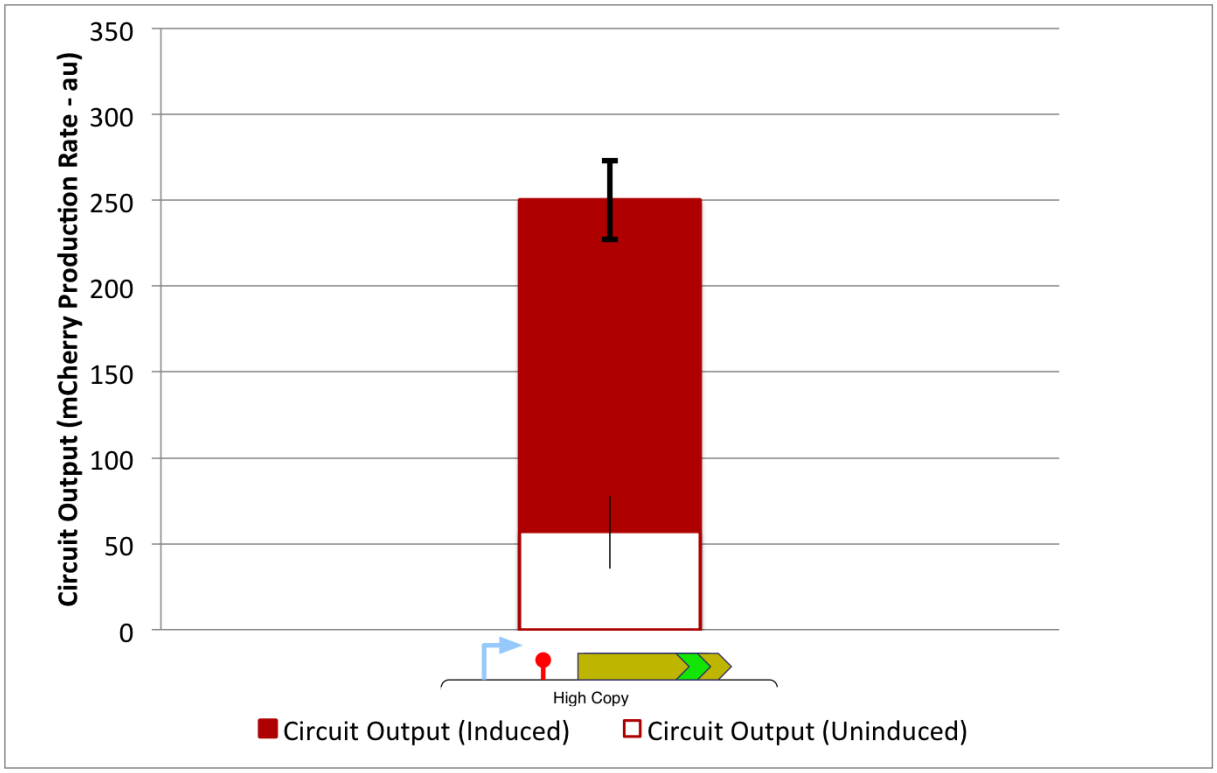
5.7.4 Key Metrics

Figure 5.11 shows the growth rates, monitor outputs and circuit outputs for DH10G cells containing the reference construct H08 (see Figure 5.6) both induced and uninduced. The data shown in these graphs corresponds to the data in Figures 5.8b, 5.9c and 5.10c at the vertical green lines. These data are calculated over a one hour window 30 minutes either side of the mid-point denoted on the x-axis (between 70 and 130 minutes centred at 100 minutes). This approach is used throughout the rest of the data shown in this results section where time-series plots are not shown but the column graphs all indicate the calculated values between 70 and 130 minutes.

In order to represent these data, we chose to visualise snapshot metrics in column graphs rather than show the entire time series as by doing this it is much easier to visualise the differences when different circuits are used. Previous characterisation has shown by the time these metrics are taken that the metrics have approximately reached a steady state as the cells grow through mid-log phase. Therefore, these metric snapshots can be said to be representative across all time points after approximately an hour of growth. The key results from this data have been discussed in more detail above. Throughout the remainder of this results section only these key metrics will be shown for the characterisation of the constructs.



(a) Growth Rate and Monitor Output



(b) Circuit Output

Figure 5.11: Growth rate, monitor output and circuit output comparison for reference construct at 100 mins. Taken as snapshot during growth in M9 media supplemented with 0.4% fructosea) OD levels b) growth rate calculated over 60 minute window, dashed green line indicates time of growth rate, monitor output and circuit output snapshot.

5.7.5 Identifying Causes of Burden in Plasmid Based System

It is highly informative to understand where in the system we have designed most of the burden on the shared resources is coming from. In order to do this, a set of different plasmids were constructed and tested. These plasmids were:

- Empty pSB1C3 backbone
- pSB1C3 containing AraBAD promoter unit
- pSB1C3 containing full reference construct

Figure 5.12 shows that DH10B (no monitor) and DH10G (DH10B with monitor) cells have very similar growth rates (as also shown in Section **??**). pSB1C3, pSB1C3 with AraBAD and the reference construct all have very similar monitor output and growth rate. The monitor output rates are approximately 20% lower than DH10G cells and the growth rates are approximately 6-12% lower. The fact these constructs all have very similar behaviours shows that this decrease is due to the presence of the backbone.

The reference construct plasmid contains twice as much DNA as the empty pSB1C3 plasmid and since they both impose very similar burden levels and have similar decreases in growth rate relative to DH10G, it can be argued that the extra DNA replication machinery required does not impose an extra burden on the cell. The difference in monitor output and growth rate when the pSB1C3 plasmid is inserted into the cell is either due to the production of the Chloramphenicol resistance protein or a factor of having chloramphenicol in the media. Residual chloramphenicol that has not been degraded by the chloramphenicol acetyltransferase protein expressed by the plasmid may be inhibiting some ribosomal activity. This is unlikey however, due to the dosage of chloramphenicol used and the fact pSB1C3 is a high copy plasmid. Therefore it is most likely that this burden placed on cells by the presence of the plasmid is caused by the expression of the resistance marker protein and origin of replication machinery.

5.7.6 Comparison of Promoter Strengths

Figure 5.13 shows the differences between two constructs with weak promoters and strong promoters. These two constructs are identical except for the promoter used and they are high-copy (pSB1C3 backbone) with fast codons and medium strength RBS. The weak promoter

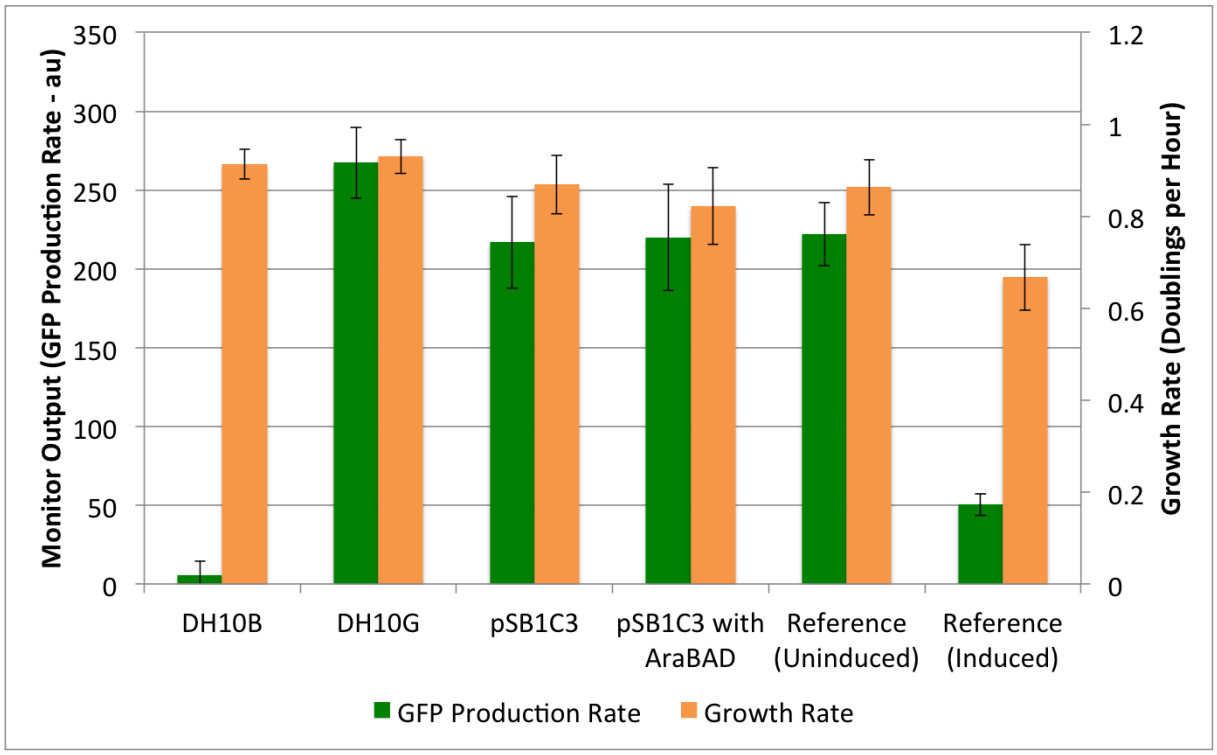


Figure 5.12: Growth rate and monitor output of reference construct parts, showing empty (no monitor) DH10B cells, DH10G cells, DH10G cells containing the pSB1C3 backbone, DH10G cells containing pSB1C3 backbone with AraBAD promoter unit, reference construct uninduced and reference construct induced.

version is the reference construct H08 whereas the strong promoter is construct H11.

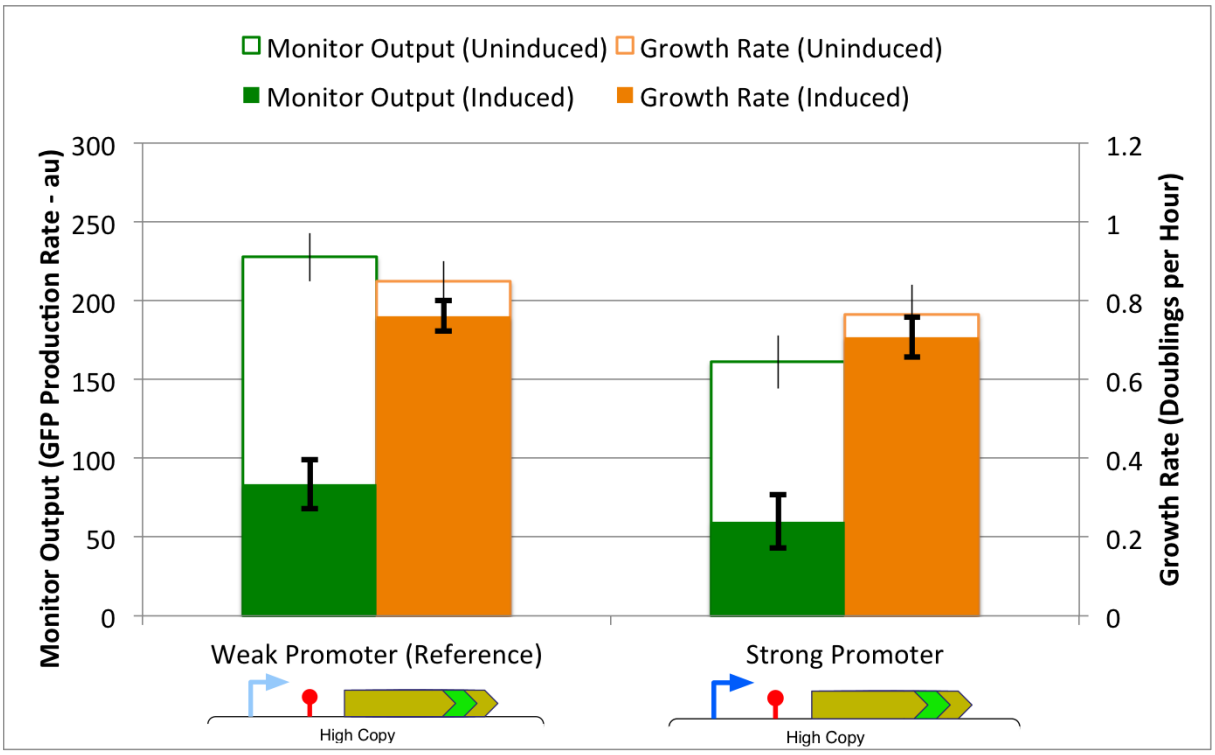
It can be clearly seen that when the circuits are uninduced the amount of monitor output is less for the strong promoter than the weak promoter (as shown by comparing the columns with green outline and white fill). This indicates that an extra burden is being placed on the shared resources by this construct even when it is not induced. Looking at the amount of protein being produced by the two circuits when uninduced it can be seen that there is a larger amount of leakage from the strong promoter, causing more protein to be produced, which corresponds with the decrease in monitor output for this construct (as shown by the columns with red outline and white fill). The growth rates of the two constructs are similar and within the range of standard deviations of each other (see columns with orange outline and white fill). This is a clear example of where our monitor is able to outperform simple growth rate measurements in terms of being able to detect burden on shared resources.

When the circuits are induced the strong promoter construct produces approximately 2.4x as much circuit protein than when uninduced (see solid red columns). This causes a slight decrease in growth rate (see solid orange columns), though standard deviations overlap, and

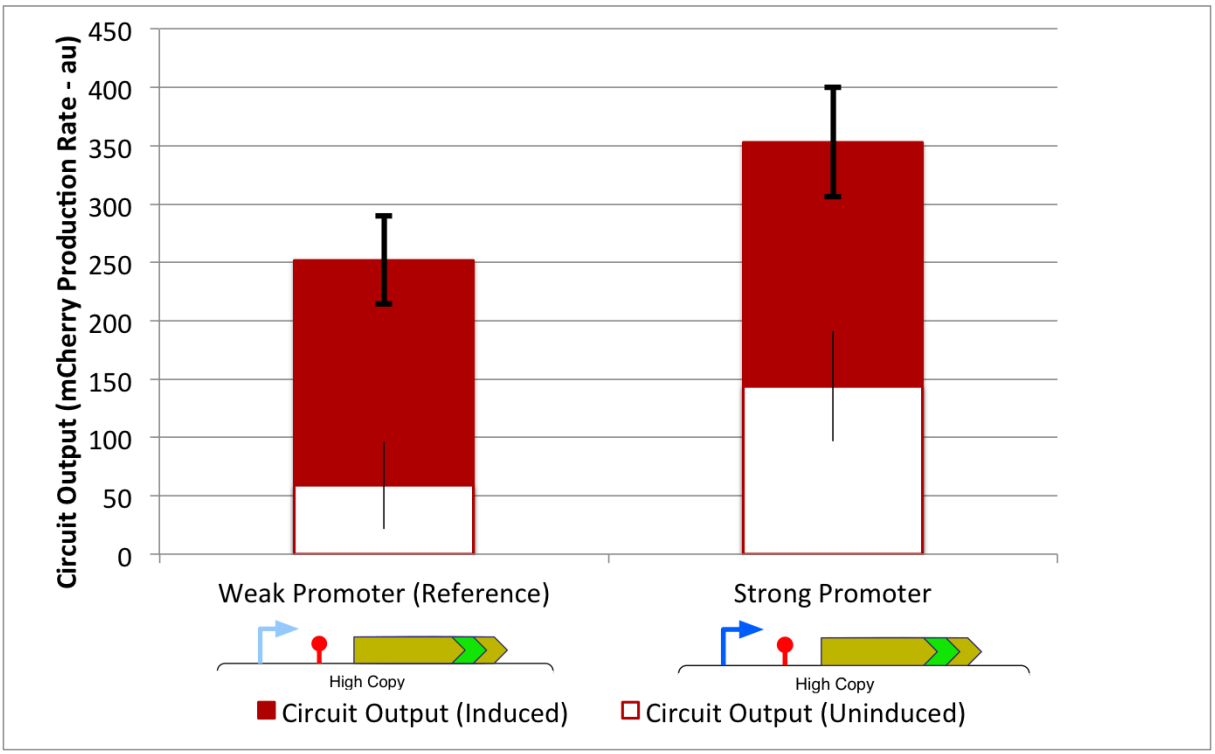
a larger (approximately 70%) decrease in the amount of monitor protein produced (see solid green columns). When compared to the induced weak promoter construct, there is an approximate 40% increase in circuit output. The monitor output is approximately 25% less for the strong promoter construct. When looking at the exact numbers it can be calculated that the circuit output ratio between the two promoters is approximately the reciprocal of the monitor outputs.

Circuit Output Ratio (weak promoter : strong promoter) = 1 : 1.40

Monitor Output Ratio (weak promoter : strong promoter) = 1 : 0.72 $\approx$ 1/1.40



(a) Growth Rate and Monitor Output

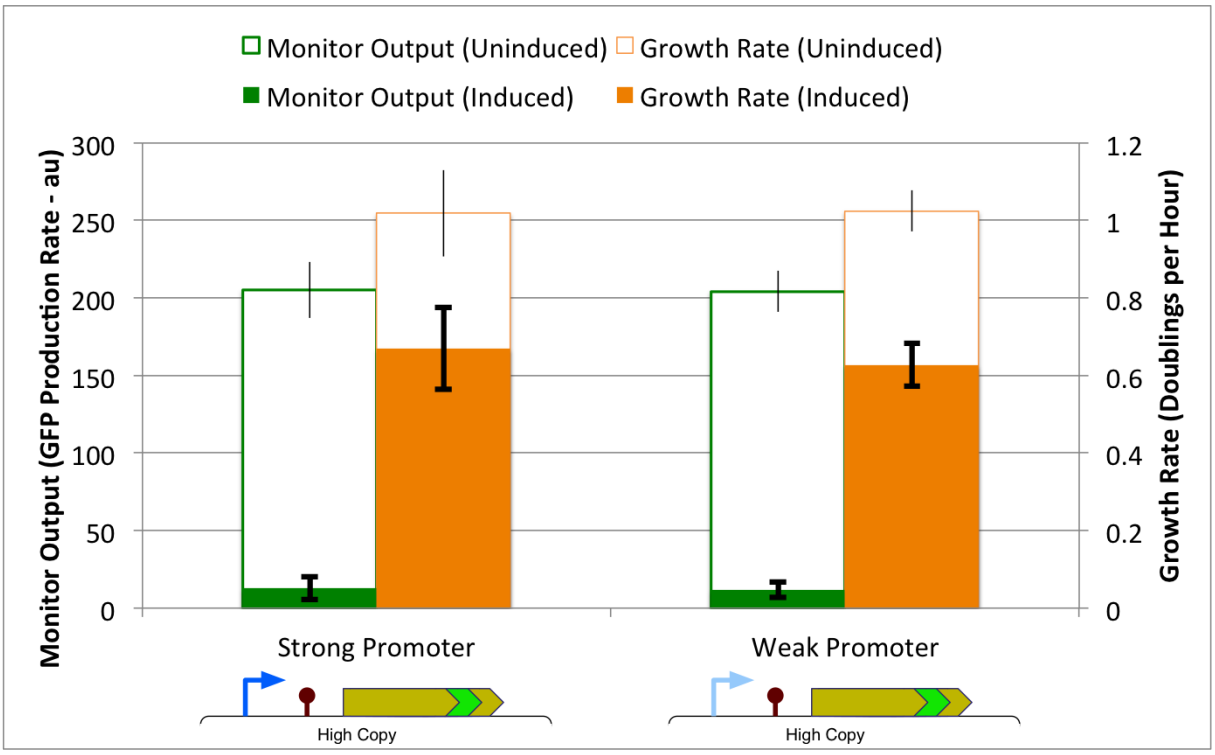


(b) Circuit Output

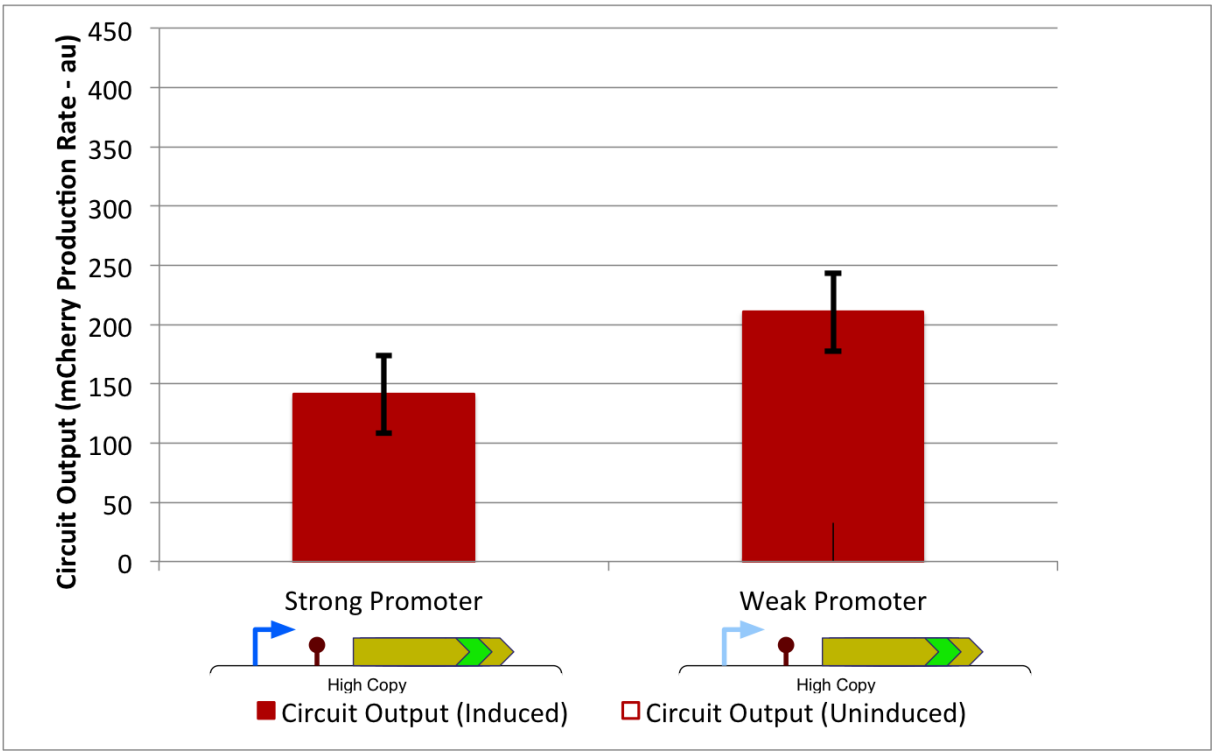
Figure 5.13: Promoter strength comparison for reference construct and strong promoter variant. *Strong Promoter* construct is H11 and *Weak Promoter* construct is reference construct H08. a) Growth rate and monitor output for both constructs induced and uninduced. b) Circuit output for both constructs induced and uninduced.

Figure **??** shows data for different constructs to those seen above in Figure 5.13. These constructs also have high-copy backbones and fast codons but have high strength RBS sequences (constructs H07 and H10 for weak and strong promoter respectively). These versions show no leakage when the promoters are uninduced, which may be due to interactions between the RBS sequence and the promoter region. Both of the uninduced constructs grow at exactly the same rate and have the same monitor output which corresponds to neither producing any protein through leakage.

When the constructs are induced, the strong promoter construct produces protein at a rate that is approximately 50% higher. The rate of protein production is lower, however, than for the constructs where the RBS strength is lower (see Figure 5.13). This characteristic of diminished output from very strong RBS sequences is discussed further in Section **??**. The Growth rates for the two constructs are very similar at an approximate decrease of 40%. The monitor outputs are both very similar at a decrease of protein production of approximately 95%. This represents a very catastrophic impact on the shared resource pool for both promoter strengths with the strong RBS. This implies that the difference in the amount of transcripts (which we expect to be a two-fold difference from promoter characterisation - see Section **??**) does not cause a large difference in the total amount of ribosomes recruited onto the circuit transcripts, but does affect the amount of protein produced.



(a) Growth Rate and Monitor Output



(b) Circuit Output

Figure 5.14: Promoter strength comparison for strong RBS constructs. *Strong Promoter* construct is H10 and *Weak Promoter* construct is H07. a) Growth rate and monitor output for both constructs induced and uninduced. b) Circuit output for both constructs induced and uninduced.

5.7.7 Comparison of RBS Strengths

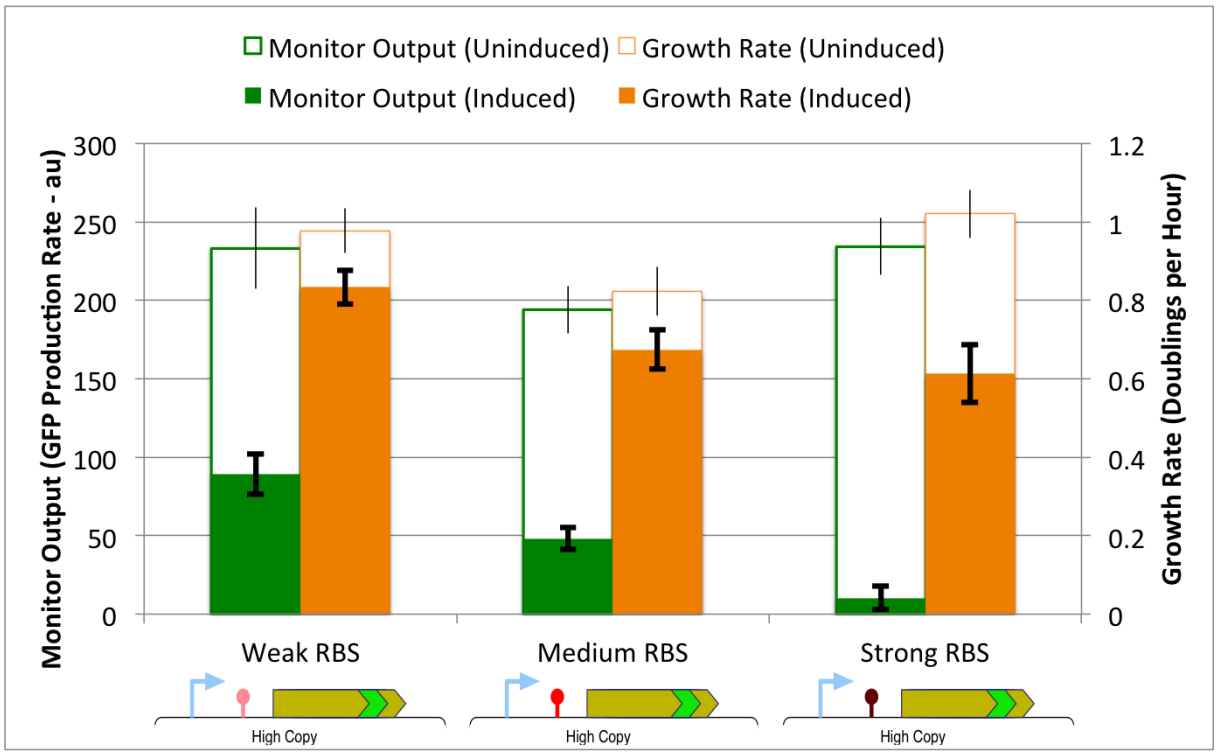
The relationship between RBS strength, the rate at which a protein expresses and the amount of cellular resources used to do so is a complex one. Figure 5.15 shows the key metrics for three different constructs with varying RBS strength. These constructs are all on high-copy plasmids and have the weak promoter version and fast codon coding sequence. The strong, medium and weak RBS constructs are constructs H07, H08 and H09 correspondingly.

When the constructs are not induced there is a 15-20% decrease in the growth rate for the cells with the medium RBS construct as well as a similar decrease in the amount of output from the capacity monitor relative to the cells with either the weak RBS construct or the strong RBS construct. Cells containing the medium RBS construct are producing protein from the circuit even when uninduced, whereas cells containing the other constructs are not. This fits with data seen in previous graphs where there is leakage from circuits with the medium RBS but not from circuits with other RBS versions (compare Figures 5.13b and 5.14b for examples). When characterising the two versions of the P_{BAD} promoter (Figure 5.4), there was very little leakage from either promoter. This is most likely due to an interaction between the medium strength RBS sequence and the promoter region or polymerases which cause there to be this leakage from the promoter.

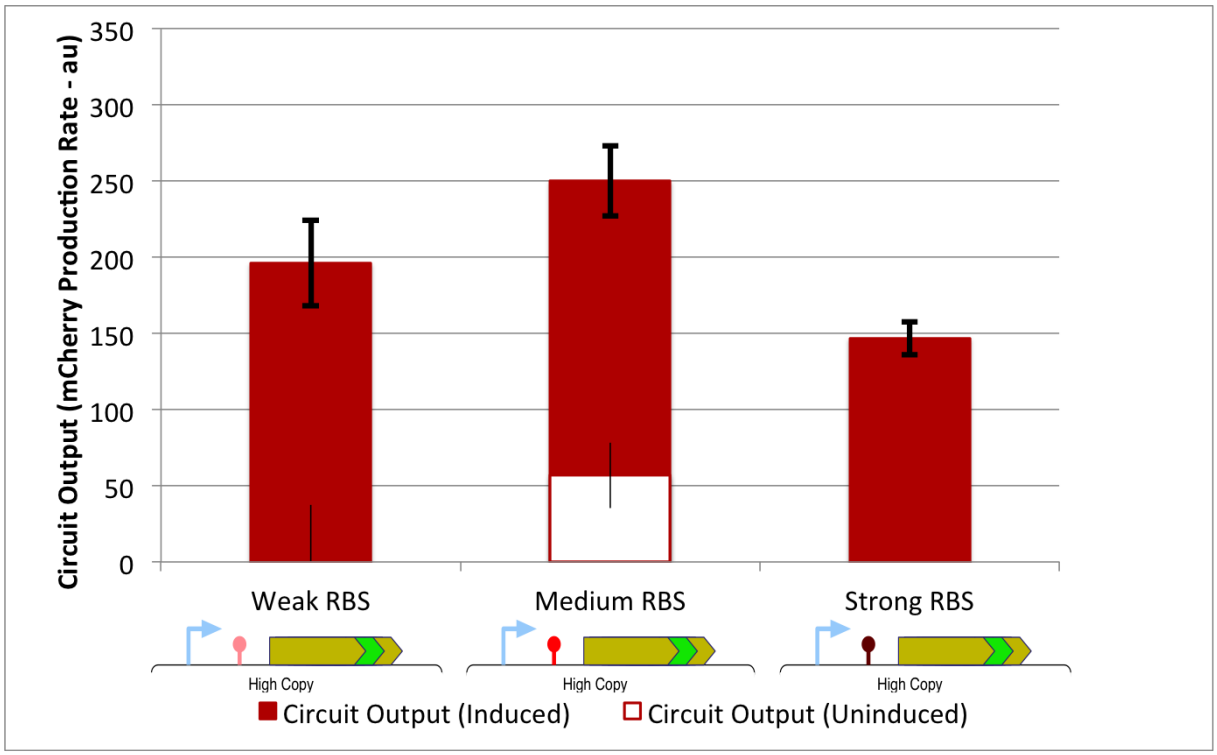
Induction causes all circuits to start producing protein. As would be predicted from only knowing the relative strengths of the RBS regions, the weak RBS construct causes less protein production than the construct with the medium strength RBS. This corresponds to a decrease in both the growth rate and monitor output of approximately 15% and 60% respectively for the weak RBS construct and 15% and 75% respectively for the medium RBS construct. This is due to the same number of transcripts (from the same promoter induced at the same level) loading ribosomes and initiating translation at different rates, with a higher rate of initiation for the medium RBS and thus more ribosomes being sequestered from the free ribosome pool onto transcripts and a decrease in ribosomal availability, which is then reflected in the rate of protein production from the capacity monitor.

The strong RBS construct causes the largest decrease in monitor output and growth rate when induced (approximately 95% and 40% respectively). This does not, however, correspond to a circuit production rate that is higher than the weak and medium strength RBS constructs. This

is most likely due to a catastrophic decrease in ribosomes leading to cell death for a portion of the population (which would explain a simultaneous decrease in growth rate, monitor output and circuit output across the population).



(a) Growth Rate and Monitor Output



(b) Circuit Output

Figure 5.15: RBS comparison for reference construct and strong promoter variant. *Weak RBS* construct is H09, *Medium RBS* construct is the reference construct H08 and *Strong RBS* construct is H07. a) Growth rate and monitor output for all constructs induced and uninduced. b) Circuit output for all constructs induced and uninduced.

5.7.8 Comparison of Copy Numbers

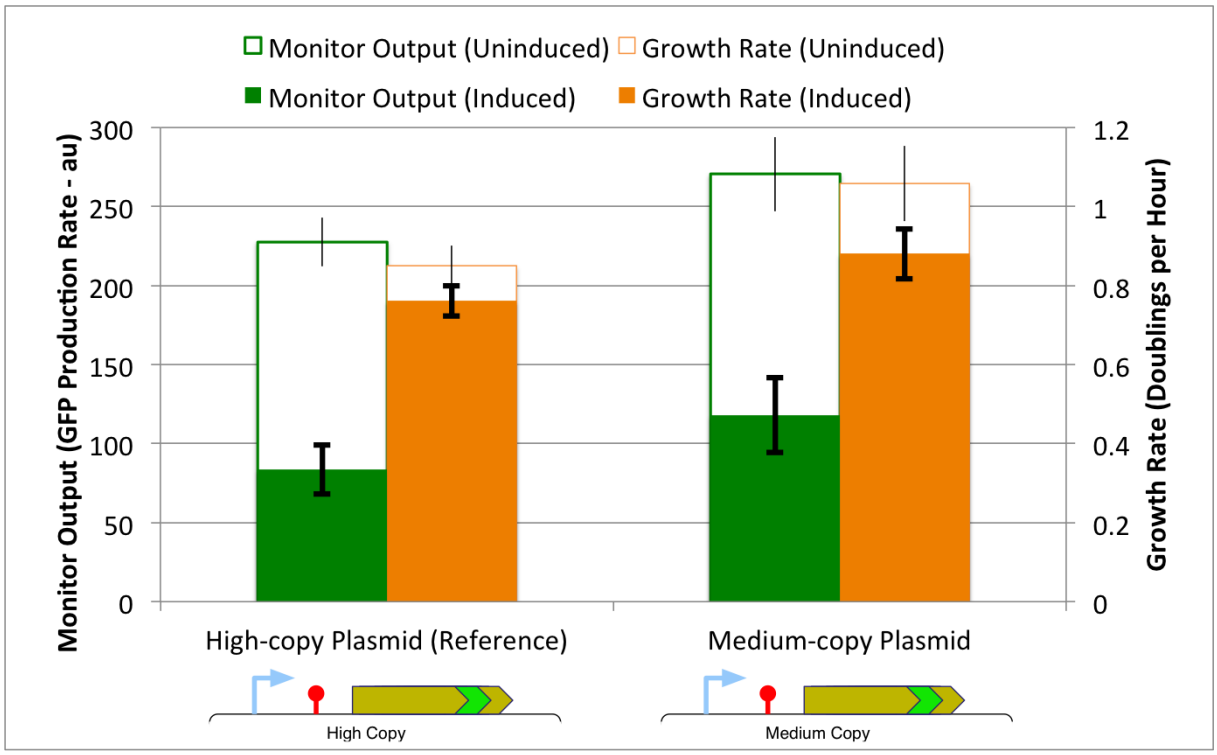
Changing the copy number is achieved by changing the plasmid backbone into which constructs are placed. In this comparison we have focussed on the difference between circuits placed in high- and medium-copy backbones since we expect there to be an order of magnitude difference in the copy number of these plasmids (100-300 and 10-12 copies per cell for high- and medium-copy plasmids respectively). Figure 5.14 shows the key metrics for the reference construct in high copy pSB1C3 backbone compared to the same construct in a medium copy backbone. Changing the plasmid backbone does not only change the amount of circuit DNA present in the cell, it also impacts on the number of copies of the antibiotic marker as well as the type and amount of any regulatory mechanisms required for the origin of replication. The constructs tested in this comparison are the reference construct H08 and the equivalent construct in a medium copy plasmid M08.

Both constructs being tested contain the medium strength RBS and therefore show leakage (see Section 5.7.7 for more details) when uninduced. When uninduced, the medium copy plasmid shows a higher growth rate. This is likely due to a number of factors, including the lower amount of DNA replication resources required to maintain a lower copy number as well as the lower amount of chloramphenicol resistance protein produced from a lower copy number backbone. These differences are also reflected in the monitor output, where the higher copy number plasmid produces GFP at a rate that is 15-20% less than the medium copy.

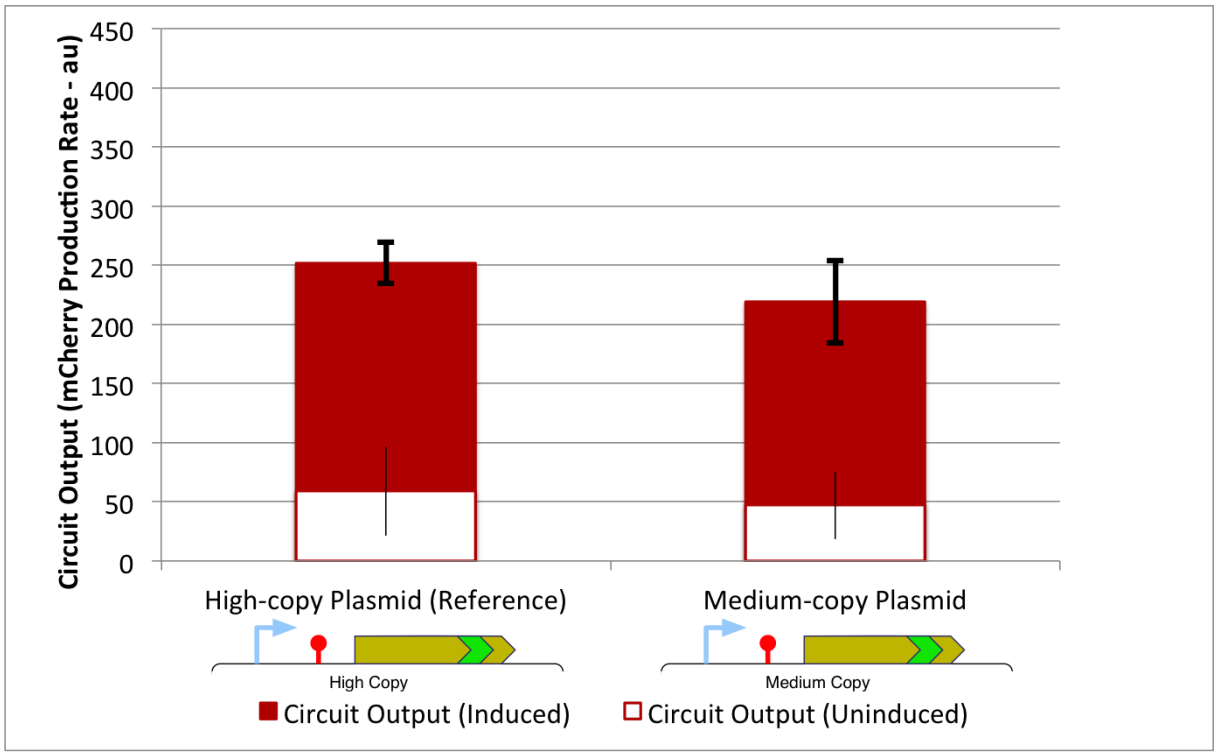
When induced, both constructs show a very similar proportional increase in circuit output relative to when uninduced (approximately 4.3 fold and 4.7 fold for high and medium copy respectively). The rate of protein production for the medium copy construct is approximately 10% lower than that from the high copy and the standard deviations of the two overlap significantly. This could be due to there being less competition for resources from elements in the plasmid backbone, such as the origin of replication, chloramphenicol resistance marker and constitutive AraC from the AraBAD promoter unit. This shows that a lower copy plasmid is able to use sufficiently little of the cells resources for replication and maintenance (RNA and protein for origin and resistance marker) that there is enough capacity for the main circuit to generate protein at a rate that is comparable to that of a much higher copy plasmid.

When induced, the growth rate of the medium copy plasmid drops by a larger amount than

the high copy construct (16% compared to 10%) though the absolute growth rate remains higher. The drop in monitor output is greater for the high copy constructs with a decrease of 63% compared to 56%. Whilst these figures show that the lower copy plasmid has a higher growth rate, higher monitor output and lower circuit output, the differences are relatively small compared to those seen when altering other control points and do not correspond to expected values for an order of magnitude difference in copy number as reported.



(a) Growth Rate and Monitor Output



(b) Circuit Output

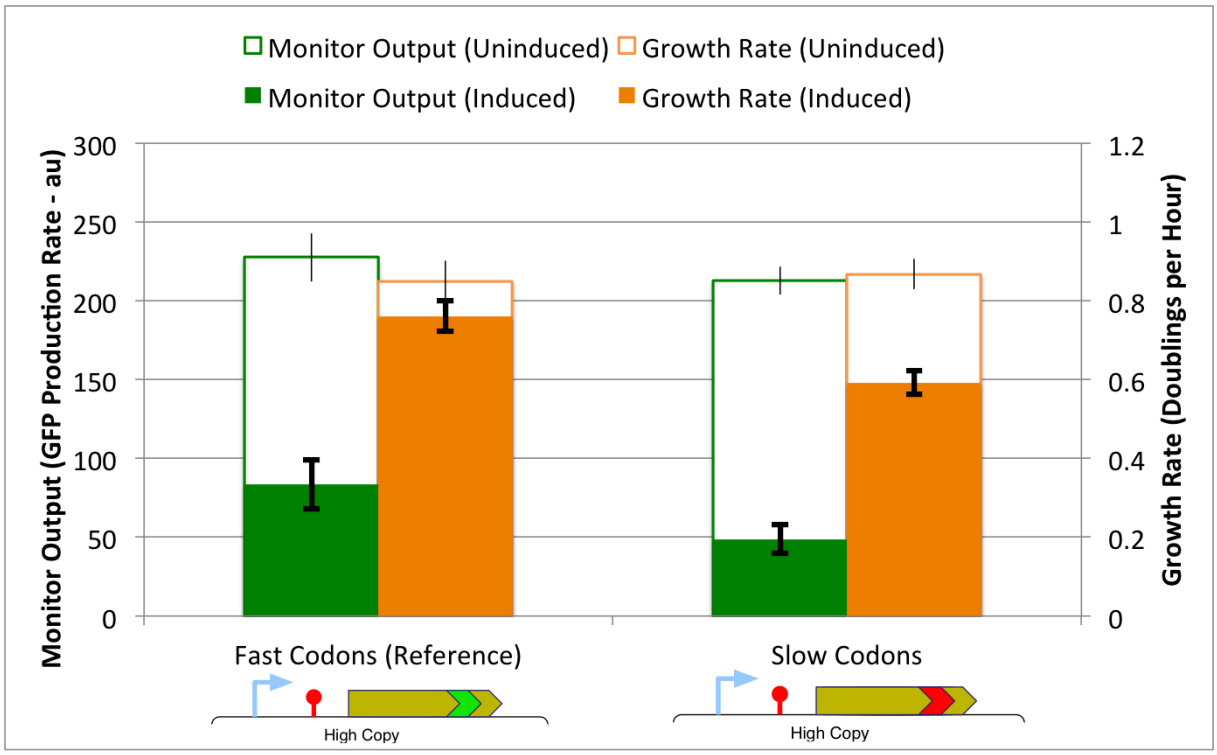
Figure 5.16: Copy number comparison for reference construct and medium copy variant. *High copy* construct is H08, *Medium copy* construct is the construct M08. a) Growth rate and monitor output for all constructs induced and uninduced. b) Circuit output for all constructs induced and uninduced.

5.7.9 Comparison of Codon Usage

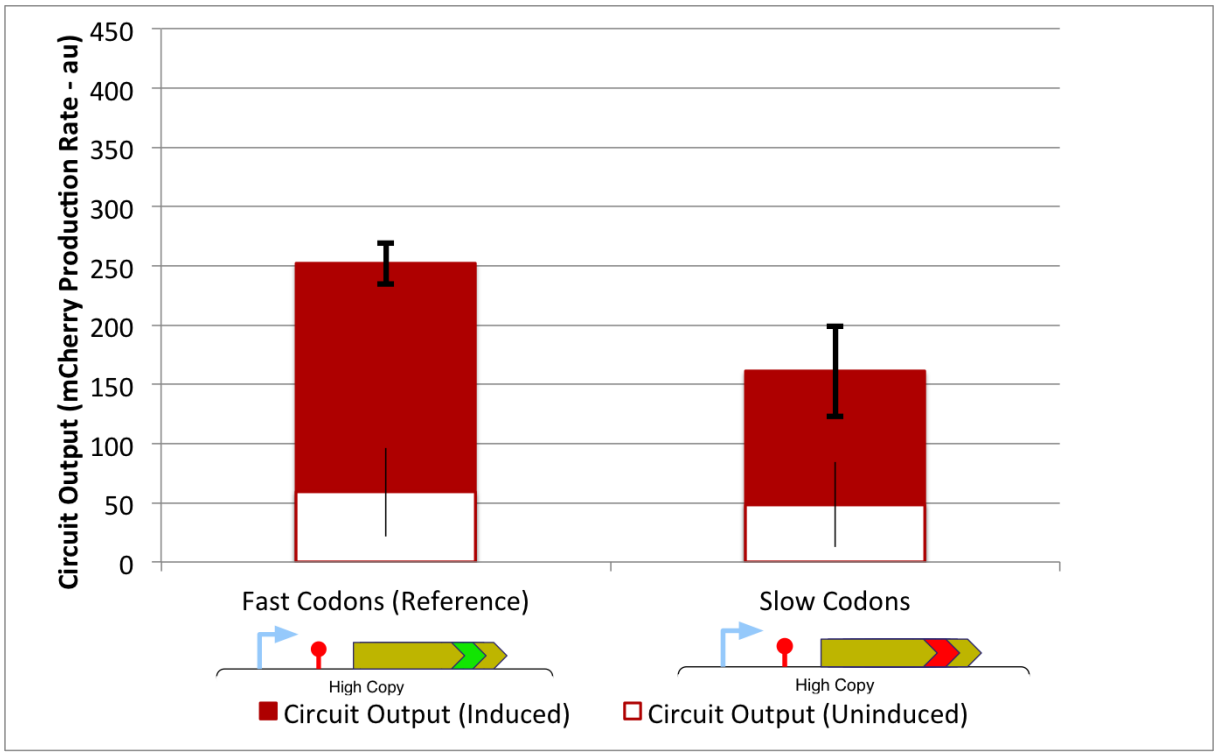
Figures 5.17 and 5.18 show the impact of introducing slow codons and anti Shine-Dalgarno sequence into the end of VioB, before the mCherry fusion. Figure 5.17 shows the key metrics for the reference construct H08 as well as another construct that is identical apart from having slow codons in the coding region (H02). It can be seen that when the circuits are uninduced the growth rates are very similar, as are the monitor outputs (though the slow codon version has slightly lower monitor output). The leaking output from the circuit is also very similar for both constructs.

When the circuits are induced a much larger difference is observed. It can be clearly seen that the slow codon construct performs 'worse' across all metrics. The amount of protein produced by the circuit is 36% less for the slow codon construct than the fast codon construct. In addition, the growth rate and monitor output are lower for the slow codon construct by 22% and 42% respectively. This is likely due to the slow codons causing ribosomes to move less quickly along the transcript, meaning that the flux of ribosomes along the mRNA is lower, and the circuit output is lower. The slow codons also mean that the amount of time each ribosome that is recruited onto the transcript spends in elongation is longer and so for an equivalent rate of translational initiation (i.e. same RBS strength) more ribosomes will be used up at any point in time, thus meaning more ribosomes are sequestered from the free pool and the monitor output decreases.

When the RBS strength is increased and the medium RBS is replaced with the strong RBS, the fast codon version shows a very large decrease in monitor output, as discussed in Section 5.7.7. Ass seen in Figure 5.18, the cells containing the slow codon construct show approximately zero output and very low decreases in monitor output and growth rate (8% and 4% respectively) with overlapping error bars. This behaviour is very similar to what would be expected from cells that do not have a circuit present. This is most likely due to the circuit having mutated so that the circuit element of the plasmid has been removed or no longer produces protein. This adaptation shows that high RBS strengths combined with slow codons can cause levels of burden that render cells unviable, meaning they evolve to cope with this.

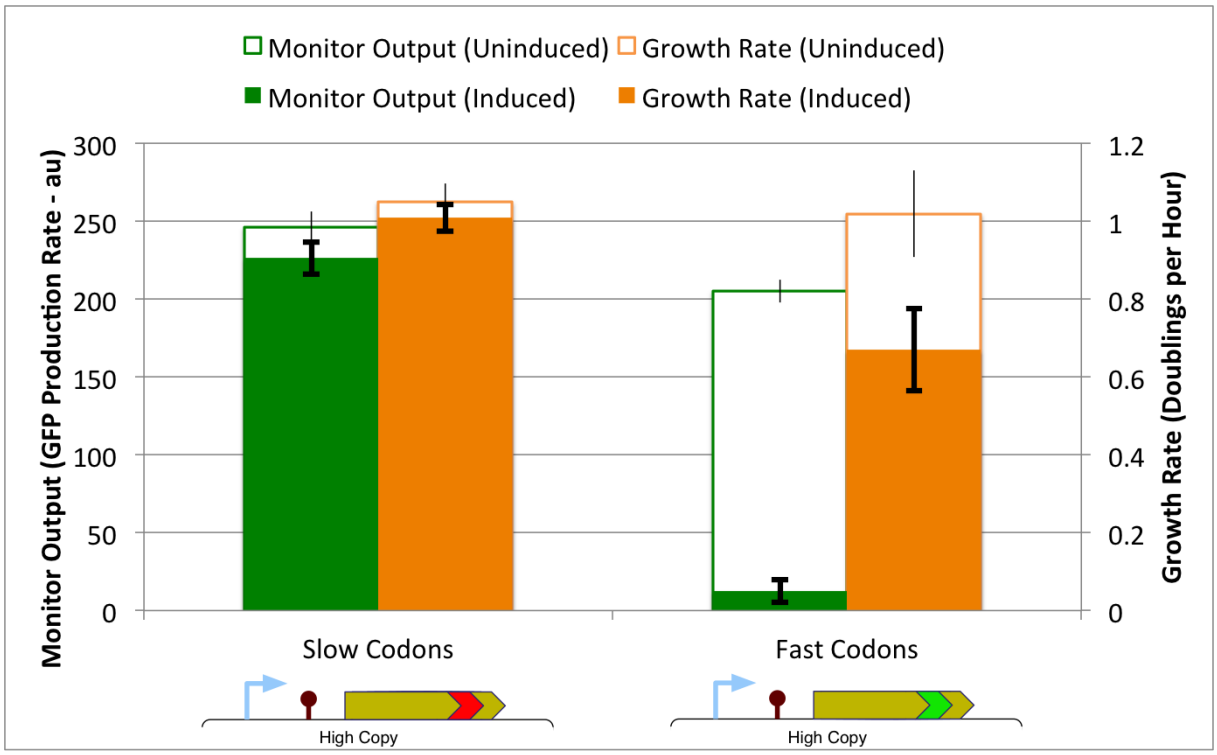


(a) Growth Rate and Monitor Output

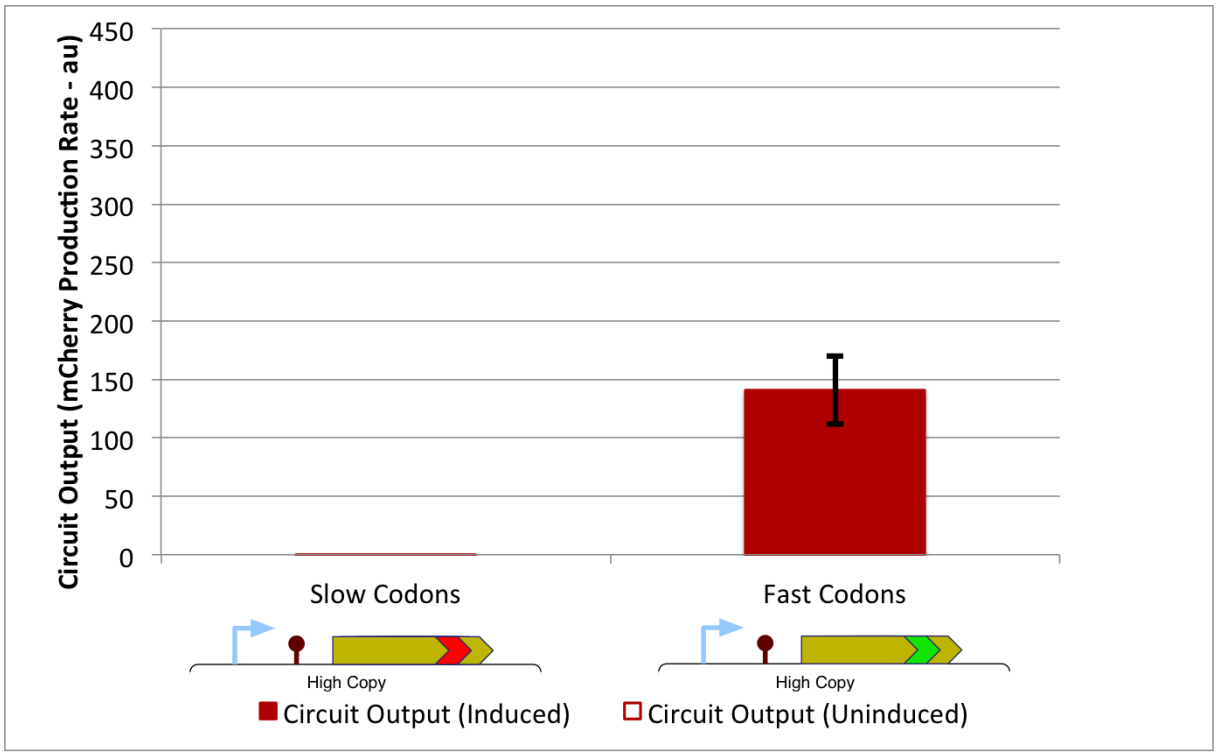


(b) Circuit Output

Figure 5.17: Codon usage comparison for reference construct and slow codon variant. *Fast codon* construct is reference H08, *Slow codon* construct is H02. a) Growth rate and monitor output for all constructs induced and uninduced. b) Circuit output for all constructs induced and uninduced.



(a) Growth Rate and Monitor Output



(b) Circuit Output

Figure 5.18: Codon usage comparison for strong RBS constructs. *Fast codon* construct is H07, *Slow codon* construct is H01. a) Growth rate and monitor output for all constructs induced and uninduced. b) Circuit output for all constructs induced and uninduced.

5.8 Obtaining Similar Circuit Output with Different Burden Levels

It is important to not only investigate what the impact of changes in single controls points are on the cell, but to also look at how changing different control points together can provide improvements in the system. Figure 5.19 shows how the 4 combination of 2 promoters and 2 RBSes allows a range of outputs and burdens from a synthetic circuit. It can be clearly seen that strong promoters cause a higher circuit output than the weak promoter variants with the same RBS, as well as a lower monitor output (higher burden on resources). Similarly strong RBS constructs give higher circuit output and lower monitor output than the weak RBS constructs wight he same promoter.

An interesting result is that the construct with weak promoter and strong RBS has approximately the same level of circuit output as the construct with strong promoter and weak RBS. While the outputs are very similar for these constructs, the level of monitor output is approximately double for the construct with the weak RBS and strong promoter. This suggests that the weak RBS/strong promoter construct uses less resources per protein produced than the strong RBS/weak promoter construct. The weak RBS/strong promoter combination is more *efficient* - a term we explore further below.

The reason for this difference is most likely related to the efficiency of transcripts in terms of the ratio between the average rate of protein production per transcript and the number of ribosomes per transcript:

$$\text{Efficiency} = \frac{\text{Rate of circuit protein production (proteins per second)}}{\text{Ribosomes used (average number of ribosomes across transcripts)}} \tag{5.1}$$

In terms of measurable quantities that we can use to estimate this efficiency we argue that the rate of protein production can be used to represent the circuit output rate and the number of ribosomes used can be approximated as the reciprocal of the capacity as represented by the rate of capacity monitor output. In this case, what we define as efficiency is proportional to both the circuit output as well as the capacity remaining in the cell. Using this method we get an equation of the form:

$$\text{efficiency} = CM_c \tag{5.2}$$

where C is the circuit protein production rate and M_c is the monitor output of cells that have been transformed with the circuit. Section 5.7.7 shows the relationship between RBS strength and efficiency of a circuit as defined above. Here we see that increasing RBS strength from very low levels causes an increase in this efficiency up until a maximal point after which the efficiency decreased until it reaches a plateau. The situation observed here indicates that the weak RBS corresponds to a higher efficiency than the strong RBS (though lower total output) and the stronger promoter means there is a larger number of transcripts which in sum give the same rate of protein production as the strong RBS/weak promoter construct. This is done with more efficient transcripts meaning in total less ribosomes are being used to produce the same level of protein output.

This result is very interesting as it shows that using principles outlined in this project we can improve the way circuits are designed to provide the same output whilst decreasing resource usage, thus reducing 'waste'.

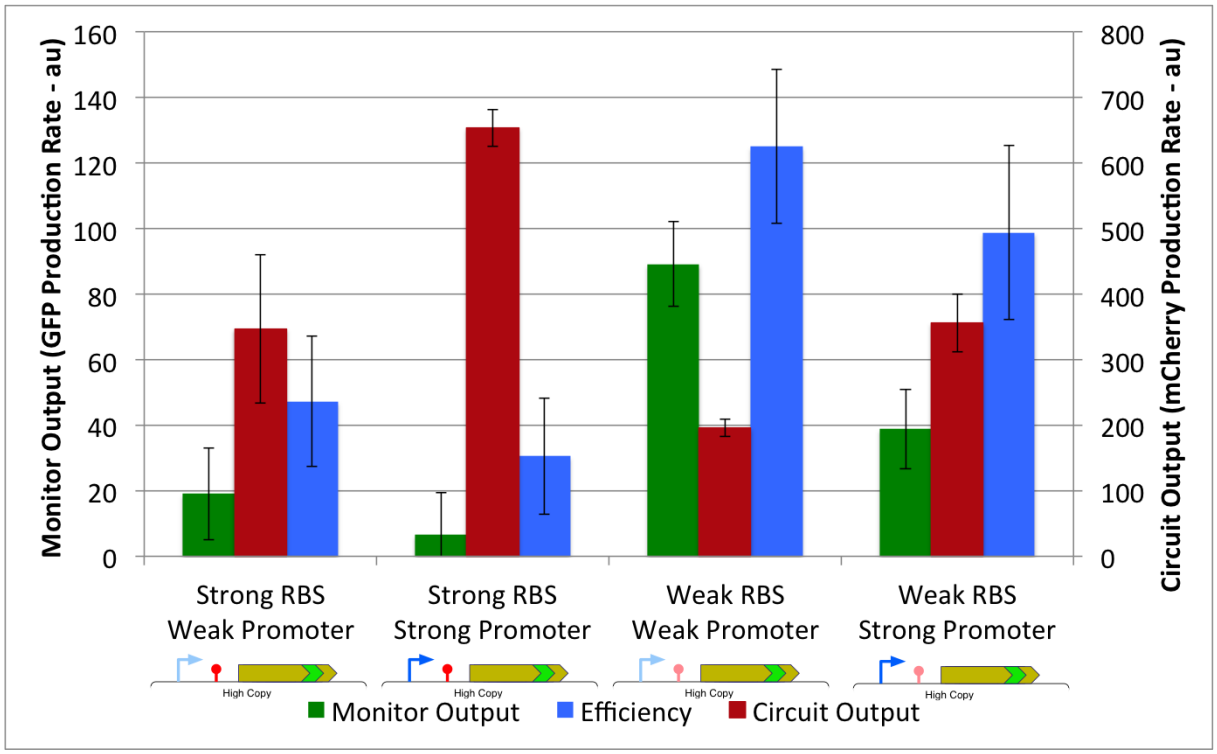
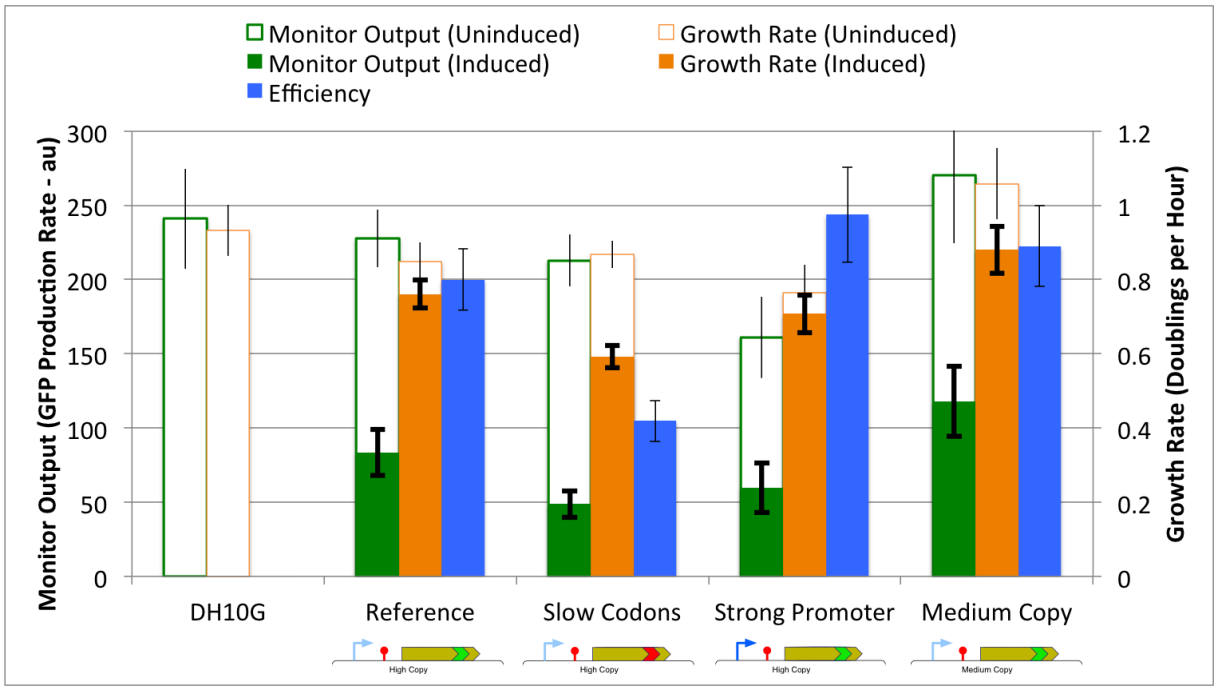


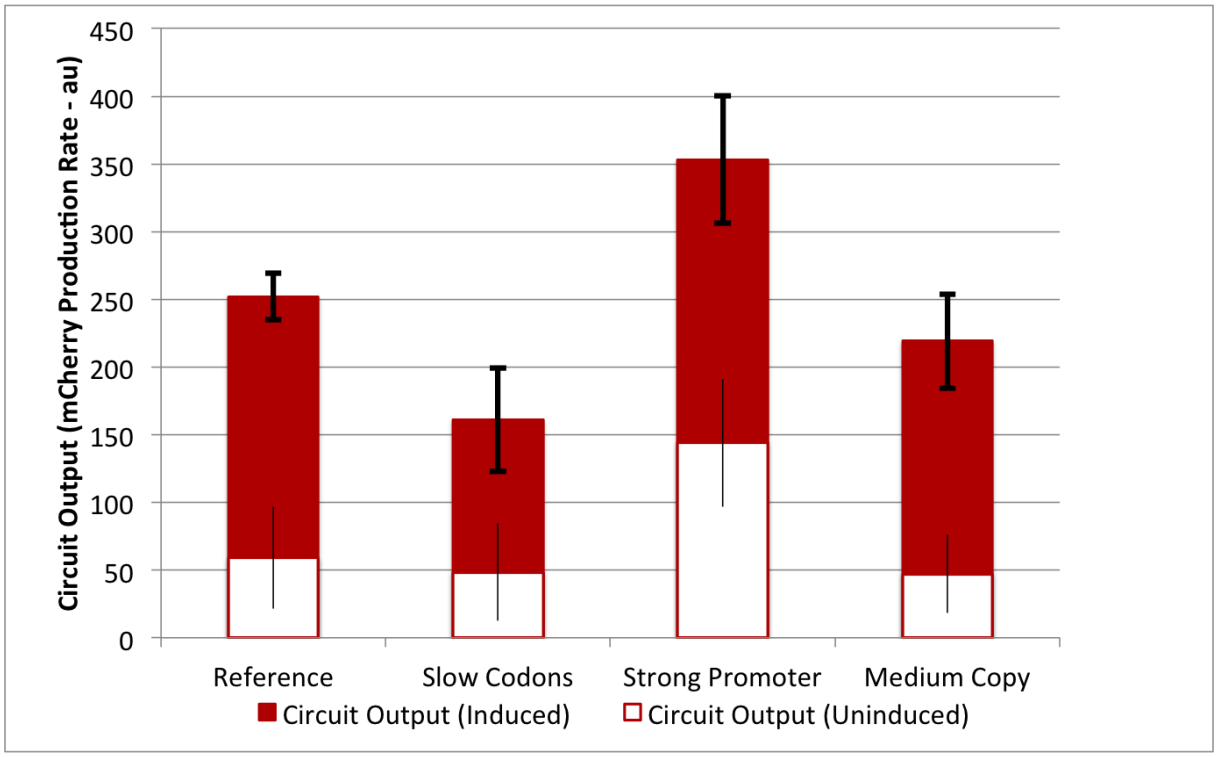
Figure 5.19: Obtaining similar circuit output with different burden levels. This figure shows circuit output and monitor output for all four combinations of medium and weak RBS with strong and weak promoter.

5.8.1 Overview

Figure 5.21 shows a comparison between a number of constructs that have been shown above in Figures 5.13, 5.16 and 5.17. This figure given an overview of the impact of changing different control points relative to the reference construct. Slow codons give no benefit in terms of growth rate, monitor output or circuit output and in the context of this project it appears that it is always beneficial to use a codon optimised coding region without slow codons or anti shine-delgano sequences. The stronger promoter causes a slight decrease in both monitor output and growth rate, but also causes a 40% increase in monitor output, therefore being a good design strategy if circuit output is the most important consideration. Using a lower copy plasmid gives the best growth rate and monitor output, both of which are higher than the empty DH10G cells.



(a) Growth Rate and Monitor Output



(b) Circuit Output

Figure 5.20: Overview of key metrics for reference construct and similar constructs shows how changing the codon usage, promoter strength and copy number affect a) monitor output and growth rate and b) circuit output.

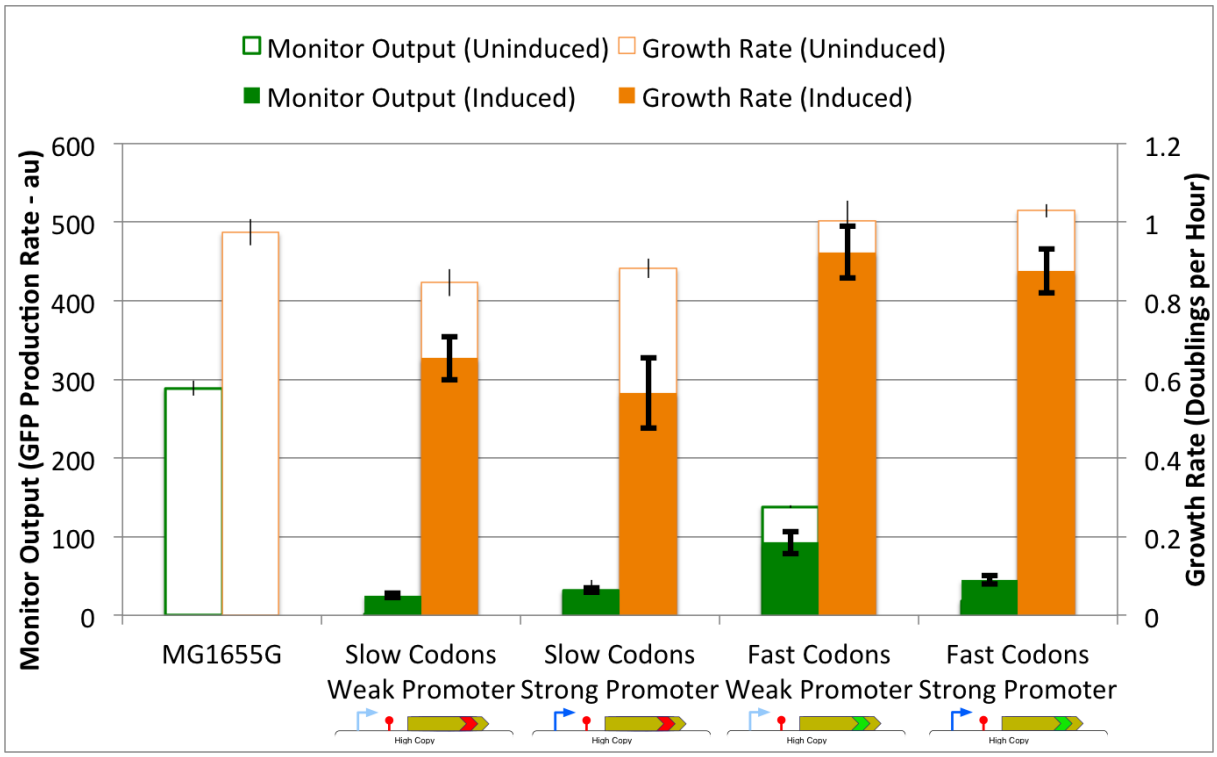
5.9 MG1655 - Impact of the Stringent Response

So far all of the circuits tested have been characterised in DH10B cells. As mentioned previously, these cells do not have the stringent response phenotype (see Section 1.1.5 for more details) and therefore would be expected to behave differently when experiencing 'burden'. We performed the same testing protocol used above for DH10B cells on MG1655 cells that had a capacity monitor inserted into the genome. The circuits tested were all versions with high-copy backbone and medium strength RBS.

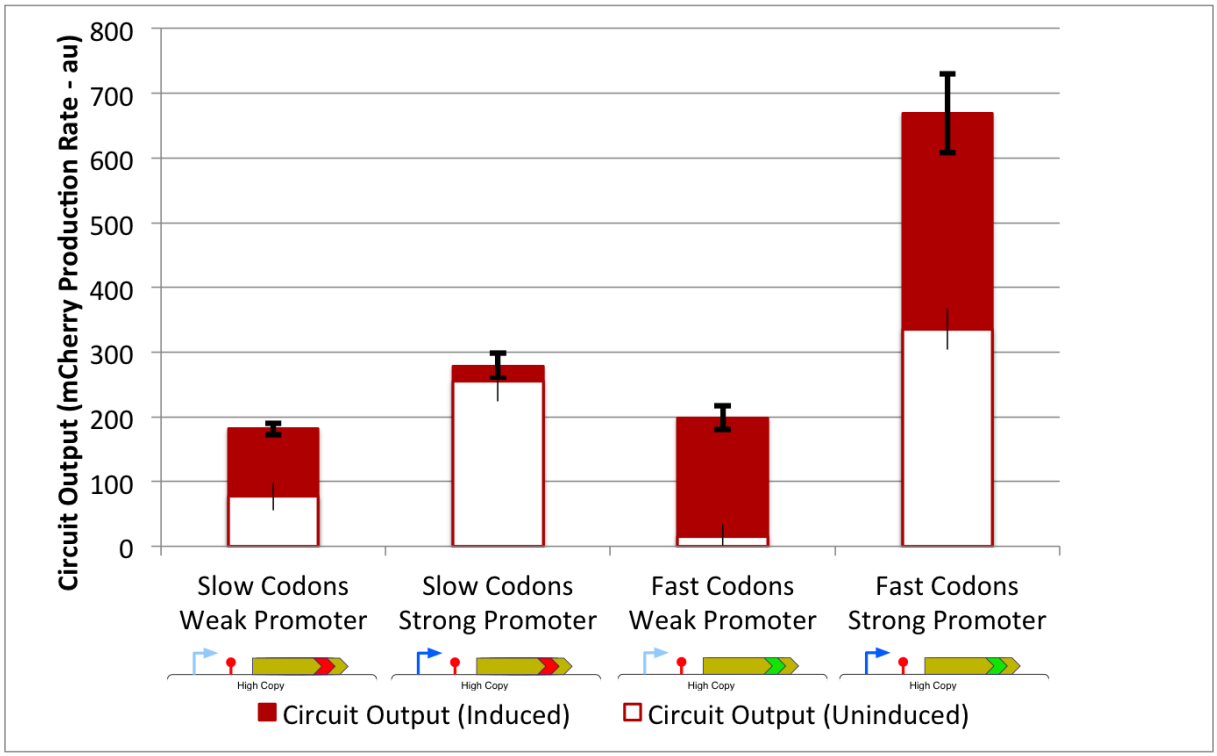
Figure 5.21a shows the monitor output and growth rate for untransformed MG1655 cells as well as ones transformed with the constructs detailed above. It can be clearly seen that when uninduced the cells with fast codons have no drop in growth rate, whereas cells containing the slow codon constructs have a decrease of approximately 5%. However, we see a large decrease in the rate of monitor output with cells containing the weak promoter/fast codons circuit decreasing by approximately 50% and all others decreasing by approximately 90%. When we consider the level of leakage of the circuit, as shown in Figure 5.21b, we see that the weak promoter/fast codons circuit has very little leakage, whereas the others have a noticeable level of leakage.

When looking at the induced circuits, the rate of circuit output for the fast codon constructs as higher than for the corresponding slow codon circuits, which matches what we have observed in DH10B cells. In addition, the growth rate of cells containing slow codon circuits is also lower compared to those containing fast codon circuits. For the strong promoter circuits the fast codon circuit has an output which is is an approximately 2.3x higher than the slow codon circuit, whereas both the fast and slow codon versions have similar rates of circuit output for the weak promoter.

These results indicate that MG1655 cells heavily down regulate the production of unnecessary protein when heterologous protein production is detected though the usage of shared resources. However, when compared to DH10B cells, the level of circuit output is similar. This indicates that the stringent response may be more effective at down regulating the production of protein from genomic DNA.



(a) Growth Rate and Monitor Output



(b) Circuit Output

Figure 5.21: Snapshot of key metrics for MG1655G cells (MG1655 with capacity monitor) transformed with all high copy/medium RBS constructs. Grown for 3 hours in M9 media supplemented with 0.4% fructose in 200 µl volumes in 96-well plate. Snapshot taken at 100 minutes after growth started. a) growth rate and monitor output, b) circuit output.

5.10 Impact on RNA Levels

In order to separate the impact of expressing a synthetic DNA on the shared resource pool into transcription and translation resources, RNA levels were quantified and compared to protein levels. Total RNA levels per cell were calculated and are shown against growth rates in Figure 5.22 and are shown to be very proportional with an R-squared value of 92.7% when fitted with a linear regression passing through the origin. This proportional relationship between cellular RNA levels and growth rate is well known and was previously reported in[?].

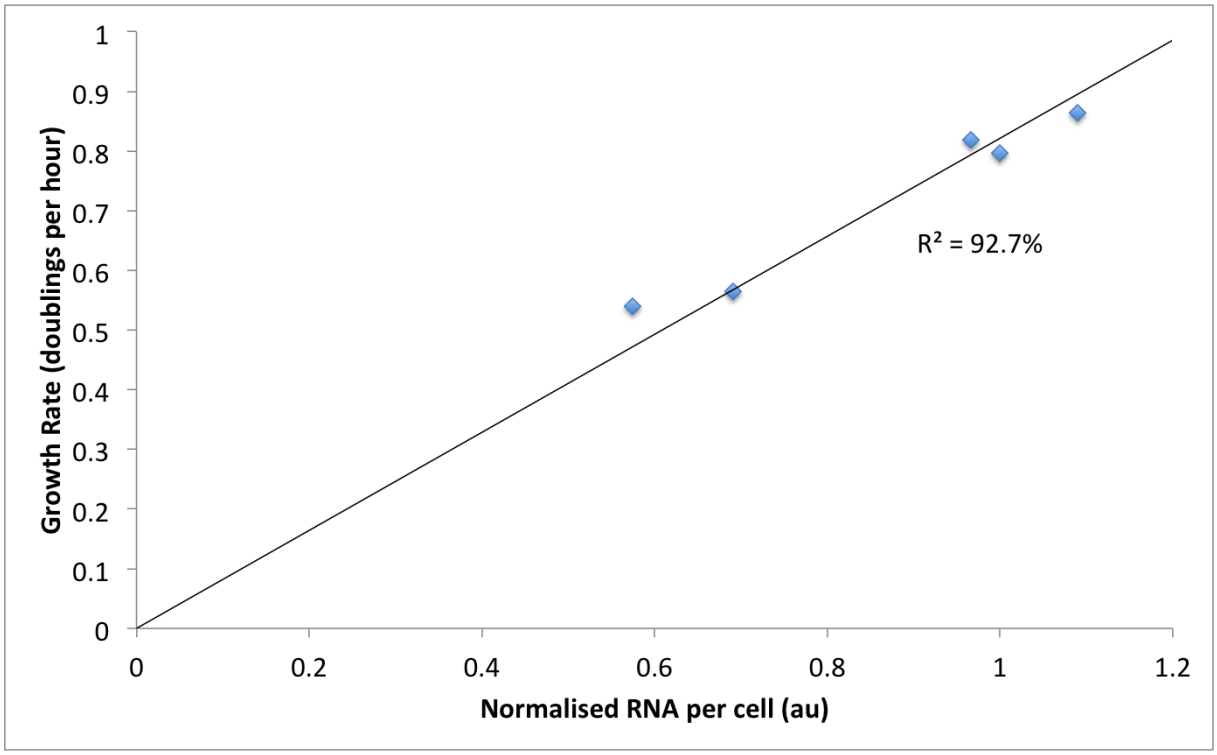


Figure 5.22: Scatter diagram of cellular RNA levels (normalised against DH10G cells), as estimated by nano-drop of cellular RNA extraction, against growth rate for DH10G cells transformed with reference construct H08 and strong promoter variant H11. RNA levels and growth rate are taken after 3 hour growth in 96-well plate in 200 μ M9 media supplemented with 0.4% fructose.

Relative cellular amounts of the capacity monitor mRNA were measured using qPCR relative to the *gapA* housekeeping gene. A number of housekeeping genes were tested in an as-yet unreported study and *gapA* was shown to be a stable housekeeping gene when looking at cells with different amount of burden placed on the shared resources pool. The proportion of monitor mRNA per cell was multiplied by the estimated total RNA per cell to estimate the absolute amount of monitor mRNA per cell. These figures were then compared to protein production rates to give an estimate of the translational rates within the cell.

Figure 5.23 shows the monitor output and calculated translation rates within the reference construct H08 and the equivalent construct with a strong promoter (H11). These are shown for these constructs both induced and uninduced. This data clearly shows that the changes in monitor output are highly correlated with the rate of transcription and therefore that the main bottleneck in resources that causes the 'burden' measured by our monitor is on the translational resources. These findings reflect what has been reported elsewhere in the literature CITE.

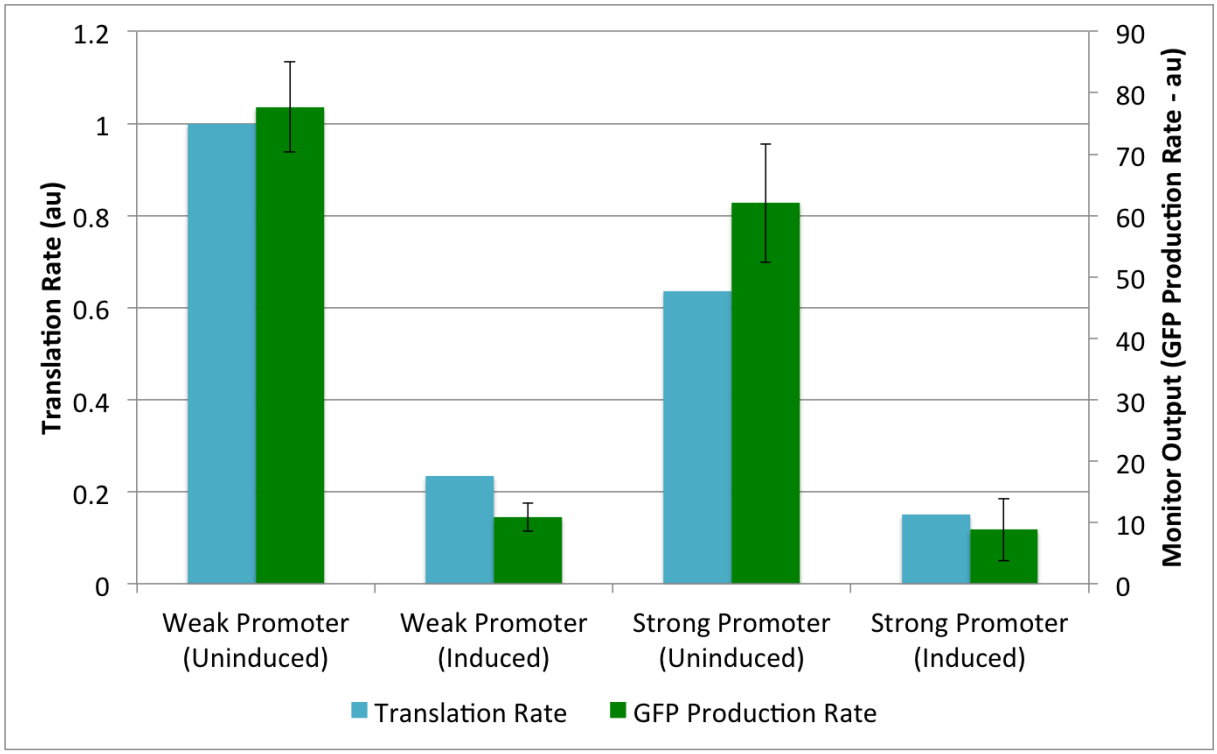


Figure 5.23: Translation rates and monitor outputs shown for the reference construct (H08) and the equivalent construct with the strong promoter version (H11). This clearly shows that changes in monitor output are highly correlated to the translation rate within cells.

5.11 Relationship Between Growth Rate and Other Metrics

A key consideration in this project is understanding if and how the capacity monitor that has been implemented is superior to the growth rate as an indicator of the burden placed on shared resources by a synthetic circuit. The growth rate of cells is a complex function of many factors such as proteome, growth media, temperature etc. Figure 5.24 shows the relationship between growth rate and monitor output for all of the constructs shown so far in this section as well other data so that all high-copy constructs are shown as well as medium copy constructs with medium strength RBS. All of the data shown are for induced constructs. The red line shows a

linear regression for this data with an x-intercept at approximately 0.4. It is possible to have cell growth without any cellular capacity, this simply means that all cells resources are diverted into growth processes and there are none available for monitor protein production. The regression has an R-squared value of 0.23. This indicates that there is a weak correlation between growth rate and monitor output, and there is a lot of unexplained variance between the two variables.

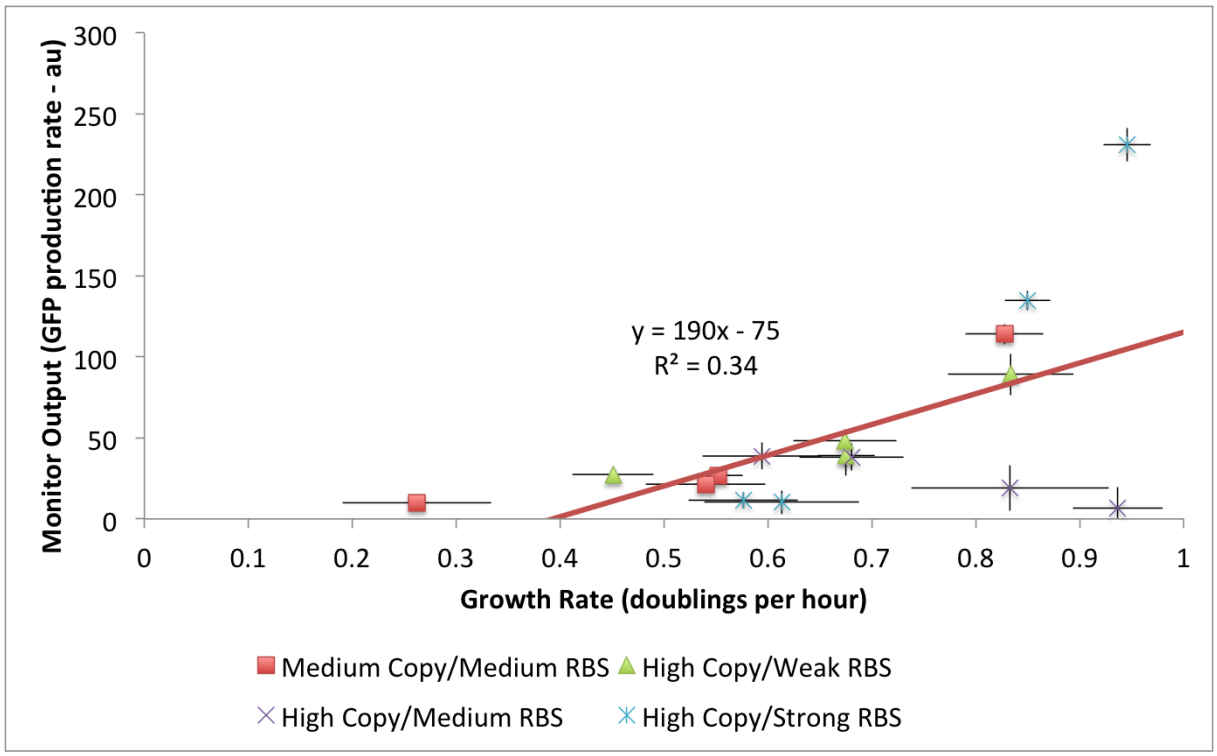


Figure 5.24: Scatter plot showing the relationship between growth rate and monitor output for all high copy constructs and medium copy constructs with medium RBS. Red line indicates a linear regression with R-squared value and equation of line shown. Error bars indicate standard deviations over 6 repeats.

We see in Figure 5.25 that there is very little correlation between the circuit output and growth rate where there is an R-squared values for the linear regression of 0.08. Figure 5.26 shows that growth rate and circuit efficiency also have very little correlation with an R-squared linear regression value of 0.09.

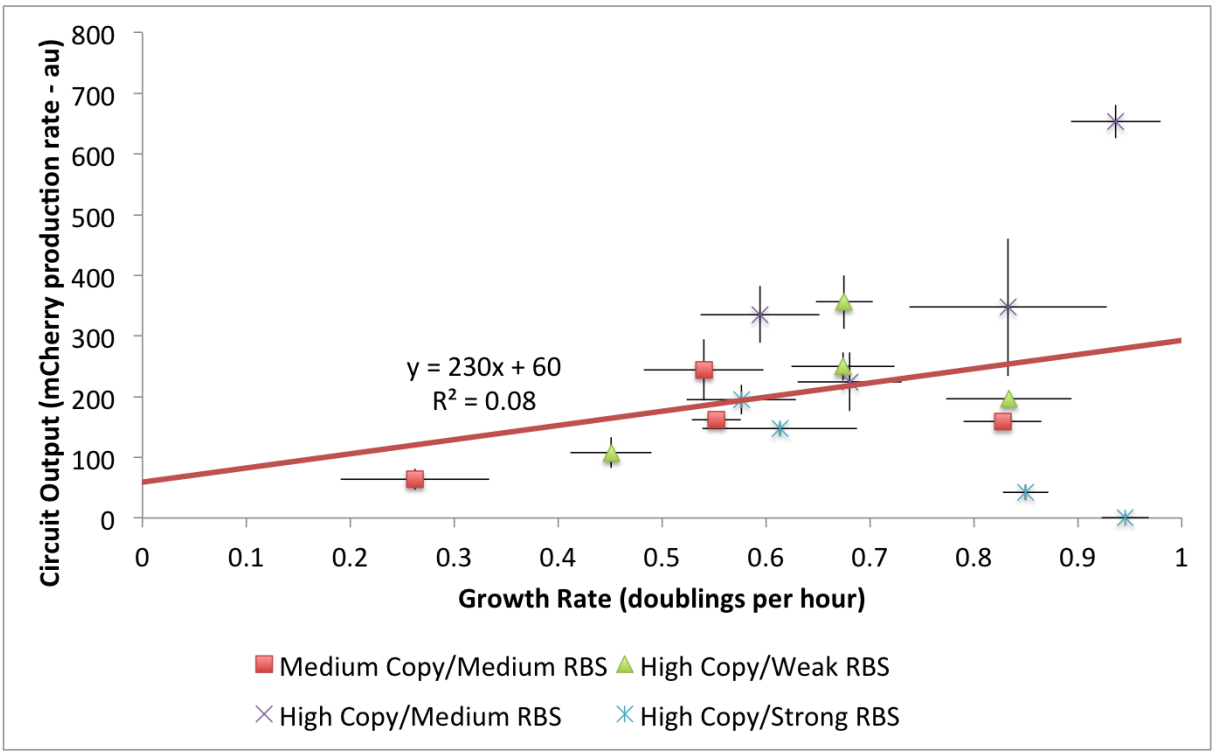


Figure 5.25: Scatter plot showing the relationship between growth rate and circuit output for all high copy constructs and medium copy constructs with medium RBS. Red line indicates a linear regression with R-squared value and equation of line shown. Error bars indicate standard deviations over 6 repeats.

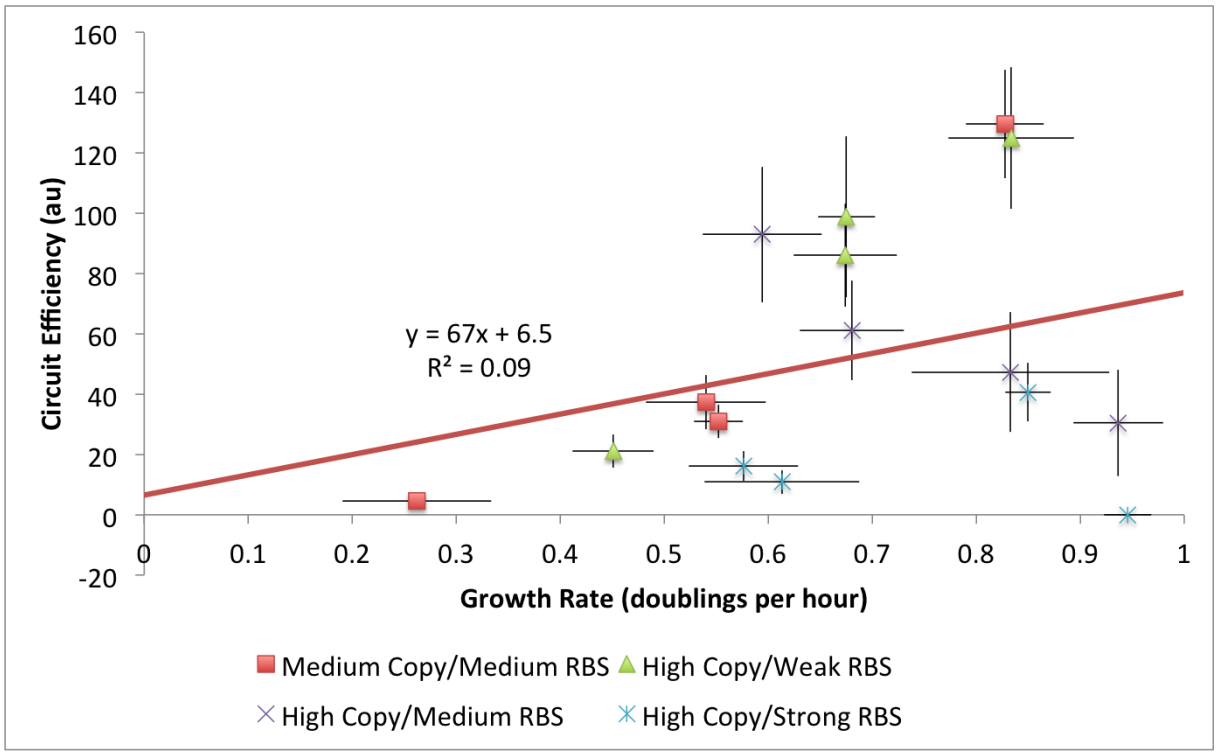


Figure 5.26: Scatter plot showing the relationship between growth rate and circuit efficiency for all high copy constructs and medium copy constructs with medium RBS. Red line indicates a linear regression with R-squared value and equation of line shown. Error bars indicate standard deviations over 6 repeats.

5.12 Conclusion

The results shown in this chapter show that we have been able to successfully build a library of constructs that have been used in conjunction with the capacity monitor to shed light on the impact of certain circuit design choices. We see a very complex interaction between gene expression control points that point us towards some design principles and help us uncover important considerations when designing genetic circuits.

The control points we have investigated are copy number, promoter strength, RBS strength and codon usage. We have clearly seen that increasing promoter strength leads to an increase in gene expression as well as a decreased in cellular capacity. RBS strength has a similar impact, however when particularly strong RBS sequences were used we saw a decreased circuit output with a decrease in capacity, most likely due to a cellular response to the burden being placed on the cell. Codon usage was shown to be very important in both the gene expression levels as well as the burden placed on the cell. Slow codons appear to be a poor design choice when

considering either the rate of protein production or the burden placed on the cell. Increasing the copy number of the circuit increases the circuit output, however the additional resources required to maintain a higher number of plasmids (because of origin of replication or antibiotic resistance marker) means that this increase comes at a high burden 'cost'.

We are able to uncover a design principle that allows two circuits to have the same rate of protein production whilst causing different levels of burden. In this situation we saw that the combination of a weak promoter with stronger RBS is less efficient (a term we define in this chapter) than the equivalent construct with a strong promoter and weak RBS.

Comparing the impact of circuits on DH10B cells and MG1655 cells we saw that MG1655 cells had a much larger decrease in capacity. This may have been due to the stringent response allowing the cells to detect the production of extra protein and adapt to cope with this by down regulating the monitor promoter. However, we observed that slow codons are still shown to be a poor design choice in MG1655 cells, indicating that this is a poor design choice across different strains of *E. coli*.

Using RNA quantification with qPCR techniques we were able to estimate the relative total RNA levels in cells containing different constructs as well as estimate the translation rates within the cells. The results showed that the growth rates and total RNA levels per cell are highly proportional, as indicated by Klumpp et al.[?]. This is observed when growth rate is mediated through changes in growth media in Klumpp at al. and therefore is not indicative that competition for transcriptional machinery is causing the decrease in total RNA. When these figures are combined with monitor protein production rates we are able to estimate translation rates for the monitor protein and we see that this is highly correlated with the rate of protein production level, indicating that changes in the capacity monitor are predominantly due to changes in translational resources.

We compared growth rate to circuit output, monitor output and circuit efficiency for the range of constructs tested in this chapter and we observed very low correlation between growth rate and each of these metrics. This indicates that growth rate is a poor indicator of cellular capacity in terms of shared resources and that the system we have developed is a much more accurate way of estimating this.

Chapter 6

Results: Modelling Burden Caused by Gene Expression

In this chapter we motivate and develop a model of gene expression. This model focuses on the translational aspect of gene expression as this is where both the literature and the experimental data from this project suggest the crucial interaction between gene expression and shared resources lies. We aim to use this model to simulate similar conditions to those tested experimentally and identify how closely *in silico* matches *in vivo*.

6.1 Basic Gene expression Model

We first propose a basic model of translation that only takes into account number of ribosomes and transcripts. This does not take into account any features of the RNA such as codon usage, RNA length etc. Therefore this model in isolation is not sufficient for our needs. We need to be able to incorporate these additional features.

$$\text{Pol} + \text{DNA} \underset{\alpha^1_-}{\overset{\alpha^1_+}{\rightleftharpoons}} \text{S}$$

$$\text{S} \xrightarrow{\beta^1} \text{Pol} + \text{DNA} + \text{RNA}$$

$$\text{Rib} + \text{RNA} \underset{\alpha^2_-}{\overset{\alpha^2_+}{\rightleftharpoons}} \text{L}$$

$$\text{L} \xrightarrow{\beta^2} \text{Rib} + \text{RNA} + \text{P}$$

where Pol is an RNA polymerase, DNA is a strand of DNA, S is a transcriptional complex of DNA and RNA polymerase, RNA is an mRNA transcript, Rib is a ribosomes, L is a translational complex of RNA and ribosomes and P is a protein.

6.2 Full Elongation Model

Both the literature and the wet-lab results from this project indicate that the bottleneck in resource availability in cells expressing heterologous protein is greatest at the ribosomal level[]. Therefore we chose to build a model that focuses on ribosomal availability. In order to include the ability for coding regions to have different codon profiles and for ribosomal traffic jams[] to be able to be modelled it was necessary to go beyond the 'one step' model shown above.

To create a model that could include these key features we had to look closely at the elongation process in translation. This is a complex process that consists of multiple steps every time the polypeptide is elongated and the ribosome moves along the transcript. Since obtaining the values for the parameters associated with these individual processes was not possible we collapsed them down in a way that each time the ribosome moved one codon along the transcript it was a single process. It was not possible to obtain these rates for the constructs used in this project either, however we were able to use the literature to estimate the rates and roughly model how the differences in codon usage might be reflected in the elongation rates used in the model.

We derived this model using a random-walk approach, though the same model has been derived elsewhere using more mechanistic and deterministic approaches[]. We confirmed that the model made sense by putting in parameters found in the literature and making sure that the outputs reflected what would be expected in vivo. Subsequently we modelled what the effects of changing the control points mentioned in above sections would have on the key metrics. Growth rate is not included in this model as the complexity of the interaction between resource availability and growth rate means it is not possible to model this interaction within the scope of this project.

6.2.1 Derivation

A model of translation was built up where the movement of individual ribosomes is treated as a random walk which occurs as follows:

Assumptions

Assumption 1. *There are a fixed number of ribosomes R.*

Assumption 2. *There is a single species of transcripts of which there is a constant number M.*

Assumption 3. *Each transcript is identical and is of length L codons.*

Assumption 4. *Ribosomes can reversibly bind to the RBS of a transcript.*

Assumption 5. *Once elongation is initiated and a ribosome has moved to the first codon of the transcript, it must continue unidirectionally along the transcript until it reaches the stop codon.*

Assumption 6. *When a ribosome reaches the stop codon it will release it and become a 'free ribosome' again and a protein will be produced.*

Assumption 7. *We approximate the size of ribosomes to be such that they only occupy a single codon (or RBS) along a transcript and neighbouring codons (and RBSes) can be occupied by separate ribosomes.*

Assumption 8. *No two ribosomes can occupy the same codon or RBS.*

Assumption 9. *Ribosomes move along transcript one codon at a time and cannot move to the next codon if it is occupied by another ribosome.*

Assumption 10. *Ribosomes move from one codon to the next at a fixed and constant rate if the next codon is not occupied.*

Assumption 11. *There is a large number of total ribosomes, so* $R >> 1$

Assumption 12. *Time is modelled discretely with intervals of* δt*. A maximum of one state transition for each ribosome may occur during this time (i.e. maximum one elongation step). Transitions from one elongation state to the next are single steps and as time is modelled discretely with intervals* δt

Assumption 13. *All transcripts are identical and so the probability of a ribosome* r *being in elongation stage* $i \in \{0, ..., L\}$ *at time* t *is the same for any mRNA:*

$$\mathbb{P}(E^r_{m,i}, t) = \mathbb{P}(E^r_{s,i}, t) \qquad \forall \ m, s \in \{1, ..., M\}$$

Assumption 14. *All ribosomes are identical and so have the same probability distribution:*

$$\mathbb{P}(E^r_{m,i}, t) = \mathbb{P}(E^q_{m,i}, t) \qquad \forall \ r, q \in \{1, ..., R\}$$

Assumption 15. *The event of a new transcript creation is defined as the moment an mRNA is finished being transcribed. Ribosomes move along mRNA closely following the RNA polymerase as it transcribes, and are already moving along the mRNA before it has finished being fully transcribed[]. Therefore upon the creation of a new mRNA we can approximate that it is already covered in ribosomes and that we can make a steady-state approximation over these equations.*

Assumption 16. *The process is Markovian and at any point in time the position of one ribosome is independent of the positions of others at that point in time.*

While we acknowledge some of these assumptions do not accurately represent the reality of the complex biological process of translation, we make them in order to simplify the model in a way that we do not anticipate will affect the core behaviours of the translation dynamics. For example, we know that a ribosome occupies space that covers more than one codon at a time, however by approximating it as occupying the space of only one codon we simplify the model significantly. What we lose in the accuracy of this specific detail we more than make up for in the increased ease with which we can work with the model.

Events

1. $E^r_{m,i}$ is the event of ribosome r being on transcript m in elongation state i (i.e. at the i^{th} codon) for $i \in \{1, ..., L\}$.
2. $E^r_{m,0}$ is the event of ribosome r being on the RBS of transcript m.
3. Rib^r is the event of ribosome r not being on any transcript (i.e. in the free ribosome pool).

For any ribosome 'r' from a pool of R ribosomes, if it is freely available (Rib^r) it can bind to the RBS of mRNA 'm' ($E^r_{m,0}$) and from this state it can either unbind and join the free ribosome pool again, or translation can be initiated and it moves into the initial state of elongation ($E^r_{m,1}$). From this the only path the ribosome can take is to go from the i$^{\text{th}}$ stage of elongation ($E^r_{m,i}$) to the i+1$^{\text{th}}$ stage of elongation ($E^r_{m,i+1}$) until it reaches the final elongation stage which, without

loss of generality, can be the L$^{\text{th}}$ stage ($E^r_{m,L}$). From this, translation finishes, a full protein is produced and the ribosome returns to the free ribosome pool.

$\mathbb{P}(E^r_{m,i}, t)$ is the probability that event $E^r_{m,i}$ occurs at time t. For the random walk we consider a discrete time distribution with steps of length δt and define the following 'rates':

Definition 1. *We define the 'unblocked' rates:*

$$\begin{aligned}
\alpha^+ &= \lim_{\delta t \to 0} \frac{\mathbb{P}\big(E^r_{m,0}, t+\delta t | Rib^r, t \cap (\bigcap_{q \neq r} \neg E^q_{m,0}, t+\delta t)\big)}{\delta t} \\
\alpha^- &= \lim_{\delta t \to 0} \frac{\mathbb{P}(Rib^r, t+\delta t | E^r_{m,0}, t)}{\delta t} \\
\beta_i &= \lim_{\delta t \to 0} \frac{\mathbb{P}\big(E^r_{m,i+1}, t+\delta t | E^r_{m,i}, t \cap (\bigcap_{q \neq r} \neg E^q_{m,i+1}, t+\delta t)\big)}{\delta t} \qquad \textit{for } i \in \{0, ..., L-1\} \\
\beta_L &= \lim_{\delta t \to 0} \frac{\mathbb{P}(Rib^r, t+\delta t | E^r_{m,L}, t)}{\delta t}
\end{aligned}$$

where α^+ is the binding rate of a ribosome to an RBS, α^- is the unbinding rate of a ribosome from an RBS and the β_i values are the rates of elongation at which a ribosome moves to the next codon (if it is not blocked).

The Model

The system can be displayed mathematically as:

$$\begin{aligned}
\mathbb{P}(Rib^r, t+\delta t) &= \mathbb{P}(Rib^r, t+\delta t | Rib^r, t)\mathbb{P}(Rib^r, t) \\
&\quad + \sum_{s=1}^{M} \mathbb{P}(Rib^r, t+\delta t | E^r_{s,0}, t)\mathbb{P}(E^r_{s,0}, t) \\
&\quad + \sum_{s=1}^{M} \mathbb{P}(Rib^r, t+\delta t | E^r_{s,L}, t)\mathbb{P}(E^r_{s,L}, t) \qquad (6.1a)
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(E^r_{m,0}, t+\delta t) &= \mathbb{P}(E^r_{m,0}, t+\delta t | E^r_{m,0}, t)\mathbb{P}(E^r_{m,0}, t) \\
&\quad + \mathbb{P}(E^r_{m,0}, t+\delta t | Rib^r, t)\mathbb{P}(Rib^r, t) \qquad (6.1b)
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(E^r_{m,1}, t+\delta t) &= \mathbb{P}(E^r_{m,1}, t+\delta t | E^r_{m,1}, t)\mathbb{P}(E^r_{m,1}, t) \\
&\quad + \mathbb{P}(E^r_{m,1}, t+\delta t | E^r_{m,0}, t)\mathbb{P}(E^r_{m,0}, t) \qquad (6.1c)
\end{aligned}$$

$$\vdots$$

$$\mathbb{P}(E^r_{m,i}, t+\delta t) = \mathbb{P}(E^r_{m,i}, t+\delta t | E^r_{m,i}, t)\mathbb{P}(E^r_{m,i}, t) + \mathbb{P}(E^r_{m,i}, t+\delta t | E^r_{m,i-1}, t)\mathbb{P}(E^r_{m,i-1}, t) \qquad \forall i \in \{2, ..., L-1\} \tag{6.1d}$$

$$\vdots$$

$$\mathbb{P}(E^r_{m,L}, t+\delta t) = \mathbb{P}(E^r_{m,L}, t+\delta t | E^r_{m,L}, t)\mathbb{P}(E^r_{m,L}, t) + \mathbb{P}(E^r_{m,L}, t+\delta t | E^r_{m,L-1}, t)\mathbb{P}(E^r_{m,L-1}, t) \tag{6.1e}$$

We next rewrite the probability of a ribosome staying in the same state as being equal to 1 minus the probability of it moving out of that state:

$$\mathbb{P}(Rib^r, t+\delta t) = \left(1 - \sum_{s=1}^{M} \mathbb{P}(E^r_{s,0}, t+\delta t | Rib^r, t)\right) \mathbb{P}(Rib^r, t) + \sum_{s=1}^{M} \mathbb{P}(Rib^r, t+\delta t | E^r_{s,0}, t)\mathbb{P}(E^r_{s,0}, t) + \sum_{s=1}^{M} \mathbb{P}(Rib^r, t+\delta t | E^r_{s,L}, t)\mathbb{P}(E^r_{s,L}, t) \tag{6.2a}$$

$$\mathbb{P}(E^r_{m,0}, t+\delta t) = \left(1 - \mathbb{P}(Rib^r, t+\delta t | E^r_{m,0}, t) - \mathbb{P}(E^r_{m,1}, t+\delta t | E^r_{m,0}, t)\right) \mathbb{P}(E^r_{m,0}, t) + \mathbb{P}(E^r_{m,0}, t+\delta t | Rib^r, t)\mathbb{P}(Rib^r, t) \tag{6.2b}$$

$$\mathbb{P}(E^r_{m,1}, t+\delta t) = \left(1 - \mathbb{P}(E^r_{m,2}, t+\delta t | E^r_{m,1}, t)\right) \mathbb{P}(E^r_{m,1}, t) + \mathbb{P}(E^r_{m,1}, t+\delta t | E^r_{m,0}, t)\mathbb{P}(E^r_{m,0}, t) \tag{6.2c}$$

$$\vdots$$

$$\mathbb{P}(E^r_{m,i}, t+\delta t) = \left(1 - \mathbb{P}(E^r_{m,i+1}, t+\delta t | E^r_{m,i}, t)\right) \mathbb{P}(E^r_{m,i}, t) + \mathbb{P}(E^r_{m,i}, t+\delta t | E^r_{m,i-1}, t)\mathbb{P}(E^r_{m,i-1}, t) \qquad \forall i \in \{2, ..., L-1\} \tag{6.2d}$$

$$\vdots$$

$$\mathbb{P}(E^r_{m,L}, t+\delta t) = \left(1 - \mathbb{P}(Rib^r, t+\delta t | E^r_{m,L}, t)\right) \mathbb{P}(E^r_{m,L}, t) + \mathbb{P}(E^r_{m,L}, t+\delta t | E^r_{m,L-1}, t)\mathbb{P}(E^r_{m,L-1}, t) \tag{6.2e}$$

Rearranging gives:

$$\begin{aligned}
\mathbb{P}(Rib^r, t+\delta t) - \mathbb{P}(Rib^r, t) &= -\sum_{s=1}^{M} \mathbb{P}(E^r_{s,0}, t+\delta t | Rib^r, t)\mathbb{P}(Rib^r, t) \\
&\quad + \sum_{s=1}^{M} \mathbb{P}(Rib^r, t+\delta t | E^r_{s,0}, t)\mathbb{P}(E^r_{s,0}, t) \\
&\quad + \sum_{s=1}^{M} \mathbb{P}(Rib^r, t+\delta t | E^r_{s,L}, t)\mathbb{P}(E^r_{s,L}, t) \qquad (6.3a)
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(E^r_{m,0}, t+\delta t) - \mathbb{P}(E^r_{m,0}, t) &= -\Big(\mathbb{P}(Rib^r, t+\delta t | E^r_{m,0}, t) - \mathbb{P}(E^r_{m,1}, t+\delta t | E^r_{m,0}, t)\Big)\mathbb{P}(E^r_{m,0}, t) \\
&\quad + \mathbb{P}(E^r_{m,0}, t+\delta t | Rib^r, t)\mathbb{P}(Rib^r, t) \qquad (6.3b)
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(E^r_{m,1}, t+\delta t) - \mathbb{P}(E^r_{m,1}, t) &= -\mathbb{P}(E^r_{m,2}, t+\delta t | E^r_{m,1}, t)\mathbb{P}(E^r_{m,1}, t) \\
&\quad + \mathbb{P}(E^r_{m,1}, t+\delta t | E^r_{m,0}, t)\mathbb{P}(E^r_{m,0}, t) \qquad (6.3c) \\
&\quad \vdots \\
\mathbb{P}(E^r_{m,i}, t+\delta t) - \mathbb{P}(E^r_{m,i}, t) &= -\mathbb{P}(E^r_{m,i+1}, t+\delta t | E^r_{m,i}, t)\mathbb{P}(E^r_{m,i}, t) \quad \text{for } i \in \{2, ..., L-1\} \\
&\quad + \mathbb{P}(E^r_{m,i}, t+\delta t | E^r_{m,i-1}, t)\mathbb{P}(E^r_{m,i-1}, t) \qquad (6.3d) \\
&\quad \vdots \\
\mathbb{P}(E^r_{m,L}, t+\delta t) - \mathbb{P}(E^r_{m,L}, t) &= -\mathbb{P}(Rib^r, t+\delta t | E^r_{m,L}, t)\mathbb{P}(E^r_{m,L}, t) \\
&\quad + \mathbb{P}(E^r_{m,L}, t+\delta t | E^r_{m,L-1}, t)\mathbb{P}(E^r_{m,L-1}, t) \qquad (6.3e)
\end{aligned}$$

For all events $E^r_{m,i}$ at a time $t+\delta t$ we have that the probability $\mathbb{P}(E^r_{m,i}, t+\delta t|X)$ for any event X can be split into two subsets, one where there is a ribosome in elongation state i on mRNA m at time t and one where there is not:

$$\begin{aligned}
\mathbb{P}(E^r_{m,i}, t|X)\mathbb{P}(X) &= \mathbb{P}(E^r_{m,i}, t|X \cap (\bigcup_{q \neq r} E^q_{m,i}, t))\mathbb{P}(X \cap (\bigcup_{q \neq r} E^q_{m,i}, t)) \\
&\quad + \mathbb{P}(E^r_{m,i}, t|X \cap (\bigcap_{q \neq r} \neg E^q_{m,i}, t))\mathbb{P}(X \cap (\bigcap_{q \neq r} \neg E^q_{m,i}, t)) \qquad (6.4)
\end{aligned}$$

Since the probability of two ribosomes being in the same state on the same mRNA is zero we

must have that:

$$\mathbb{P}(E^r_{m,i},t|X)\mathbb{P}(X) \quad = \quad \mathbb{P}(E^r_{m,i},t|X\cap(\bigcap_{q\neq r}\neg E^q_{m,i},t))\mathbb{P}(X\cap(\bigcap_{q\neq r}\neg E^q_{m,i},t)) \tag{6.5}$$

which can be rewritten as:

$$\mathbb{P}(E^r_{m,i},t|X)\mathbb{P}(X) \quad = \quad \mathbb{P}(E^r_{m,i},t|X\cap(\bigcap_{q\neq r}\neg E^q_{m,i},t))\mathbb{P}(\bigcap_{q\neq r}\neg E^q_{m,i},t|X)\mathbb{P}(X) \tag{6.6}$$

Due to mutual exclusivity, we know that the probability of no other ribosomes being there is equal to 1 minus the sum of the probabilities of each other ribosome being there:

$$\mathbb{P}(E^r_{m,i},t|X)\mathbb{P}(X) \quad = \quad \mathbb{P}(E^r_{m,i},t|X\cap(\bigcap_{q\neq r}\neg E^q_{m,i},t))(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,i},t|X))\mathbb{P}(X) \tag{6.7}$$

Combining this with equations (6.3) gives:

$$\begin{aligned}\mathbb{P}(Rib^r,t+\delta t)-\mathbb{P}(Rib^r,t) \quad = \quad & -\sum_{s=1}^{M}\Bigg(\mathbb{P}\big(E^r_{s,0},t+\delta t|Rib^r,t\cap(\bigcap_{q\neq r}\neg E^q_{s,0},t+\delta t)\big)\\ & \qquad\cdot\Big(1-\sum_{q\neq r}\mathbb{P}(E^q_{s,0},t+\delta t|Rib^r,t)\Big)\mathbb{P}(Rib^r,t)\Bigg)\\ & +\sum_{s=1}^{M}\mathbb{P}(Rib^r,t+\delta t|E^r_{s,0},t)\mathbb{P}(E^r_{s,0},t)\\ & +\sum_{s=1}^{M}\mathbb{P}(Rib^r,t+\delta t|E^r_{s,L},t)\mathbb{P}(E^r_{s,L},t)\end{aligned} \tag{6.8a}$$

$$\begin{aligned}\mathbb{P}(E^r_{m,0},t+\delta t)-\mathbb{P}(E^r_{m,0},t) \quad = \quad & -\mathbb{P}(Rib^r,t+\delta t|E^r_{m,0},t)\mathbb{P}(E^r_{m,0},t)\\ & -\Bigg(\mathbb{P}\big(E^r_{m,1},t+\delta t|E^r_{m,0},t\cap(\bigcap_{q\neq r}\neg E^q_{m,1},t+\delta t)\big)\\ & \qquad\cdot\Big(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,1},t|E^r_{m,0},t)\Big)\mathbb{P}(E^r_{m,0},t)\Bigg)\\ & +\Bigg(\mathbb{P}\big(E^r_{m,0},t+\delta t|Rib^r,t\cap(\bigcap_{q\neq r}\neg E^q_{m,0},t+\delta t)\big)\\ & \qquad\cdot\Big(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,0},t|Rib^r,t)\Big)\mathbb{P}(Rib^r,t)\Bigg)\end{aligned} \tag{6.8b}$$

$$\mathbb{P}(E^r_{m,1},t+\delta t)-\mathbb{P}(E^r_{m,1},t) \quad = \quad -\Bigg(\mathbb{P}\big(E^r_{m,2},t+\delta t|E^r_{m,1},t\cap(\bigcap_{q\neq r}\neg E^q_{m,2},t+\delta t)\big)$$

$$
\begin{aligned}
&\cdot\Big(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,2},t|E^r_{m,1},t)\Big)\mathbb{P}(E^r_{m,1},t)\Big)\\
&+\Big(\mathbb{P}\big(E^r_{m,1},t+\delta t|E^r_{m,0},t\cap(\bigcap_{q\neq r}\neg E^q_{m,1},t+\delta t)\big)\\
&\cdot\Big(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,1},t|E^r_{m,0},t)\Big)\mathbb{P}(E^r_{m,0},t)\Big) \qquad (6.8c)
\end{aligned}
$$

$$\vdots$$

$$
\begin{aligned}
\mathbb{P}(E^r_{m,i},t+\delta t)-\mathbb{P}(E^r_{m,i},t) &= -\Big(\mathbb{P}\big(E^r_{m,i+1},t+\delta t|E^r_{m,i},t\cap(\bigcap_{q\neq r}\neg E^q_{m,i+1},t+\delta t)\big)\\
&\cdot\Big(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,i+1},t|E^r_{m,i},t)\Big)\mathbb{P}(E^r_{m,i},t)\Big)\\
&+\Big(\mathbb{P}\big(E^r_{m,i},t+\delta t|E^r_{m,i-1},t\cap(\bigcap_{q\neq r}\neg E^q_{m,i},t+\delta t)\big)\\
&\cdot\Big(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,i},t|E^r_{m,i-1},t)\Big)\mathbb{P}(E^r_{m,i-1},t)\Big) \qquad (6.8d)\\
&\forall i\in\{2,...,L-1\} \qquad (6.8e)
\end{aligned}
$$

$$\vdots$$

$$
\begin{aligned}
\mathbb{P}(E^r_{m,L},t+\delta t)-\mathbb{P}(E^r_{m,L},t) &= -\mathbb{P}(Rib^r,t+\delta t|E^r_{m,L},t)\mathbb{P}(E^r_{m,L},t)\\
&+\Big(\mathbb{P}\big(E^r_{m,L},t+\delta t|E^r_{m,L-1},t\cap(\bigcap_{q\neq r}\neg E^q_{m,L},t+\delta t)\big)\\
&\cdot\Big(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,L},t|E^r_{m,L-1},t)\Big)\mathbb{P}(E^r_{m,L-1},t)\Big) \qquad (6.8f)
\end{aligned}
$$

Dividing both sides by δt and taking $\lim_{\delta t\to 0}$ as well as taking the definitions of the rates as mentioned in Definition 1 gives:

$$
\begin{aligned}
\frac{d\mathbb{P}(Rib^r,t)}{dt} &= \sum_{s=1}^{M}\alpha^-\mathbb{P}(E^r_{s,0},t)\\
&-\sum_{s=1}^{M}\alpha^+\mathbb{P}(Rib^r,t)(1-\sum_{q\neq r}\mathbb{P}(E^q_{s,0},t))\\
&+\sum_{s=1}^{M}\beta_L\mathbb{P}(E^r_{s,L},t) \qquad (6.9a)\\
\frac{d\mathbb{P}(E^r_{m,0},t)}{dt} &= -\alpha^-\mathbb{P}(E^r_{m,0},t)\\
&+\alpha^+\mathbb{P}(Rib^r,t)(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,0},t))\\
&-\beta_0\mathbb{P}(E^r_{m,0},t))(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,1},t)) \qquad (6.9b)
\end{aligned}
$$

$$\begin{aligned}
\frac{d\mathbb{P}(E^r_{m,1},t)}{dt} &= \beta_0\mathbb{P}(E^r_{m,0},t)(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,1},t)) \\
&\quad -\beta_1\mathbb{P}(E^r_{m,1},t)(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,2},t)) \qquad (6.9c)\\
&\vdots \\
\frac{d\mathbb{P}(E^r_{m,i},t)}{dt} &= \beta_{i-1}\mathbb{P}(E^r_{m,i-1},t)(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,i},t)) \\
&\quad -\beta_i\mathbb{P}(E^r_{m,i},t)(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,i+1},t)) \qquad (6.9d)\\
&\vdots \\
\frac{d\mathbb{P}(E^r_{m,L},t)}{dt} &= \beta_{L-1}\mathbb{P}(E^r_{m,L-1},t)(1-\sum_{q\neq r}\mathbb{P}(E^q_{m,L},t)) \\
&\quad -\beta_L\mathbb{P}(E^r_{m,L},t) \qquad (6.9e)
\end{aligned}$$

Using assumptions (4) and (5) we can take the sums and along with using assumption (6) and saying $R >> 1 \implies R-1 \simeq R$ (the minimal number of ribosomes we have modelled this system with is 1000, and often higher meaning this assumption holds) we get:

$$\begin{aligned}
\frac{d\mathbb{P}(Rib^r,t)}{dt} &= M\cdot\alpha^-\mathbb{P}(E^r_{m,0},t) \\
&\quad -M\cdot\alpha^+\mathbb{P}(Rib^r,t)(1-R\cdot\mathbb{P}(E^r_{m,0},t)) \\
&\quad +M\cdot\beta_L\mathbb{P}(E^r_{m,L},t) \qquad (6.10a)\\
\frac{d\mathbb{P}(E^r_{m,0},t)}{dt} &= -\alpha^-\mathbb{P}(E^r_{m,0},t) \\
&\quad +\alpha^+\mathbb{P}(Rib^r,t)(1-R\cdot\mathbb{P}(E^r_{m,0},t)) \\
&\quad -\beta_0\mathbb{P}(E^r_{m,0},t))(1-R\cdot\mathbb{P}(E^r_{m,1},t)) \qquad (6.10b)\\
\frac{d\mathbb{P}(E^r_{m,1},t)}{dt} &= \beta_0\mathbb{P}(E^r_{m,0},t)(1-R\cdot\mathbb{P}(E^r_{m,1},t)) \\
&\quad -\beta_1\mathbb{P}(E^r_{m,1},t)(1-R\cdot\mathbb{P}(E^r_{m,2},t)) \qquad (6.10c)\\
&\vdots \\
\frac{d\mathbb{P}(E^r_{m,i},t)}{dt} &= \beta_{i-1}\mathbb{P}(E^r_{m,i-1},t)(1-R\cdot\mathbb{P}(E^r_{m,i},t)) \\
&\quad -\beta_i\mathbb{P}(E^r_{m,i},t)(1-R\cdot\mathbb{P}(E^r_{m,i+1},t)) \qquad (6.10d)\\
&\vdots \\
\frac{d\mathbb{P}(E^r_{m,L},t)}{dt} &= \beta_{L-1}\mathbb{P}(E^r_{m,L-1},t)(1-R\cdot\mathbb{P}(E^r_{m,L},t)) \\
&\quad -\beta_L\mathbb{P}(E^r_{m,L},t) \qquad (6.10e)
\end{aligned}$$

A set of randoms variables $X_{m,i}$ ($i \in \{0...L\}$) is defined as follows:

$$X_{m,i}(t) = \begin{cases} 1 & \text{if there is a ribosome present in elongation stage '}i\text{' on mRNA '}m\text{' at time '}t\text{'} \\ 0 & \text{if there is no ribosome present in elongation stage '}i\text{' on mRNA '}m\text{' at time '}t\text{'} \end{cases}$$

The random variable $F(t)$ represents the number of ribosomes not on a transcript at time t. At any time 't', using assumption (17) on independence of ribosome positions:

$$\mathbb{P}(X_{m,i}(t) = 1) = \sum_r \mathbb{P}(E^r_{m,i}, t)$$

and

$$\mathbb{P}(X_{m,i}(t) = 0) = 1 - \sum_r \mathbb{P}(E^r_{m,i}, t)$$

By the definition of expectation:

$$\mathbb{E}(X_{m,i}(t)) = 1 \cdot \mathbb{P}(X_{m,i}(t) = 1) + 0 \cdot \mathbb{P}(X_{m,i}(t) = 0)$$

so that,

$$\mathbb{E}(X_{m,i}(t)) = \sum_r \mathbb{P}(E^r_{m,i}, t)$$

Using assumption (1) we get:

$$\mathbb{E}(X_{m,i}(t)) = R \cdot \mathbb{P}(E^r_{m,i}, t) \tag{6.11}$$

where R is the total number of ribosomes. We further define the random variable $X_i(t)$ as the sum of random variables $X_{m,i}(t)$ across all mRNA, i.e. the total number of ribosomes in position i across all transcripts:

$$X_i(t) = \sum_m X_{m,i}(t)$$

which, taking expectations, gives

$$\mathbb{E}(X_i(t)) = \sum_m \mathbb{E}(X_{m,i}(t))$$

Combining with equation (6.11) gives

$$\mathbb{E}(X_i(t)) = \sum_m R \cdot \mathbb{P}(E^r_{m,i}, t) \tag{6.12}$$

Now, using assumption (2), we have:

$$\mathbb{E}(X_i(t)) = MR \cdot \mathbb{P}(E^r_{m,i}, t) \tag{6.13}$$

We define the variable $Y_i(t)$ to be the expectation of the random variable $X_i(t)$

$$Y_i(t) = \mathbb{E}(X_i(t)) \tag{6.14}$$

Therefore,

$$Y_i(t) = MR \cdot \mathbb{P}(E^r_{m,i}, t) \quad \forall i \in \{0, ..., L\} \tag{6.15}$$

We also investigate the variance of $X_i(t)$ to be confident that the system will reliably behave as the expectation dictates. The variance of each $X_{m,i}(t)$ is equal to:

$$Var(X_{m,i}(t)) \quad = \quad \mathbb{E}\left(X_{m,i}(t)^2\right) - \mathbb{E}\left(X_{m,i}(t)\right)^2 \tag{6.16}$$

However, since $X_{m,i}(t)$ can only take the values 0 or 1, it must be true that $X_{m,i}(t)^2 = X_{m,i}(t)$ and so, if we let $\mu = \mathbb{E}(X_{m,i}(t))$ it must be the case that:

$$Var(X_{m,i}(t)) \quad = \quad \mu - \mu^2 \tag{6.17}$$

$X_i(t)$ is a random variable that represents the sum of a population of independent, identically distributed (IID) random variables $X_{m,i}(t)$. From the variance of a population of independent, identically distributed random variables we get:

$$Var(X_i(t)) = \frac{\mu - \mu^2}{M} \tag{6.18}$$

This gives us an estimate of the cell to cell variance we would expect in these values from this model, however there are many other factors not included in this model which cause both cell to cell variations as well as population level variations in circuit output. This indicates that the behaviour of the circuit becomes less noisy as the number of transcripts increases since the variance per cell is inversely proportional to the number of transcripts. This is an interesting result and suggests that a stronger promoter would cause lower cell to cell variation.

It is tempting to include a 'deterministic' variance for the simulation we perform to estimate what the cell to cell variation would be. However, our simulations are performed to represent populations of identical cells (similar to the in vivo results shown) and therefore expected cell to cell variation cannot be compared to the population to population variance we see from in vivo results. This is because in vivo populations contain very large numbers of cells in which any variation as predicted by our model would be silenced, and the variation observed in vivo is due to additional factors not included in our model.

The random variable F which is the number of free ribosomes can be calculated as the total number of ribosomes minus the expected total number of ribosomes on transcripts:

$$F(t) = R - \sum_i X_i(t) \tag{6.19}$$

and letting $G(t)$ be the expectation of the random variable $F(t)$

$$G(t) = \mathbb{E}(F(t)) \tag{6.20}$$

by combining (6.15) and (6.19) with (6.10) and dropping the (t) from notation by letting $X_i = X_i(t)$ and $F = F(t)$ we are left with:

$$\frac{dG}{dt} = -M\alpha^+ G(1 - Y_0/M) + \alpha^- Y_0 + \beta_L Y_L \tag{6.21a}$$

$$\frac{dY_0}{dt} = M\alpha^+ G(1 - Y_0/M) - \alpha^- Y_0 - \beta_0 Y_0(1 - Y_1/M) \tag{6.21b}$$

$$\frac{dY_1}{dt} = \beta_0 Y_0(1 - Y_1/M) - \beta_1 Y_1(1 - Y_2/M) \tag{6.21c}$$

$$\vdots$$

$$\frac{dY_1}{dt} = \beta_{i-1} Y_{i-1}(1 - Y_i/M) - \beta_i Y_i(1 - Y_{i+1}/M) \tag{6.21d}$$

$$\vdots$$

$$\frac{dY_L}{dt} = \beta_{L-1} Y_{L-1}(1 - Y_L/M) - \beta_L Y_L \tag{6.21e}$$

We assume that the system is in steady state in exponential growth where each transcript has a steady state distribution of ribosomes on it (Assumption 15). The steady state equations are:

$$M\alpha^+ G(1 - Y_0/M) = \alpha^- Y_0 + \beta_L Y_L \tag{6.22}$$

$$\begin{aligned} \beta_0 Y_0(1 - Y_1/M) &= \beta_1 Y_1(1 - Y_2/M) \\ &= \vdots \\ &= \beta_{i-1} Y_{i-1}(1 - Y_i/M) \\ &= \vdots \\ &= \beta_{L-1} Y_{L-1}(1 - Y_L/M) \\ &= \beta_L Y_L \end{aligned} \tag{6.23}$$

6.2.2 Solving the Steady State Model

Rearranging Equation (6.23) and letting

$$Rib = G$$

$$L = Y_0$$

$$F_i = X_i$$

we can define functions f_{Rib}, f_L and f_k $\quad \forall k \in \{1, ..., m-1\}$:

$$
\begin{aligned}
Rib &= f_{Rib}(L, F_1) && \text{(6.24a)} \\
L &= f_L(F_1, F_2) && \text{(6.24b)} \\
F_k &= f_k(F_{k+1}, F_{k+2}) \qquad \forall k \in \{1, ..., m-2\} && \text{(6.24c)} \\
F_{m-1} &= f_{m-1}(F_m) && \text{(6.24d)}
\end{aligned}
$$

These functions can be rewritten as:

$$
\begin{aligned}
Rib &= g_{Rib}(F_m) && \text{(6.25a)} \\
L &= g_L(F_m) && \text{(6.25b)} \\
F_k &= g_k(F_m) \qquad \forall k \in \{1, ..., m-2\} && \text{(6.25c)} \\
F_{m-1} &= g_{m-1}(F_m) && \text{(6.25d)}
\end{aligned}
$$

6.2.3 Asserting monotonicity in the model

In order to assert that there is a unique solution to set of equations 6.25 we must ensure that the total number of ribosomes can be expressed as a strictly monotonically increasing function of ribosomes in the free pool. In order to do this we must use the inverse function theorem and show that each variable is a strictly monotonically increasing function of free ribosomes.

$$
\begin{aligned}
F_{m-1} &= g_{m-1}(F_m) && \text{(6.26a)} \\
&= \frac{\beta_m F_m}{\beta_{m-1}(R^T - F_m)} && \text{(6.26b)}
\end{aligned}
$$

$$\frac{dF_{m-1}}{dF_m} = \frac{(\beta_m/\beta_{m-1})R^T}{(R^T - F_m)^2} \tag{6.27a}$$

$$> 0 \tag{6.27b}$$

So F_{m-1} is a strictly monotonically increasing function of F_m. For F_{m-2}:

$$F_{m-2} = g_{m-2}(F_m) \tag{6.28a}$$

$$= \frac{\beta_m F_m}{\beta_{m-2}(R^T - F_{m-1})} \tag{6.28b}$$

$$= \frac{\beta_m F_m}{\beta_{m-2}(R^T - \frac{\beta_m F_m}{\beta_{m-1}(R^T - F_m)})} \tag{6.28c}$$

$$\frac{dF_{m-2}}{dF_m} = (\beta_m/\beta_{m-2})\frac{R^T - (F_{m-1} - F_m\frac{dF_{m-1}}{dF_m})}{(R^T - F_{m-1})^2} \tag{6.29a}$$

But we have that:

$$F_{m-1} - F_m\frac{dF_{m-1}}{dF_m} = \frac{\beta_m}{\beta_{m-2}}\frac{F_m}{(R^T - F_m)} - F_m\frac{R^T}{(R^T - F_m)^2} \tag{6.30a}$$

$$= \frac{\beta_m}{\beta_{m-2}}\frac{F_m R^T - F_m^2 - F_m R^T}{(R^T - F_m)^2} \tag{6.30b}$$

$$= -\frac{\beta_m}{\beta_{m-2}}\frac{F_m^2}{(R^T - F_m)^2} \tag{6.30c}$$

$$\leq 0 \tag{6.30d}$$

Therefore

$$\frac{dF_{m-2}}{dF_m} > 0 \tag{6.31}$$

and F_{m-2} is a strictly monotonically increasing function of F_m. Generally for F_i: ASSUME:

$$\frac{dF_{i+2}}{dF_m} > 0 \tag{6.32}$$

$$F_i = g_i(F_m) \tag{6.33a}$$

$$= \quad (\beta_m/\beta_i)\frac{F_m}{R^T - F_{i+1}} \tag{6.33b}$$

Differentiating:

$$\frac{dF_i}{dF_m} \quad = \quad \frac{\beta_m}{\beta_i}\frac{R^T - (F_{i+1} - F_m\frac{dF_{i+1}}{dF_m})}{(R^T - F_{i+1})^2} \tag{6.34a}$$

Expanding and taking into account Equation 6.32:

$$F_{i+1} - F_m\frac{dF_{i+1}}{dF_m} \quad = \quad \frac{\beta_m}{\beta_{i+2}}(\frac{F_m}{(R^T - F_{i+2})} - F_m\frac{R^T - F_{i+2} + F_m\frac{dF_{i+2}}{dF_m}}{(R^T - F_{i+2})^2}) \tag{6.35a}$$

$$= \quad \frac{\beta_m}{\beta_{i+2}}\frac{R^T F_m - F_m F_{i+2} - F_m R^T + F_m F_{i+2} - F_M^2\frac{dF_{i+2}}{dF_m}}{(R^T - F_{i+2})^2} \tag{6.35b}$$

$$= \quad -\frac{\beta_m}{\beta_{i+2}}\frac{F_M^2\frac{dF_{i+2}}{dF_m}}{(R^T - F_{i+2})^2} \tag{6.35c}$$

$$\leq \quad 0 \tag{6.35d}$$

Which gives us that:

$$\frac{dF_i}{dF_m} > 0 \tag{6.36}$$

Since $\frac{dF_m}{dF_m} > 0$ and $\frac{dF_{m-1}}{dF_m} > 0$ we can use inductive reasoning to get that g_{Rib}, g_L and g_k for $\forall k \in \{1, ..., m\}$ are all strictly monotonically increasing functions in F_m and therefore (by the inverse function theorem) have inverse functions. This means we can rewrite all the variables as strictly monotonically increasing functions of Rib:

$$Rib \quad = \quad h_{Rib}(Rib) \tag{6.37a}$$

$$L \quad = \quad h_L(Rib) \tag{6.37b}$$

$$F_k \quad = \quad h_k(Rib) \qquad \forall k \in \{1, ..., m\} \tag{6.37c}$$

Conservation of ribosomes gives that:

$$Rib + h_L(Rib) + \sum_{k=1}^{m} h_k(Rib) = Rib^T \tag{6.38}$$

The left-hand-side of this equation is a sum of strictly monotonically increasing functions and therefore is itself a strictly monotonically increasing function of Rib and tends to $+\infty$ as Rib and

tends to $+\infty$. Therefore, for any Rib^T we have a unique solution for Rib, and from equations (h functions) unique solutions in L and F_k for $\forall k \in \{1, ..., m\}$.

6.3 Simulating Circuit and Monitor Behaviour

Equations 6.37 and 6.38 cannot be solved analytically for systems that are large enough to be representative of real synthetic circuits. Therefore, the model must be simulated to understand how changing the control points affects the circuit behaviour and free ribosome pool.

A python script was built that allowed this model to be simulated. It consists of two classes, `Circuit` and `Cell`. The `Circuit` class describes individual circuits and allows a user to define the number of transcripts, elongation rates, binding and unbinding affinities for RBS. This can be done for any number of circuits. The `Cell` class allows user to define a model of a cell including the total number of ribosomes available in the cell as well as which circuits it contains. These classes have attributes and methods that allow simulations of the system to be run. The method `simulate` on the `Cell` class allows the simulation of the cell to be run and gives a dictionary output that describes the number of free ribosomes remaining in the cell as well as the distribution of ribosomes for each circuit. This script uses functions provided by the scipy python package, which thus must be installed a priori.

```
from scipy.optimize import fsolve


################################
# Ribosomal Competition Model #
################################


class Cell(object):
    """
    Object respresenting a cell.

    free_ribosomes = integer number of ribosomes available in the cell
        for synthetic circuits to use.
```

```
    circuits = list of synthetic gene circuits in cell.

    """
    def __init__(self, free_ribosomes=1000, circuits=[]):
        self.free_ribosomes = free_ribosomes
        self.circuits = circuits
        self.array = [0] #free ribosomes
        for circuit in circuits:
            self.array += [0]

    def conservation(self,p):
        return (sum(p)-self.free_ribosomes,)

    def equation(self,p):
        circuit_lengths = []
        q = [p[0]]
        shift_counter = 1
        for circuit in self.circuits:
            length = circuit.length + 1
            q += [[p[shift_counter:shift_counter+length]]]
            shift_counter += length
        eqns = self.conservation(p)
        for i, circuit in enumerate(self.circuits):
            eqns = eqns + circuit.equation(q,i)
        return eqns

    def simulate(self):
        initial_conditions = [self.free_ribosomes]
        for circuit in self.circuits:
            initial_conditions += circuit.initial_conditions
        solution = fsolve(self.equation, initial_conditions)
        result = {'free_ribosomes': solution[0], 'circuits': []}
```

```
            shift_counter = 1
            for circuit in self.circuits:
                length = circuit.length+1
                result['circuits'] += [solution[shift_counter:shift_counter +
                                                length]]
                shift_counter += length
            return result


class Circuit(object):
    """
    Object representing synthetic gene circuits that will be placed into
    a cell.

    total_transcripts = integer number of transcripts for this circuit
        in the cell. Is a function of both copy number and promoter
        strength.

    alpha_plus = rate at which free ribosomes bind to RBS.

    alpha_minus = rate at which ribosomes unbind from RBS.

    betas = list of rates for ribosomes moving along transcript. betas[0]
        represents the rate at which ribosome moves from RBS to initial
        elongation state. betas[i] represents rate at which ribosomes
        moves from position i to position i+1 if unblocked (or into free
        ribosome pool for i = length(betas)).

    RBS_strength = single number that replaces alpha_plus, alpha_minus
        and betas[0] if defined.

    """
```

```
def __init__(self, total_transcripts, alpha_plus_scale=0.00001,
             alpha_minus_scale=60, length=100, betas=None,
             RBS_strength=None, speed=20):
    self.total_transcripts = total_transcripts
    if betas:
        self.betas = betas
        self.length = len(betas)-1
    else:
        self.betas = [speed for i in range(length+1)]
        self.length = length
    if RBS_strength:
        self.betas[0] = RBS_strength
    self.alpha_plus = alpha_plus_scale*speed*self.betas[0]
    self.alpha_minus = alpha_minus_scale*speed/self.betas[0]
    self.initial_conditions = [0 for i in range(self.length+1)]]


def equation(self,q,index):
    """
    Provides ODE equation set for this species solving fsolve
    p is the equation input
    index is index of equation in list for master fsolve
    """
    L = self.length
    T = self.total_transcripts
    a_p = self.alpha_plus
    a_m = self.alpha_minus
    b = self.betas
    x = q[0]
    y = q[index+1][0]
    eqns = ( a_p * x * (T - y[0])
                - a_m * y[0]
                - b[0] * y[0] * (1 - y[1]/T) ,)
```

```
for eqn in (( b[i] * y[i] * (1 - y[i+1]/T)
                - b[i+1] * y[i+1] * (1 - y[i+2]/T) ,)
                for i in range(L-1)):
    eqns = eqns + eqn
eqns = eqns + (b[L-1] * y[L-1] * (1 - y[L]/T) - b[L] * y[L],)
return eqns
```

It is trivial to extend the model described in Section 6.2.2 to a system of two (or more) circuits. A simulation was done of a two circuit system in a way where one circuit represented the monitor and the other represented a synthetic circuit. Simulation of this model allows predictions to be made about changes in the behaviour of a synthetic circuit when the key control points discussed in Section **??** are altered as well as how the expected output from the monitor changes. These results can then be compared to the wet-lab data from Chapter **??** to verify their validity.

6.3.1 Parameter and Unit Checking

In order to test this model, we start by performing a simulation with realistic values that are observed in the actual system. Using parameters obtained from BioNumbers[], we run a simulation to test whether the output values observed are within realistic bounds. Table 6.1 shows the parameters. These roughly represent a medium copy plasmid (25-50 copies per cell) with a medium promoter (2-4 transcripts per promoter in cell at any time) giving 100 transcripts, 1000 available ribosomes (5% of total cellular ribosomes at 20,000), a 900 bp CDS (300 amino acids long) that has been codon optimised so elongation rate at each codon is 20 codons per second for all codons.

Parameter	**Model Parameter**	**Value**	**Units**
Codon speed (elongation rate)	β_i for all i	20	ribosomes^{-1} s^{-1}
Transcripts	R^T	100	mRNA cell^{-1}
Available ribosomes	Rib_T	1000	ribosomes cell^{-1}
Transcript length	m	300	codons
Ribosome-RBS binding rate	α_+	0.0001	ribsome^{-1} RBS^{-1} s^{-1}
Ribosome-RBS unbinding rate	α_-	200	ribsome-RBS-complex^{-1} s^{-1}

Table 6.1: Model parameters used for testing model validity

Running a simulation of a single circuit with ribosomes gives a circuit that produces proteins at an average rate of 35.04 proteins per second. This uses an average of 537.97 ribosomes at any point in time which is 2.5% of all cellular ribosomes. This appears to be the correct order of magnitude since there are approximately 4000 genes, of which perhaps half are active. This gives 2000 active promoters with approximately 2-4 transcripts per promoter with a total of 4000-8000 cellular transcripts per cell. The 100 transcripts from the synthetic construct constitute 1.25-2.5% of the total cellular transcripts and therefore we would expect the same proportion of the total cellular ribosomes to be on circuit transcripts.

6.4 Modelling Control Points

We use a simple two circuit simulation to investigate the effect of changing the parameters associated with the different control points. Since we do not know the exact biological parameters for the systems we are investigating we cannot expect an accurate quantitative prediction of the impacts of specific changes. However, we can do a comparative investigation where we look at the qualitative and relative changes in system behaviour when we change the control point parameters.

6.4.1 Promoter Strength and Copy Number

The model being considered in this project only models ribosomal availability and therefore when considering the number of circuit transcripts it is independent of the mechanisms that cause changes in the amount of mRNA. Plasmid backbones cause different levels of 'background' burden on the cell and are not considered as part of this model due the the higher levels of complexity of different origins of replication and resistance markers.

A suitable approach would be to characterise the behaviour of the backbone as demonstrated in Section **??** and use this modelling approach to predict how to best optimise the design of the circuit contained in the plasmid given a set of constraints.

In this modelling approach the plasmid copy number and promoter strength are compounded into a single variable - the number of transcripts. Figure 6.1 shows the amount of circuit output and monitor output for the model system for a range of transcript numbers. At low levels of

transcripts (< 400 per cell) the relationship between transcript number and circuit output is approximately proportional. Similarly, the relationship between the number of transcripts and monitor output is approximately linear in this region. This indicates that for a given number of ribosomes and for transcript numbers in this range, all transcripts use a similar number of ribosomes to produce protein at a similar rate.

As the number of transcripts increases, the system becomes saturated with respect to transcripts and large increases in the number of transcripts cause relatively small increases and decreases in circuit output and monitor output respectively.

The vertical yellow lines in Figure 6.1 show where the data shown in Figure 6.2 lies on this graph.

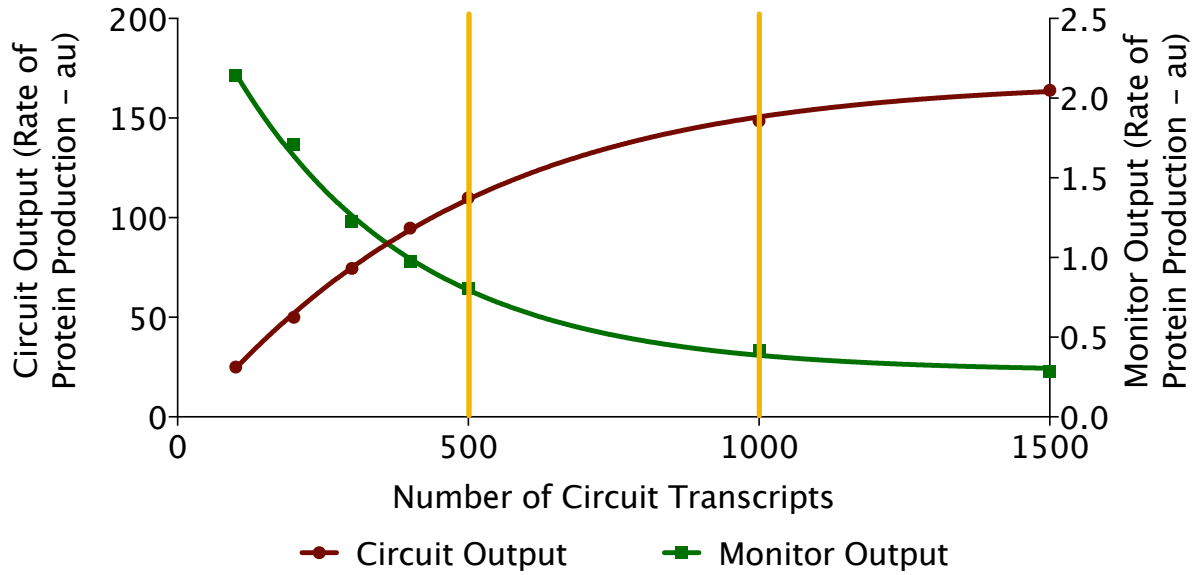


Figure 6.1: Modelled impact of transcript number on circuit and monitor outputs. This figure shows both both monitor output and circuit output for a range of circuit transcript numbers. Lines represent best fit of hill curves using GraphPad Prism with no parameter constraints.

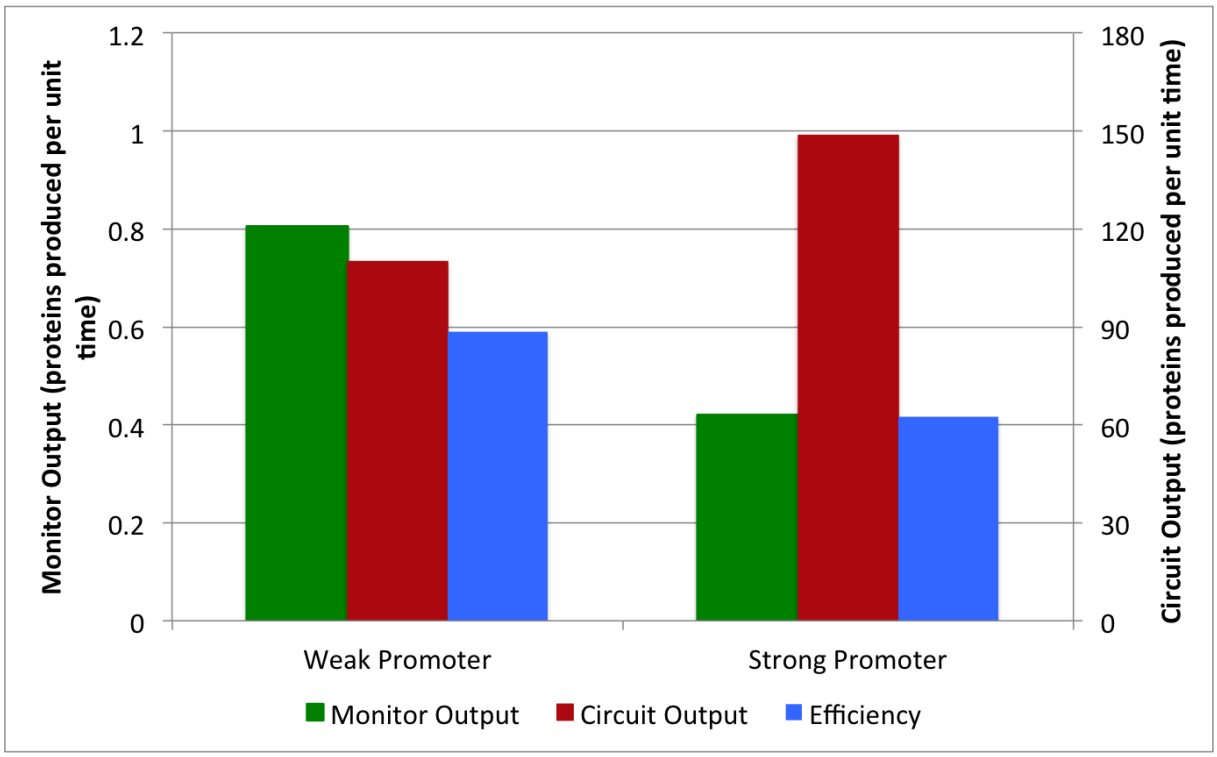


Figure 6.2: Modelled impact of transcript number on circuit and monitor output. This figure shows that a higher number of transcripts causes higher circuit output and lower monitor output.

6.4.2 RBS Strength and Codon Usage

We modelled the system with a range of different RBS strengths as well as two different codons usages. The fast codon version has uniform elongation rates of 1 along a 100 codon transcript and the slow codon version has a uniform elongation rates of 1 along a 100 codon transcript with the exception of elongation rates of 0.5 for codons 85 to 95.

Both codon usage and RBS have a large impact on the behaviour of the circuit. Figure 6.3 shows how both codon usage and RBS strength affect the monitor output and circuit output. For both codon usages, as the RBS strength strength increases at low levels (< 0.4) the relationship between circuit output and RBS strength is approximately linear. As the RBS strength continues to increase, the circuit output reaches a saturation level. Slower codons heavily impact the maximum output of the circuit. This is due to slower codons imposing a lower maximum flux of ribosomes through the system. Also, for slower codons this saturation is reached at lower RBS strength. This intuitively makes sense since slower codons will cause a lower maximum flux through the system and a higher rate of recruitment of ribosomes onto the transcript will cause this maximum to be reached.

In terms of monitor output, for RBS strengths < 1, the relationship between RBS strength and monitor output is approximately linear. For higher RBS strengths, the monitor output tends to a lower asymptote. The slower codon circuit causes a decrease in monitor output.

The yellow lines represent the time point at which the data represented in Figure 6.4 are considered, while the dashed blue line represents the time point at which the data represented in Figure 6.5 are considered. We can see that this data qualitatively matches the results seen in the wet-lab in Figures 5.15 and 5.17. However, we are unable to capture the phenomenon of reduced circuit output at the highest RBS strengths. This is because we are not including cellular response and adaptation in this model, though it would be a very interesting thing to include into the model, which we plan to do in future work.

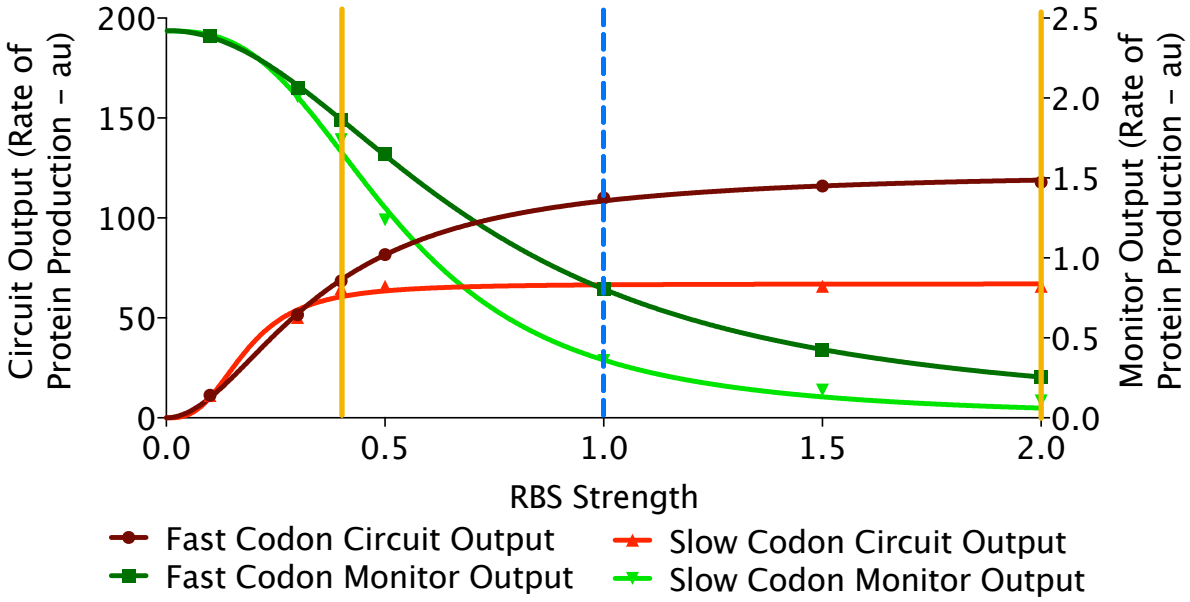


Figure 6.3: Modelled impact of RBS strength and codon usage on circuit and monitor outputs. Shows both monitor output and circuit output for a range of RBS strengths for two different codon usages. Lines represent best fit of hill curves using GraphPad Prism with no parameter constraints.

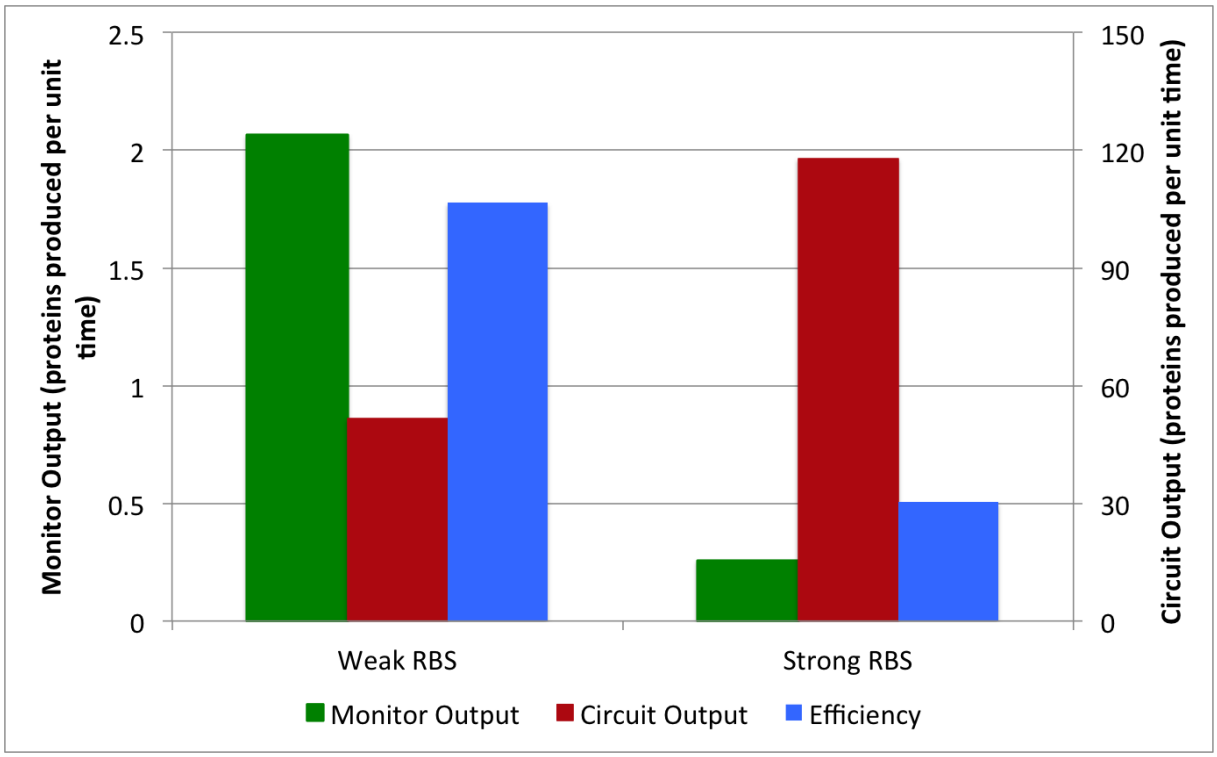


Figure 6.4: Modelled impact of RBS strength on circuit and monitor output shows that a stronger RBS causes higher circuit output and lower monitor output.

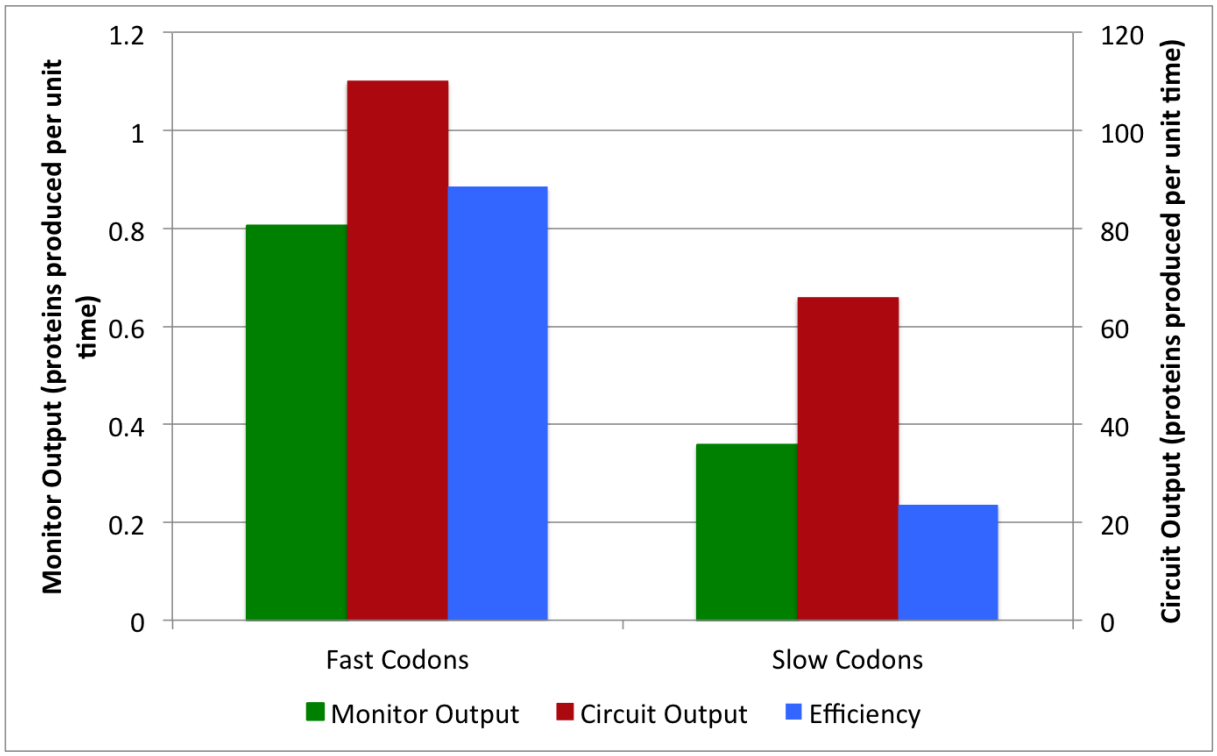


Figure 6.5: Modelled impact of codon usage on circuit and monitor output shows that slower codons in the circuit cause lower circuit output as well as lower monitor output.

6.5 Obtaining Similar Circuit Output with Different Burden Levels

A particularly interesting result from the wet-lab data was to see that it was possible to design two circuits with similar circuit output but different monitor output (resource usage). This was done by changing both the RBS and promoter together so that in one case a stronger promoter was used with a weaker RBS and in the other case a weaker promoter was used with a stronger RBS.

We simulated this using RBS strengths and transcript numbers that were above the regions where we saw a proportional behaviour between the variables and circuit output. The weak RBS had strength 0.4 and the strong RBS had strength of 2, a 5-fold difference (approximately the same difference as predicted by the Salis RBS calculator for the two RBS strengths used in the wet-lab experiment). The number of transcripts used was 300 for the weak promoter and 500 for the strong promoter. This is approximately the same difference in output that we saw when the two P_{BAD} promoters were characterised and the values correspond to relatively strong promoters on high-copy plasmids.

Figure 6.6 shows the data obtained from the modelling and shows that the circuit output for the strong promoter/weak RBS and weak promoter/strong RBS is approximately the same, however the output from the capacity monitor is higher for the strong promoter/weak RBS version. The weak promoter/weak RBS construct has the lowest circuit output and highest monitor output whilst the strong promoter/weak RBS construct has the highest circuit output and lowest monitor output.

Comparing these results to the wet-lab data shown in Figure 6.7 we can see that whilst there is not an exact quantitative match between the simulation and the experimental data, there is definitely a qualitative match. This shows that our model is able to predict some of the more unexpected and interesting behaviours that arise from changing control points together. It also means our model may be used to help design constructs in a way that minimises burden.

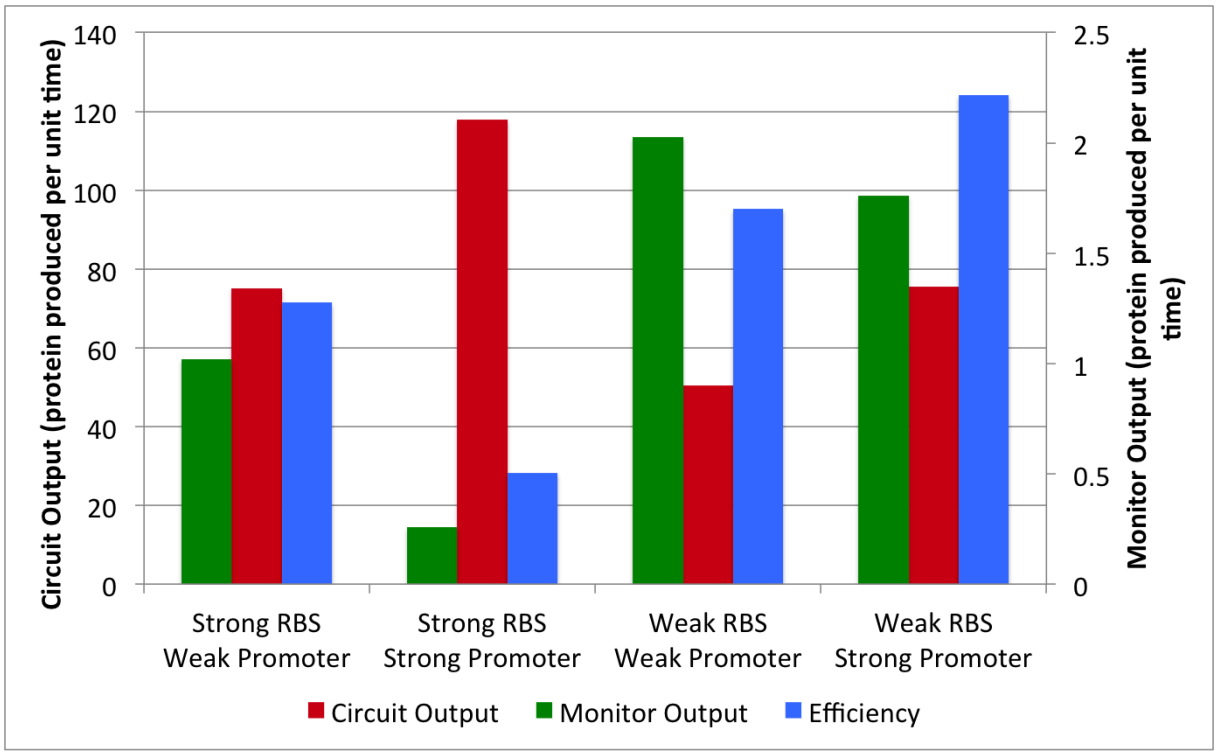


Figure 6.6: Obtaining Similar Circuit Output with Different Burden Levels - Modelling

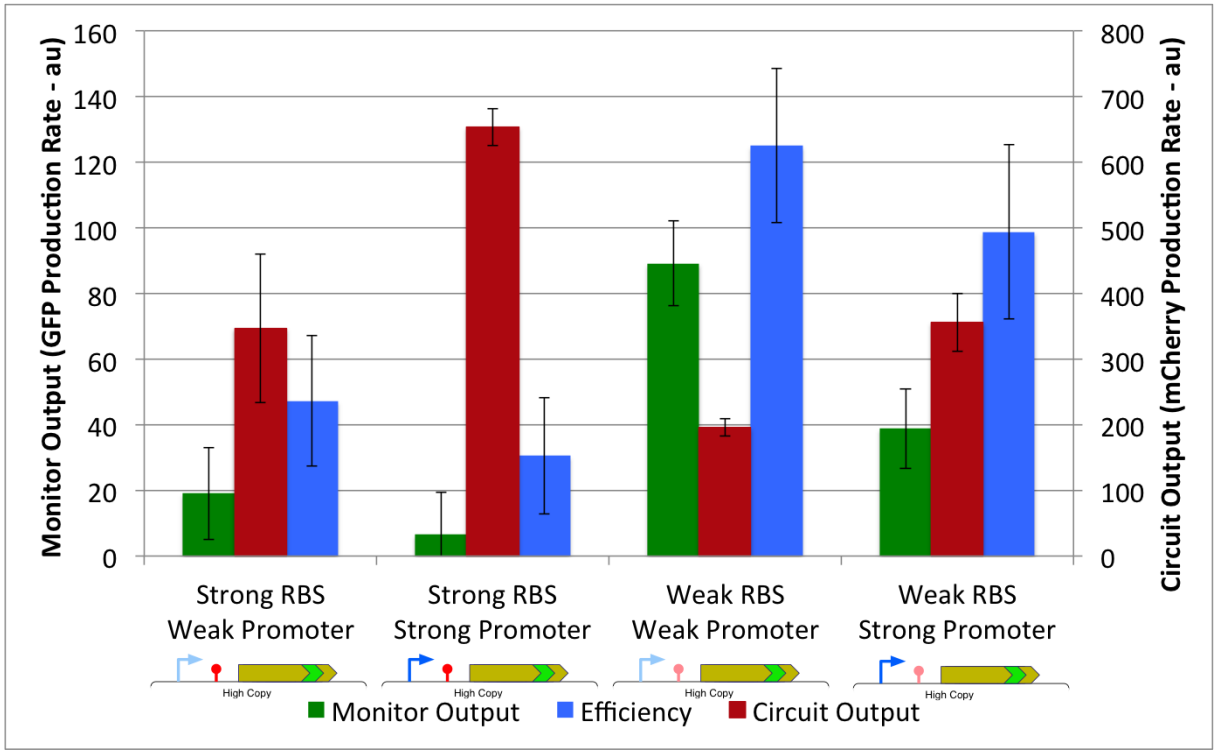


Figure 6.7: Obtaining Similar Circuit Output with Different Burden Levels - Experimental

6.6 Optimising the Monitor

Since our monitor was designed before any of the modelling and simulation of control point experimental testing is it likely that it is itself not the optimal design. We wanted to investigate how we might be able to obtain better performance by changing the design. Our results show that the faster the codons in the coding region, the less burden and the higher the output will be for a circuit. The capacity monitor sfGFP has already been codon optimised so it is unlikely that the codon usage can be improved upon.

The RBS we chose for the circuit was designed by maximising the predicted strength on the Salis RBS calculator. Figure 6.8 shows how the RBS strength affects the sensitivity of the monitor to changes in ribosomal availability. It is clear that the higher the RBS strength, the less sensitive the monitor will be to changes in ribosomal availability, especially at higher levels. It is also important to take into consideration the slope of the curve, since for a shallower slope the noise in the monitor output may make it difficult to accurately calculate the amount of free ribosomes available in the cell.

From the experiment data shown above we are confident that our monitor is certainly sufficient for the type of investigation that we have performed in this project, however if it were to be used in specific industrial and biotechnological contexts, these are important considerations that should be taken into account when improving the monitor.

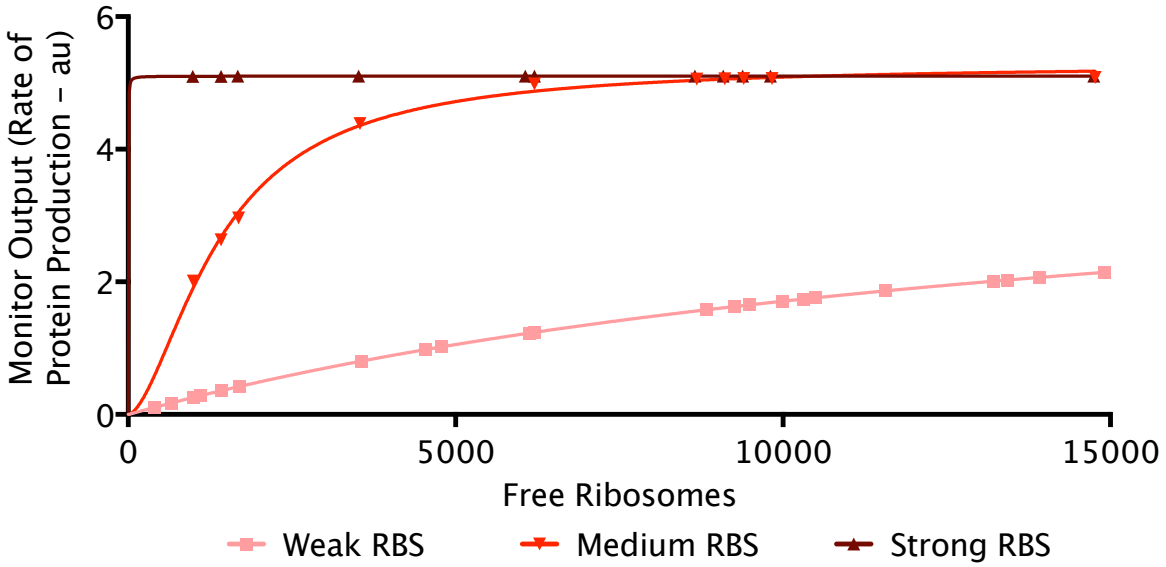


Figure 6.8: Affect of RBS strength on monitor sensitivity shows the monitor output levels for different free ribosome numbers for different monitor RBS strengths. Lines represent best fit of hill curves using GraphPad Prism with no parameter constraints.

6.7 Conclusion

In this chapter we have shown the development of a model of translation. This has been designed so that it is able to incorporate the effect of codon usage on gene expression as well as ribosomal usage. This model assumes that the competition for transcriptional resources is less important than the competition for translational resources and can be neglected.

We used a random walk approach to model the behaviour of ribosomes within a cell moving from a free pool and reversibly binding to a transcript before moving unidirectionally along the transcript. This approach was then developed into a deterministic steady-state model using expectations. We then proved that there was a unique solution to these equations. However, unfortunately for any realistic circuit we might want to model these equations are not analytically solvable and therefore we must use numerical methods to simulate their behaviour.

We provide a python script that is able to simulate a cell with an arbitrary number of mRNA species where the length, codon speed, RBS strength and number of transcripts call all be defined. We then ran a simulation of this with biologically realistic numbers and obtained outputs that were within realistic bounds.

This model was subsequently used to predict the impact of changing the number of transcripts

(to reflect a change in copy number or promoter strength). These results showed that there are diminishing returns for protein production levels as transcript numbers are increased and that both monitor output and circuit output tend towards asymptotes. The simulations also qualitatively reflected the results observed in the experimental data.

The model also predicted that increases in RBS strength would lead to saturating increases in circuit output as well as decreases in monitor output. Unfortunately, since the model did not include any cellular feedback we were not able to observe a decreased circuit output for particularly high RBS strengths. We also observed that introducing slow codons into the transcript caused a decrease in both monitor output and circuit output, reflecting the experimental results.

Crucially, the model was also able to reflect the ability for two circuits to have the same circuit output whilst causing different levels of burden. A construct with low transcript numbers and high RBS strength (weak promoter, strong RBS) was shown to cause a higher level of burden than a circuit with a higher number of transcripts (stronger promoter) and weaker RBS that gave the same circuit output. This shows that our model may be used to uncover additional non-intuitive circuit behaviours.

We also modelled the impact of different RBS strengths for the circuit monitor to test what the impact of this might be on the sensitivity of the circuit to changes in ribosomal availability. We see that increased RBS strengths are predicted to make the circuit less sensitive to changes in ribosomal availability. However, the slope of the curve and the ability to accurately identify differences in monitor output when noise is considered should be thought about in the experimental implementation of this result.

Chapter 7

Conclusion and Discussion

7.1 Overview

The overriding aim of this project was to gain a greater understanding of the interactions between a synthetic cell and its host chassis cell. The work done has focused on the interactions through shared resources. Here we defined shared resources as the cellular machinery and building blocks such as ribosomes, polymerases, amino acids etc. Looking back on the objectives of this project (Section 1.2) we note that there were three core modules to this project.

7.1.1 Module 1: Capacity Monitor

We were able to successfully design and implement a device that monitored the cell to detect the production of heterologous protein as well as cellular adaptation to media shifts. A key question arising from this is whether it is actually the capacity for the cell to produce additional proteins that we are quantifying.

Detecting Capacity

The device expresses codon-optimised superfolder GFP from a fully-synthetic constitutive promoter controlled by a synthetically designed RBS. Since our device is constitutively expressed, we expect that any changes in the expression rate are due to changes in the global shared

resources rather than any specific regulatory factor. Therefore we argue that the monitor is detecting the burden that additional protein is putting on the shared resources of the cell.

We do not expect that there is a direct proportionality between the output of the monitor and any particular resource (though we do believe that a key factor impacting the monitor output is ribosome availability). However, a greater output from the monitor would indicate that there are more resources available within the cell and therefore if genes were added to the synthetic circuit, a higher monitor output would indicate that these genes would be expressed at a higher level.

Compatibility with Synthetic Circuits

The monitor device we built was integrated into the genome of *E. coli* DH10B and MG1655 at the λ-site. This genomic integration means that there are no compatibility issues with any other (non-λ) integrations or any plasmid origins of replication. However, the CRIM[?] integration methodology leaves the resistance marker of the CRIM plasmid within the genome. In the case of the λ-site CRIM plasmid (pAH63) the resistance marker is kanamycin. This means that our capacity monitor is not compatible with synthetic circuits that use kanamycin as either part of the circuit or as part of the DNA used to introduce them to the cell.

The use of GFP as the reporter protein means that the device is incompatible with synthetic circuits that use GFP. This is potentially quite a large issue, due to the prevalence of GFP as a reporter protein in synthetic biology. However, the functionality of the monitor device does not depend on the excitation/emission of the protein used as a reporter and therefore implementing a similar device (or range of devices) with different fluorescent proteins (that fulfil requirements such as long half-life etc) would allow a larger range of synthetic circuits to be characterised using capacity monitors. This is potential further work that could be done for this project and is discussed in more detail later.

Copy Number and Degradation Tags

We investigated a rage of 12 potential monitor devices and characterised the production rates and degradation rates of GFP across them as well as their impact on the growth rates of the

cells. It was decided that the optimal design was the device with no degradation tag integrated into the genome.

By integrating the monitor onto the genome and having it at a single-copy, we have shown that we are able to avoid any large impact on the cell (as indicated by growth rate). This has been done whilst maintaining the ability to detect GFP at levels that allow us to accurate estimate the production rate. GFP production rates are most accurately estimated for proteins without degradation tags. We have shown that tagging GFP with SsrA degradation proteins leads to unpredictable and unreliable protein degradation rates and we therefore chose to avoid using them in our device. These results are due to competition for the native cellular degradation machinery and have been shown in the literature[].

Testing the Monitor

We performed a number of tests to confirm that changes in the rate of protein output from our monitor were due to changes in the amount of shared protein expression resources. By inducing the expression of protein from a synthetic circuit we confirm that the production of heterologous protein is causing the reduction in rate of protein output from the capacity monitor device. By controlling the time of induction and the carbon source in the media we are able to identify that the monitor device is able to detect the usage of shared resources due to a number of factors.

The capacity monitor is able to detect decreases in shared resources from both the production of additional heterologous protein from a synthetic circuit. When the cells underwent a diauxic shift we observed an additional decrease in capacity monitor output when the cells adapt to a new carbon source.

A key result from this module is that we were able to develop both a device for monitoring the capacity in the shared resource pool and a protocol for characterising circuits in terms of their impact on shared resources, as well as the cell's response to additional stresses.

7.1.2 Module 2: Investigating the Impact of Various Control Points

The aim of this module was to gain a greater understanding of how changing key genetic control points affects both the protein output and resource usage of a synthetic circuit. The control points we investigated were the copy number, promoter strength, RBS strength and codon usage. By successfully creating a combinatorial library of constructs with all of these control points varied we were able to gain an insight into the impact of changing these control points both in isolation, and in tandem.

Characterising this library of constructs in DH10B *E. coli* cells was done by estimating the GFP production rates from the capacity monitor using regular (every 10 minutes) readings of both OD 600 and GFP fluorescence. The results have shown there is a complex relationship between changes in the genetic control points for a circuit and its behaviour in terms of both output and the burden it places on shared resources.

Promoter Strength

We used two strength P_{BAD} variants - a wild-type 'weak' promoter (this is weak relative to the modified variant, though P_{BAD} is a strong promoter when compared to most native *E. coli* promoters) and a modified 'strong' version. Changing the promoter strength alters the amount of mRNA that are transcribed from each promoter. As would be expected, an increase in promoter strength causes an increase in the output of the circuit as well as a corresponding decrease in capacity in the cell.

RBS Strength

Increases in RBS strength cause increases in the amount of burden a circuit places on the shared resource pool. At lower RBS strengths, these increases also correspond to an increase in the protein production rate from a circuit. However, above a certain point increases in the RBS strength start to cause decreases in the protein production rate. This is a result that has been reported elsewhere in the literature and is likely due to cells adapting to the depleted resources. The continued decrease in capacity is likely due to the fact the cell is using more ribosomes to produce proteins to change its proteome to adapt to the burden caused by the

synthetic circuit.

When characterising the different constructs in our library we observed that there was significantly higher leakage for constructs with the medium strength RBS when the promoters were uninduced. The fact this occurs with the medium strength RBS but not the strong or weak RBS suggests that it is not a function of the RBS strength. The reason for this is that if it was a function of the strength we would also expect the same (or more) leakage from either the strong or weak promoter as well as the fact that the RBS strength only matters post-transcription when translation is being initiated. Therefore this behaviour is most likely due to sequence-specific interactions whereby the sequence of the medium strength RBS helps recruit additional polymerases to the promoter. This could be tested by designing RBS sequences with similar strength to this RBS but with different sequences. If the leakage is due to interaction between the RBS sequence and the promoter we would expect to see similar levels of induced output but significantly lower leakage.

Copy Number

The copy number of a circuit is the average number of copies of the circuit DNA per cell. For genomic integrations this number will depend on the rate at which the cell is growing and the location on the genome with slower growth meaning there are less copies per cell and the closer to the 50' genome location, the lower the copy number. For larger copy numbers in *E. coli*, plasmid systems are often used where the copy number is controlled by the origin of replication used. In order to maintain the plasmid in cells a selection marker is used, often an antibiotic resistance gene. Since there is a copy of the resistance marker on every plasmid, the amount of protein produced increases with the copy number as does the amount of shared resources required to maintain the plasmid. Our results show that the output from a circuit on a high-copy plasmid (100-300 copies per cell) is only marginally higher than the same circuit on a medium-copy plasmid (10-12 copies per cell). The growth rate and circuit output are also higher for the medium copy plasmid. This is also the case when the circuit is uninduced, indicating that the resources required to maintain the plasmid are significantly higher for the higher copy plasmid.

Codon Usage

Another crucial control point is the codon usage in a protein coding region. The sequence of a coding region is important for a number of factors. The sequence of the 5' region is important because of two main reasons. Firstly, the sequence impacts the formation of secondary structures with the RBS that might cause it to be blocked and therefore affects the rate at which translation initiates and this is something that is taken into account in the RBS 'strength'. The second main reason is that slower codons at the start of a transcript mediate the rate at which ribosomes start moving along the mRNA in the elongation stage. This 'on ramp' helps minimise traffic jams caused by slow codons further along the transcript.

In addition to the 5' region, the sequence of the transcript is important as it dictates the rate at which ribosomes move along it. There are two reasons for this given in the literature. Firstly the codons used to encode the amino acid sequence of the protein affect which transfer RNAs (tRNAs) are used to recruit amino acids into the elongating peptide chain. Different tRNAs occur at different concentrations within the cell and their relative abundances mean their amino acids are incorporated into the elongating peptide at different rates. Certain codons in *E. coli* are considered to be 'slow codons' and are translated at a slower rate relative to other codons. Secondly, certain sequences known as anti Shine-Dalgarno sequences have a high affinity for the 16S ribosomal RNA of the translating ribosome which causes a decrease in the rate at which the ribosome is able to elongate past this sequence[]. The locations of slow codons and anti Shine-Dalgarno sequences are highly correlated as the anti Shine-Dalgarno sequence contains the same sequence as a number of slow codons.

In this study we have not made the distinction between the two factors when trying to introduce a delay in the translation rate towards the end of the VioB sequence. We introduced sow codons as well as anti Shine-Dalgarno sequences. Our experimental results show that the introduction of these sequence motifs towards the end of the VioB causes both decreased capacity as well as decreased circuit output. This is because the slow codons (a term we use to describe a sequence that contains both slow codons and anti Shine-Dalgarno sequences) are causing a decrease in the flux of ribosomes through a certain point (or set of points) along the mRNA. Each protein produced corresponds to a ribosome moving fully along the mRNA and translating an entire peptide chain and therefore, decreasing flux of proteins along the mRNA decreases the rate at which proteins are produced. Also, if ribosomes are being recruited onto the mRNA

at a higher rate than they are moving past the slow codons then this will lead to ribosomes being in 'traffic jams' at the slow codons due to a bottle neck in the flux.

The relationship between the strength of the RBS and the minimal translational rate across the mRNA (due to codon usage/anti Shine-Dalgarno sequences) impacts heavily on the efficiency of the transcript. When the RBS strength is low, the rate at which transcription is initiated is the limiting factor in protein production. This also means that traffic jams occur with much lower frequency. We hypothesise that for a given coding sequence (in a given set of growth conditions) there is a threshold for RBS strength above which, the codon usage and presence of anti Shine-Dalgarno sequences become the limiting factor in protein production rate and 'traffic jams' increasing occur, thus decreasing the efficiency of the circuit.

Whilst slow codons appear to be a poor choice in circuit design, they can be useful in ensuring the correct folding of multi-domain proteins[]. Therefore, in some circumstances their use may be required. In this case we suggest that the circuit should be designed so that the RBS strength is at, or just below, the threshold mentioned above and the correct level of expression should be controlled via the promoter strength and copy number.

The DH10B cells used in most of this project do not have the stringent response phenotype, by performing some of the same experiments in MG1655 cells we were able to observe the behaviour of wild-type cells in response to burden. We saw that MG1655 cells had a much larger decrease in capacity when heterologous protein production was induced. This may have been due to the stringent response allowing the cells to detect the production of extra protein and adapt to cope with this by down regulating the monitor promoter. We observed that slow codons are also a poor design choice in MG1655 cells, indicating that they should be avoided (where possible) across different strains of *E. coli*.

Growth rate is frequently used as a proxy for the 'health' of the cell, however our results have shown there is weak correlation between this and metrics such as cellular capacity or circuit efficiency. Given that we are using a more direct system for observing resource availability we argue that our system is a better method for observing the 'health' of a cell in the context of cellular capacity and resource availability.

7.1.3 Module 3: Modelling the Interactions

Both the literature and wet-lab results indicate that the key factor in cell-circuit interactions through shared resources is ribosomal availability[?]. Therefore we develop a model of gene expression that focuses on the translational process and includes elongation rate. The model was derived as a Markovian random walk process and developed into a deterministic steady-state model using expectations. This model enables us to simulate the ribosomal density across transcripts and in the free ribosome pool.

The parameters in the model allow us to control the number of transcripts (to represent a change in the copy number or promoter strength), the length of transcripts, elongations rates at each codon, RBS strength (RBS-ribosome binding and unbinding rates) and total number of ribosomes. When these parameters are given experimentally realistic values we are unable to solve the equations analytically and therefore we use numerical analysis to solve them. When we input example values for these parameters that we might expect *in vivo* we observe a realistic output which gives us verification of the scaling of the model.

We simulate a system with a burden monitor device and a synthetic circuit and change the parameters to reflect the experiments performed *in vivo*. The simulations closely resemble the experimental data. Some of the areas where there were differences were at higher RBS strengths. The experimental data shows that at high RBS strengths there is a decrease in the protein output, most likely due to the host cell adapting to cope with the depleted resources. The cellular behaviour and any feedback is not modelled in our system and therefore these behaviours would not be expected to manifest in the simulations of the model.

The model is able to not only predict the impact of changes of single parameters at a time, but is also able to capture more complex interactions across multiple control points. We were able to experimentally show that by balancing promoter strength and RBS strength we are able to construct two circuits that have the same output but place different levels of burden on the resource pool. By simulating this experiment in the model we get the same results whereby a strong promoter and weak RBS combination is more efficient than a strong RBS and weak promoter combination. Due to a lack of knowledge about the exact parameter values we did not get a fully quantitative match, however we were able to get a strong qualitative match in terms of both monitor output and circuit output across all four RBS/promoter combinations.

This model is clearly a very simplified way of representing gene expression and the interaction with shared resources. A number of assumptions are made in the derivation of the model and may impact on our ability to accurately model the cellular system. One assumption that we know to be incorrect is that each ribosome occupies a single codon on the mRNA. In reality a ribosome occludes a space of approximately 13 codons[], however the core dynamics of the process are unaltered by this assumption and we are able to much more easily simulate behaviour to a qualitatively /accurate level.

We know that occasionally ribosomes can terminate translation early and exit the mRNA before reaching the stop codon. This is not allowed in our model, and indeed it may be the case that ribosomal traffic jams encourage early termination and therefore less ribosomes are sequestered in these situations than predicted by our model.

Another assumption we make is that the position of a ribosome at a given point in time is independent of the position of all other ribosomes. This is an important assumption in the derivation of our model as it allows us to simplify many of the probability functions of ribosomal movement. The implication of the removal of this assumption would be that there would be a lot more recursive relationships within the model. The full model is of a similar form to one derived elsewhere in the literature using different techniques[**?**], which implies this is a valid assumption to make.

This model was implemented in a python script which allows users to simulate the behaviour of cells with any number of circuits. Users are able to define all of the key metrics for both the circuits and the cell and simulate the behaviour using a simple piece of code. This could be easily implemented in CAD or modelling software and can easily be expanded to include additional factors.

7.2 Overall Conclusions

In this project we have investigated how synthetic genetic circuits and host chassis *E. coli* cells interact through shared resources. By inserting a constitutively expressed GFP into the genome we have created a device which is able to monitor the cell's capacity for gene expression. The device was tested and shown that it can detect the burden placed on a cell's shared resources both from heterologous gene expression as well as the cell producing additional protein to

adapt to a change in growth media.

We created a library of sequences in order to test the impact of changing the key control points of synthetic circuits. We established how changes in these control points affected the circuit output and the burden it placed on the cell. An important finding was that slow codons are detrimental in terms of both of these factors. We established certain design principles that allow circuits to be designed to have the same output but place different burdens on the cell by using a combination of weak RBS and strong promoter. Our results confirmed that translation was the limiting step in gene expression for the circuits we investigated. We also showed the relationship between growth rate and cellular capacity has only a weak correlation.

By building a model of translational processes we have been able to reproduce almost all of the experimental data. In addition we have been able to predict the impact of mRNA levels and RBS strength are 'saturating'. This means that both burden levels and circuit output tend towards asymptotes as these variables are increased. We have built a programmable implementation of the model (using python) that could easily be used as a package in biological CAD/modelling software. The model has also been used to identify how future versions of the monitor might be designed by changing the RBS strength to affect the 'sensitivity' of the monitor.

7.3 Future Work and Implications

There is a range of further work that could be done to expand and complement the results of this project. This ranges from performing some additional characterisation experiments to the inclusion of additional considerations such as noise.

7.3.1 Improving the Capacity Monitor

The final monitor device design was decided upon after a number of versions were considered with different degradation tags and copy numbers. The core design such as promoter, RBS sequence and coding sequence were decided upon by simply trying to maximise the rate of protein output from the device and no variations of these were tested. This project has shown clearly that this approach is not optimal for maximising the efficiency of the circuit. It would be interesting to look at how the core monitor device could be redesigned by considering the

results of this project. We have seen that our model may be able to help guide this design taking into account factors such as sensitivity to changes in free ribosome numbers.

The capacity monitor has been introduced into the cell via a genomic integration into the λ-site where a *pit*-dependent origin of replication and a kanamycin resistance marker are also present from the CRIM system[]. The kanamycin resistance marker causes constitutive expression of *KanR* protein, which has two main effects. Firstly the production of any protein requires shared resources and since the capacity monitor should have as little impact on the cell's shared resources, the presence of any unnecessary gene is not ideal. Also, by having a kanamycin resistance gene on the genome cells are incompatible with any synthetic circuits that use kanamycin to maintain their presence (i.e. plasmids with kanamycin resistance). Genomic insertion of the core monitor device without the origin of replication of resistance marker would be an important piece of work when improving the monitor.

The choice of sfGFP as the reporter protein in the monitor device gives a number of important advantages such as having a well characterised protein that can be used in many synthetic biology labs. However, this means the monitor is incompatible with any synthetic circuit that uses GFP. Compatibility could be increased by implementing a library of monitor versions with a range of alternative fluorescent proteins with different emission and excitation wavelengths. If these monitors could be collaborated against each other it may be possible to characterise and compare any selection of genetic circuits.

7.3.2 Additional Growth Conditions and Stresses

The characterisation of the library of constructs was performed under the same growth conditions growing in volumes of 200 μl in a 96-well plate with M9 media supplemented with 0.4% fructose at 37℃. These conditions are very specific and it would be interesting to test how circuits behave under different conditions. Interesting conditions to test would be with different media such as a different carbon source or amino acid composition. By changing the amino acid availability, the rate at which elongation occurs may change, though the rate of initiation might stay the same. This would impact the balance between RBS strength and codon usage (something our modelling has shown is key) and may change whether the RBS strength or elongation speed is the rate limiting factor in protein production rate.

Additional experiments could also be done to test how the capacity of cells is affected when they undergo stress. This stress may come in different forms such as temperature stress, oxidative stress, nitrogen starvation etc. It would be interesting to look at 'steady-state' growth in these conditions as well as observing how the capacity of the cell behaves when a stress is put on the cell (such as a shift in the carbon source as shown in Chapter 4). It would be interesting to observe how well cells containing different versions of the test circuit are able to adapt to cope with these stresses and see if we can predict the nature of these adaptations, an interesting question would be if circuits that cause higher levels of burden cause cells to take longer to adapt to stressful conditions.

7.3.3 Testing in Additional Strains and Organisms

All of the experiments performed were in *E. coli* DH10B, with the exception of a comparison with MG1655 for a selection of circuits. Cardinale et al. have investigated cell-circuit interactions by investigating how a synthetic circuit behaves in different cellular contexts. Testing the construct in different strains of *E. coli* would enable us to understand the relative differences in native capacity, i.e. how much of the shared resources can we use before adversely affecting the cell? Different strains may also affect how different control points relate to each other, for example a cell with a higher native abundance of charged tRNAs may cause a different interaction between the RBS strength and codon usage in a synthetic circuit.

In addition to working with *E. coli* it would be interesting to implement a similar burden monitor in other bacteria such as *Bacillus subtilis* to investigate the different capacities and cellular responses to the addition of genetic circuits. This should be a relatively simple expansion, as long as the species the monitor is implemented in have reasonably simple ways of introducing a circuit into the genome. Attempting to implement a capacity monitor in eukaryotes such as *Saccharomyces cerevisiae* would pose additional challenges, however it would be very interesting to observe the implications of doing so.

It may be possible to build a controllable burden inducing device which can be used to impose a range of different definable burden levels on the shared resource pool. This would allow cells to be characterised in terms of how much native capacity they have as well as how they respond to a range of burdens. This information can be used with circuit characterisation data

in an extended model of cell-circuit interaction to predict how separately characterised cells and circuits might behave together. This device could be a simple inducible promoter driving the production of RNA that sequesters a defined number of ribosomes.

7.3.4 Expanding the Test Construct Library

The library of constructs we produced contained 2 promoter variants, 2 plasmid backbone variants, 3 RBS variants and 2 codon usage variants. This gave a total of 24 constructs in our library with which we were able to gain an understanding into the impact of changes in these control points. From the data obtained in characterising these circuits with the capacity monitor we were able to get a good insight into the way in which synthetic circuits interact with the shared resource pool. However, obtaining data for a broader range of variants of these control points could give us a finer, and more detailed, insight into how these control points affect circuit behaviour. We are able to use our model to get curves of circuit output and monitor output across a range of RBS and promoters strengths and it would be useful to get the corresponding *in vivo* data to make a more informed comparison. Different codon usages could be investigated by introducing 'on ramps' of slow codons towards the start of the transcript as well as changing the location and length of the slow codon regions. Using different plasmid backbones would allow us to investigate the implications of different origins of replications and selection markers on cellular capacity and might help us develop plasmids that are better optimised for minimising the impact of a circuit on shared resources.

All of the constructs we built used the VioB protein. This is a limiting factor in the sense that the decrease in monitor output may be a function of the fact we are using VioB. This is unlikely as we selected this protein for its orthogonality to the cellular metabolome and the fact it is non-toxic. Constructing alternative test circuits with different proteins would allow us to confirm that the behaviour we see is not due to the VioB protein (we see evidence of this from the circuits used in Chapter 4, though it is not directly comparable). Use of alternative proteins would also up greater potential genetic space for testing different codons usage profiles.

7.3.5 Using the Capacity Monitor to Predict Additional Circuit Behaviour

In this project we used the capacity monitor to provide a proof of principle that we can observe differences in resource availability within *E. coli* cells and to gain a greater understanding of the impact of different design choices on circuit output and resource usage. The monitor may also be used to predict how additional circuits or genes would be expressed when introduced in combination with a characterised circuit. For example take synthetic circuits X and Y that have been characterised with the capacity monitor and an additional gene (or circuit) Z. If cells containing construct X show a greater output from the monitor than cells containing Y, would gene Z be expressed at higher levels in the former when compared to the later. Such a utility for the monitor would be very useful when designing larger and more complex circuits.

7.3.6 Expanding the Concept of Optimisation

As mentioned in the introduction, until recently the concept of optimising gene expression was largely considered in the context of maximising protein output. We have seen in our results that another important consideration is the amount of resources a circuit uses. By comparing the amount of resources a circuit uses and its output we have developed a metric which we call 'efficiency' and gives an insight into how much resources a circuit is using to provide a given rate of protein production.

Whilst this is a crucial consideration, we acknowledge that there are other factors we have not considered that might need to be taken into account. For example, the noise in the output of a circuit and the cell to cell variation in its behaviour might be important factors in its application. There are not factors we have considered and when we compared two circuits with the same protein production rate but different burden levels we claimed that the one with the lower burden was more optimal. This is only the case if the aim is to reduce resource usage and it may be the case that the circuit that caused higher burden had less cell-cell variation, which might be a more important consideration in some situations. Investigating how these additional metrics can be obtained and implemented into our system could broaden the scope of how people are able to optimise circuits.

7.3.7 Using Growth Rate Decreases in Circuit Design

Tan et al. show that it is possible to create novel functionality in a circuit through interactions between the circuit and the cell through growth rate[]. It may be possible to use this principle to design circuits that have burden causing aspects that will feed back on their own behaviour through a link with growth rate. Biotechnology applications often want to maximise the yield of protein produced from a given amount of input. By slowing the growth rate of cells it is possible to cause a greater 'per cell' accumulation of protein, even if it is being produced at a lower rate. Balancing these factors may mean circuits can be designed to induce a certain level of growth retardation that causes the most amount of protein per cell. This may mean the batch takes longer to grow to the final density, but by diverting more resources to the production of the protein of interest, and away from growth, the total protein yield per unit of input may be greater.

7.3.8 Expanding the Model

The model developed in this project allows us to capture most of the behaviours observed *in vivo*, however there are a number of way in which this model could be extended or adapted to improve it. We have focused on modelling the translational process as this is what both the literature and our own results indicated was the main factor in cell-circuit interactions through shared resources. In addition we have not included any cellular behaviour such as growth rate or any feedback through cellular adaptation.

Extending the model to include important cellular metrics such as growth rate is non-trivial due to the complex relationships between the shared resource pool and growth rate. A deeper study of the literature combined with more experimental work may allow a reliable model of this link to be developed. In addition it may be possible to predict how the cell will adapt to burden being placed on shared resources through the stringent response or other mechanisms. It is unlikely that this will be possible from a first principles approach given current understanding of the cellular processes. However, it may be possible to build a framework whereby cells can be individually characterised and this data fed into a model as mentioned in Section 7.3.3.

As well as modelling the cell's behaviour it is possible to increase the level of detail to which we describe the gene expression process. Including the impact of different growth rates on mRNA

levels (as shown in Klumpp et al.[]) would be a sensible way of extending the gene expression model to include direct relationships with the growth rate. In addition, we can include the processes of transcription and DNA replication into the model to include competition for the resources involved in them. Whilst we have seen that they are not as important as competition for translational resources, their inclusion may allow us to more accurately predict how a circuit behaves.

Further analysis of the model may enable us to uncover more design principles that allow the design of more optimal circuits. We have seen that we are able to accurately reproduce how a strong promoter/weak RBS construct causes lower burden than a weak promoter/strong RBS construct when the constructs have the same output. It is likely that there are additional unexpected interactions between different control point selections we can utilise to build more optimal circuits.

We have made a number of assumptions in the development of the model, some of which we know to be untrue. These were made in order to simplify the model and make simulation easier. One important consideration is the size of the ribosome as this will impact how many ribosomes can fit on a transcript and how many are sequestered in traffic jams. Whilst we have shown that our current model is able to qualitatively reproduce *in vivo* results, a model with a more accurate representation of ribosome sizes may allow more quantitative reliability in predictions. A model incorporating this consideration has already been developed but is not shown in the project.

7.3.9 Other Future Work

We have shown a limited amount of data from quantitative PCR and RNA quantification. This has been very useful as it has allowed us to identify where the main cause of reduced capacity is coming from. However, the laborious nature of RNA extraction and qPCR means that it is very challenging to get all of the data on mRNA levels for the capacity monitor at the same level of quality as protein levels. Finding a better way of quantifying mRNA levels in cells that can be done at the same frequency and integrated into the current protocol would be highly advantageous. One potential method would be to use SPINACH RNA aptamers to quantify RNA levels[]. However, the fact it uses green fluorescence causes a slight issue as well as

the low levels of fluorescence it provides, which may mean getting accurate quantifications is difficult.

A factor we have briefly touched upon in the introduction and Section 5.7.9 is the evolution and stability of constructs. Recent papers by Sleight and Sauro indicate that the burden from gene expression as the key driver of deleterious mutation and construct instability over multiple generations[]. Our project could be used to predict the evolutionary stability and outcomes of synthetic systems based on the burden they cause. There is certainly a large amount of work to be done in solving this problem, however it is an exciting prospect.

It would be an interesting study to take extend both this work and the work of Li et al[] by trying to decouple the slow codons from anti Shine-Dalgarno sequences and investigating their separate affects on shared resource usage. This could be done by changing the 16S sequence and investigating whether similar decreases in cellular capacity could be achieved using only a modified anti Shine-Dalgarno sequence without slow codons and vice versa.

7.3.10 Design Principles

Through both *in vivo* and *in silico* work we have been able to uncover some key relationships between the circuit design and the amount of shared resources it uses. These allow us to hypothesise a methodology for optimally designing a synthetic circuit.

The initial stage in the process is to codon optimise the gene of interest. This should be done by using an algorithm such as the one developed by DNA2.0[] whilst taking into account the potential requirement for slow codons at certain locations for multi-domain proteins if they are required for the correct folding of the protein. Any anti Shine-Dalgarno sequences should also be removed. The impact of slow codons has been shown both *in vivo* and *in silico* in this project.

Once this has been done, an appropriate RBS sequence should be selected. This should be designed so that the RBS strength is such that translational initiation is the limiting factor. This means that traffic jams are avoided and any necessary slow codons do not cause adverse effects by being the limiting factor in translation rate. Dependant on loaded tRNA abundances and the growth conditions (which affect translational elongation rate) the choice of RBS strength may be different, and a weaker RBS may be selected to increase the robustness of the circuit's

efficiency when experiencing changes in growth conditions.

The next stage is to select the plasmid backbone and promoter. We have seen that a lower copy number causes less burden and impacts less on the circuit protein production rate. Therefore a plasmid backbone should be chosen with the lowest copy number so that it places the lowest burden on shared resources. The optimal origin of replication for each copy number and the best selection marker should be investigated in future work. Finally the promoter should be chosen at a strength that gives the desired rate of protein production. If this is not possible then it may be necessary to increase the copy number of the circuit.

We believe this simple approach to circuit design will allow researchers to easily improve their circuit designs and that future work will enable us to move closer to fully optimal designs.