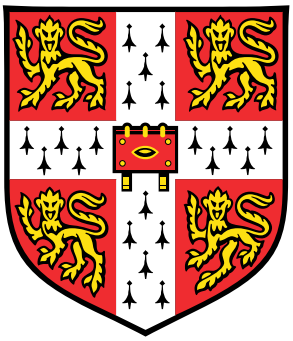
# Establishing Ratiometric Characterisation in *Bacillus Subtilis* for Biosensing Applications

Towards an Arsenic Biosensor



**Haydn James King**

Department of Pathology

University of Cambridge

This dissertation is submitted for the degree of

*Doctor of Philosophy*

Pembroke College

December 2017

Yna heb frys na braw
Llithrodd ei flewyn cringoch dros y grib;
Digwyddodd, darfu, megis seren wîb.

To Mum, who did not go gentle.

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 60,000 words including appendices, bibliography, footnotes, tables and equations.

Haydn James King
December 2017

# Acknowledgements

Firstly I would like to thank my supervisor, Dr Jim Ajioka, for his advice and guidance during the course of my PhD. His help and support has been invaluable and I am deeply grateful for the opportunity to work on this project.

I would also like to thank members of the Ajioka group, in particular Dr Peter Davenport for his expert help and advice dispensed over innumerable cups of tea, as well as Dr Orr Yarkoni for sharing his wealth of practical experience in working with *B. subtilis* as well as with flow cytometry. I owe both these gentlemen a debt of gratitude, as I could not have completed my studies without their help and support.

Next, I would like to thank the many friends I have made while at Cambridge, in particular those from the Pembroke Graduate Parlour who have made my time here so much more enjoyable. My particular thanks goes to Marion, and to those others who have helped brighten the last few months.

That Mum did not get to see me complete my studies was undoubtedly a great disappointment to her, and the thanks that I owe both her and Dad cannot be fully expressed here.

# Establishing Ratiometric Characterisation in *Bacillus Subtilis* for Biosensing Applications

Haydn James King

## Abstract

Arsenic contamination of groundwater remains a serious health concern in many areas of the world. Developing countries such as Bangladesh and Nepal are particularly affected because access to high quality water infrastructure is low. Since the 1970s, most water in these countries is sourced from shallow tube wells installed to reduce the spread of diseases associated with poor water hygiene. In this goal they were successful, however by the mid 1990s it became apparent that many of these wells were contaminated by arsenic and that these countries' rural poor were being slowly poisoned.

No simple, cheap, and reliable test for arsenic exists, and efforts to mitigate arsenic contamination have been severely limited by this over the past two decades. Government backed well-testing efforts using commercially available field kits have many issues with reliability, safety, rigour, and transparency, and have lost their urgency over the past decade, while the expensive field test kits remain out of the reach of most ordinary people in these areas. Synthetic Biology offers the technology to develop a new class of biosensor by exploiting bacteria's natural ability to sense and respond to levels of arsenic considerably lower than commercially available kits which are based on analytical chemistry.

In order to reach this goal, we must first develop our understanding of the natural response to arsenic in our chosen host, *B. subtilis*. Although we have a reasonably good qualitative understanding of the operon responsible for arsenic sensing, very little quantitative analysis has been carried out, and a robust system for ratiometric characterisation has not been established in the bacteria.

In this work, a robust platform for rapid ratiometric characterisation is established in *B. subtilis*. A rigorous mathematical model of the ars operon is developed and analysed before being verified experimentally. This new knowledge is then used to explore synthetic permutations to the natural system aimed at improving the sensor properties of the system. Finally, a biological architecture for an easily tunable biosensor with good characteristics is recommended.

# Table of contents

# List of figures

# List of tables

# Chapter 1

# Introduction

## 1.1 Arsenic

### 1.1.1 Nature and Abundance

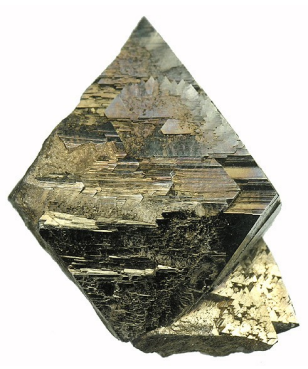
The toxic nature of arsenic has been known since ancient times. The industrial revolution saw a great increase in arsenic extraction, as deposits of the metalloid are commonly associated with copper, lead and gold ores. As production increased, uses for arsenic were found - in France it became known as *poudre de succession* ("inheritance powder"[122]) due to its common use as a poison by the ruling classes. Until the Marsh Test was published in 1836[95], no sensitive test for the tasteless and highly toxic arsenic trioxide [$As_2O_3$] existed, making it a popular murder weapon as the symptoms of arsenic poisoning are similar to Cholera to the untrained eye.

Arsenic commonly exists in four oxidation states: $-3$, 0, 3, and 5. The heavy metal can appear in both organic and inorganic forms, with inorganic forms being highly toxic to life. In aqueous, aerobic environments, arsenic is predominantly found as arsenate [$As^{V}$ as $H_2AsO_4^-$ and $HAsO_4^{2-}$] while in anaerobic environments as arsenite [$As^{III}$ as $H_3AsO_3^0$ and $H_2AsO_3^-$][121].

Natural arsenic is typically found as a mineral in combination with sulphur (for example realgar [AsS] and orpiment [$As_2S_3$]) or also with iron (as in arsenopyrite [FeAsS])[140], substances which while highly toxic, are insoluble. Due to their bright colouration (see Figure 1.1), both realgar and orpiment were commonly traded as pigments within the Roman Empire, as well as being used for fireworks and to kill plants, insects and rodents. Powdered realgar is even added to cereal wine to make "realgar wine", which is traditionally drunk during the Chinese festival of Duanwu[180].

(a) Realgar

(b) Orpiment



(c) Arsenopyrite

Fig. 1.1 Realgar, Orpiment, and Arsenopyrite crystals. Images from Rob Lavinsky/iRocks.com used under CC-BY-SA-3.0

While realgar is insoluble and thus poorly taken up by the gastrointestinal system, realgar deposits also contain up to 10% arsenate and arsenite[180]. Both oxidation states are readily taken up by the body, and are highly toxic. Arsenate, $As^{V}$, is a molecular analogue of phosphorus - arsenic belongs to the same periodic group as phosphorus - and inhibits oxidative phosphorylation, the most efficient process used by almost all aerobic organisms to generate energy from nutrients. Arsenite, $As^{III}$ is more broadly toxic and binds to sulfhydryl groups[109], the functional group found in the amino acid Cysteine, preventing the formation of disulphide bonds important in protein folding and thus impairing protein function.

Despite this knowledge, arsenic contamination still risks the health of tens of millions of people worldwide[122]. Human activity contributes significantly to environmental arsenic concentrations, with arsenic being released into the environment through a number of industrial processes such as smelting, combustion of coal, mining, hide tanning, pigment production and glass production. Inorganic arsenic compounds saw heavy use as pesticides between the 1930s and 1980s, contributing almost 10,000 metric tons per year in the U.S[121]. Prior to the introduction of dichlorodiphenyltrichloroethane (DDT) in 1947, lead arsenate [$PbHAsO_4$] was the primary pesticide used in fruit orchards, but the use of inorganic arsenic based pesticides was not banned until the 1980s and 90s[165]. As well as agricultural uses, arsenic compounds were heavily used as a fungicide for example during Chromated Copper Arsenate (CCA) pressure treatment of wood, a process used to preserve wood exposed to the environment. In the U.S., this process was used for over 60 years and remained the single greatest use of arsenic compounds in the year 2000[165]. In 2003, European Directive 2003/2/EC restricted the marketing and use of arsenic[41], banning CCA treated wood for domestic and residential constructions in favour of safer alternatives such as Alkaline Copper

Quaternary (ACQ). In 2004 the U.S. also began mandating the use of non-arsenic based wood treatments such as ACQ.

Despite these changes, arsenic compounds were still commonly used in the U.S. until more recently. Roxarsone, a derivative of phenylarsonic acid, $C_6\,H_5\,As\,(O)\,(OH)_2$, was widely used as a food additive in the poultry and pig industry as it improved weight gain and feed efficiency as well as acting as an intestinal palliative and improving pigmentation[121]. In 2006, roughly one million kilogrammes of Roxarsone was produced in the U.S.[63]. Since roxarsone is an organic arsenic compound, it is not itself highly toxic as it is not readily absorbed into the body and is thus not a direct risk. However, arsenic containing roxarsone is released into the environment, where various processes are known to convert organic arsenic compounds into highly toxic inorganic forms, and contributing to arsenic contamination of groundwater[29]. In 2013 roxarsone was voluntarily withdrawn by its manufacturers in the U.S., before having its approval for use in feedstocks withdrawn by the U.S. Food and Drug Administration (FDA) in 2015[42]. Roxarsone was never approved for use in the European Union.

### 1.1.2 Contamination of Groundwater

Prior to the 1970s, Bangladesh had one of the highest infant mortality rates in the world. This was due, in part, to poor water purification and the spread of diseases such as cholera as most drinking water was sourced above-ground and access to clean water was poor. UNICEF and the World Bank advocated the installation of tube-wells used to tap into deeper groundwater found in permeable aquifer rock. Since this water was free of contaminants, instances of diarrhoeal disease decreased significantly[22].

The first cases of arsenic induced skin lesions were noted in patients from West Bengal, but by 1987 patients from Bangladesh were also identified[137]. Water sources were eventually identified as the cause of the problem, with most contamination due to shallow tube-wells. By 2001, between 8 – 11 million tube wells had been installed in Bangladesh alone[78], with more than 80% of the population having access to water from tube-wells, ring-wells and taps[151]. Despite knowledge of the problem being widespread by the mid 1990s, new tube wells were not routinely tested for arsenic, leading to the problem being described as the "largest mass poisoning of a population in history"[151].

A 2012 study[44] estimated the total number of deaths caused by arsenic contamination of water sources in Bangladesh to be around 43,000 per year, accounting for up to 15% of all deaths in some districts. The same study estimates that between 1 and 5 million of the 90 million children expected to be born between 2000 and 2030 will eventually succumb to arsenic poisoning.

Ironically, despite the many ways in which humans release arsenic into the environment, most of these instances of arsenic poisoning result from natural arsenic deposits. A tube well typically consists of a roughly 5 cm diameter tube which extends into the ground, usually to a depth of less than 200 m[151]. Water is then pumped by hand from the underground aquifer - a layer of porous rock capable of holding water - up to the surface where it is used for drinking, washing and irrigation. At the time when many of the wells were installed, arsenic contamination was not known to be a problem, and thus routine testing of the newly installed wells for arsenic was not considered.

Indeed, much of the microbiology of how arsenic came to be present in water extracted from aquifers was not known until early in the new millennium, reviewed in Oremland and Stolz [121]. Bacterial oxidation of arsenite to arsenate was first reported in 1918, but this was largely unnoticed until 1949 when Turner [157] isolated 15 strains of Heterotrophic Arsenite Oxidizers (HAOs). HAOs oxidise $As^{III}$ encountered on the cell membrane to the less toxic $As^{V}$. While some HAOs use this as an energy source[159], it is generally considered to be a detoxification mechanism[176], protecting the cell from the more toxic arsenite and making arsenic uptake into the cell less likely as arsenate is strongly adsorbed onto the surface of many common minerals such as ferrihydrite and alumina, reducing its mobility[121].

On the other hand, Chemolithoautotrophic Arsenite Oxidizers (CAOs) are able to translate energy derived from the oxidation of arsenite for cell growth by using $As^{III}$ as an electron donor in the reduction of either Nitrogen or Oxygen, fixing $CO_2$ in the process[120, 176]. As these bacteria can derive energy from arsenic oxidation, it is hypothesised that they attack insoluble arsenic deposits such as realgar [AsS], orpiment [$As_2S_3$], and arsenopyrite [FeAsS] releasing arsenate[121]. However, little of this arsenate is likely to enter aqueous phase due to the aforementioned fact that as $As^{V}$ is strongly adsorbed onto a number of minerals, and thus has low mobility in such environments.

Bacteria are also able to reduce $As^{V}$ to $As^{III}$. It had long been known that many bacterial arsenic detoxification systems reduce arsenate to arsenite in the cytoplasm before expelling arsenite from the cell (discussed in Section 3.1). It was quite a surprise, however, to discover that certain anaerobic bacteria are capable of using $As^{V}$ as their respiratory oxidant, effectively "breathing" arsenate and producing the highly toxic arsenite[1, 85, 154]. These microbes, collectively referred to as Dissimilatory Arsenate-Reducing Prokaryotes (DARPs), have been isolated from freshwater sediments, estuaries, soda lakes, hot springs, and gold mines[119] as well as the gastrointestinal tracts of animals[61] and, crucially, subsurface aquifer materials from Bangladesh[138]. These bacteria have been shown to reduce adsorbed $As^{V}$, which, together with microbial reduction of the substrates such as Ferrihydrite, releases $As^{III}$ into the aqueous phase[115].

While microbes are the agents of the release of arsenite, human activity does play a role. The oxidation of arsenite from minerals by CAOs is aerobic, and is thus accelerated by the digging of tube-wells which provides a source of oxidants in the form of dissolved oxygen and nitrates in agricultural regions. CAOs also produce biomass by fixing $CO_2$, while microbial respiration leads to the depletion of oxygen and the onset of anoxia. In these conditions, DARPs are able to oxidise the build up of organic matter, producing $CO_2$, while respiring $As^{V}$ to produce the more toxic and more mobile $As^{III}$[121].

### 1.1.3 Response to Contamination

Contaminated groundwater was identified as the leading cause of arsenic poisoning in the early to mid nineties. Approximately 97% of the rural population of Bangladesh use groundwater for drinking, domestic and agricultural purposes[66], while larger towns and cities are mainly protected as households there are supplied with piped water. In 1997, Bangladesh's Department of Public Health Engineering (DPHE) began testing rural wells, testing 23,000 tube wells during the year. However, the field tests used were only capable of detecting arsenic at greater than $100\,\mu g l^{-1}$, twice the Bangladeshi drinking water standard of $50\,\mu g l^{-1}$. The commercially available tests became more accurate in the following years, and in 1999 the DPHE in collaboration with the British Geological Survey (BGS) found arsenic contamination above $50\,\mu g l^{-1}$ in 60 out of 64 districts surveyed[78].

Well testing continued into the next decade, with nearly 5 million wells tested between 2000 and 2006 in a joint project between the DPHE and UNICEF[73]. Based on this data, it was estimated that nearly 20% of wells nationally were contaminated with arsenic, exposing 20 million people to drinking water at or above the Bangladeshi limit of $50\,\mu g l^{-1}$[74]. The results of these tests were stored nationally, but tube wells were also painted red or green to indicate whether they were above or below the national threshold[131]. Testing each well once is not sufficient, however, as arsenic can flow within the subsurface meaning that a well which was previously contaminated may later become safe, and a well which was once safe may become contaminated[43].

While the Bangladeshi threshold for considering a water source contaminated or otherwise stands at $50\,\mu g l^{-1}$, there is evidence to suggest that ingesting water with a concentration lower than this can have serious health consequences. The World Health Organisation (WHO) has therefore adopted the lower safe threshold of $10\,\mu g l^{-1}$[44]. By 2010, estimates of the global number of people exposed to unsafe drinking water due to arsenic contamination above this lower threshold had grown to 100 million[43].

In areas of Bangladesh with considerable groundwater arsenic contamination, several mitigation options have been explored. Testing and painting of wells red or green to indicate

arsenic levels above or below the $50\,\mu g l^{-1}$ threshold has been the most effective intervention, as this allows villagers to switch their water source to a safer well. An estimated 29% of all those informed that their tube well had elevated arsenic were able to change their source of water by 2006[2], while well-switching in areas where a safe alternative is available has been observed to be as high as 75%[74]. The installation of deep tube wells has had the second largest effect, with around 12% of those told their water source was contaminated having access to a government installed deep tube well. These deeper wells draw water from deeper, older aquifers which are less likely to be contaminated with arsenic, possibly as a result of the lower incidence of organic matter necessary for the release of arsenite into the aqueous phase present at such depths[43]. While these wells are less numerous and are often shared or community wells, they often prove popular when located centrally, and often provide an impetus for privately owned wells to be re-dug to a greater depth[2].

In 2004, the Government of Bangladesh launched the National Policy for Arsenic Mitigation (NPAM)[54], aimed at forming a national strategy for mitigating arsenic contamination of wells. The policy recognised well-switching as a solution in communities where less than 40% of wells are contaminated, however deep tube wells were seen as a low priority option in favour of recommending a return to the use of surface water – where microbial contamination remains an issue[2].

NPAM recommended five other mitigation strategies which were trialled, but none of which have proven successful to date. The first was the use of shallow dug wells, but this is prone to contamination in the same way as the surface water. Filtration to remove arsenic is also an option, though as of yet it has proven challenging to reliably remove enough arsenic for the water to be safe. Removal technologies depend heavily on water quality, especially the pH[101], and so considerable testing, quality assurance and maintenance is needed for a successful filtration device, all of which are hard to deliver in rural areas. Sand filtration of pond and river water was also put forward, but again, ponds often directly receive human waste from latrines[2], making sand filtration insufficient to remove the contamination. Rainwater harvesting does have some promise during the wet season as plentiful arsenic-free water can be collected. Storage of enough of this water, without contamination, for the entirety of Bangladesh's 8-month dry season is most often not feasible for economic reasons. The installation of piped water systems is NPAM's final recommendation, which, while an attractive solution technologically, has a high financial cost associated with it. Since arsenic contamination disproportionately affects Bangladesh's rural poor, it is likely to be some considerable time before any large proportion of the affected population has access to piped water[2].

Even where installation of the new government-backed deep wells has gone ahead, problems exist. A 2005 survey[75] of 1,060 government installed wells failed to locate some 8% of them, while 36% were found to be no longer functional. Normalising for the range of technologies used across the country, the authors estimate that only about 76% of newly installed wells were working, and even the functional wells often broke down, with government policy leaving volunteers from the local community responsible for their repair.

In addition, of the 125,000 deep tube wells installed between 2007 and 2012, approximately 5% are believed to be contaminated with arsenic above Bangladesh's $50\,\mu g l^{-1}$ threshold[133]. This work was funded by UNICEF, who later performed remedial work on the affected wells, and noted, in communication with human rights group Human Rights Watch:

> An assessment by UNICEF and the Department of Public Health Engineering (DPHE) to determine why some DPHE-UNICEF supported water points were contaminated recognized that a specific clause in the government's drilling contract stipulates 'no success, no payment' which made the contractor liable for arsenic contaminated water points. This clause may have discouraged a transparent and honest process of collection and submission of water samples for testing.

Furthermore, Human Rights Watch's 2016 report[68] on the response to arsenic contamination in Bangladesh found that the siting of new deep wells has not been conducted on the basis of need. Instead, most new wells were located in areas where access to arsenic-free water was relatively good. The report found that up to 50% of new well locations were influenced by politicians rather than public health professionals, diverting vital funds away from the most needy. In addition, wells sitied in areas where hydrological or geological difficulties were encountered were often abandoned, with little to no appetite shown for overcoming these challenges.

The report paints a grim picture of the response to arsenic contamination, which Human Rights Watch describe as "a disaster which humans have caused and perpetuated". The organisation describes the Bangladeshi Government's official response to the contamination issue as 'failing', and argues that efforts to regularly test rural wells have lost their urgency since 2006.

In conclusion, Human Rights Watch make a number of recommendations to both the Government of Bangladesh and its responsible agency, the DPHE, as well as to the international community. HRW recommends that the DPHE should, amongst other things,

> Ensure all district and upazila (sub-district) DPHE offices are provided with sufficient arsenic test kits and are required to regularly use these to monitor and report on arsenic levels in public and private water supplies in the area

and that the international donor community should, amongst other things,

- Support improved government infrastructure for testing tube wells for arsenic
- Ensure independent third party assessment of project results, including random sampling and testing of water quality and rehabilitation or replacement when testing reveals water point contamination, is a mandatory component of any current and future rural water supply projects that receive multilateral or bilateral donor support.

Continued testing of drinking water sources in affected areas in Bangladesh and elsewhere is crucial to overcoming this public health disaster.

### 1.1.4 Arsenic Field Tests

In 1997, the United Nations Children's Fund (UNICEF) together with the World Health Organisation (WHO) took a joint decision to test every tube well in Bangladesh for arsenic contamination in order to better assess the scale of the problem[40, 66]. Due to the sheer number of tube wells, their remote location, and the lack of reputable laboratories within Bangladesh with the capacity to test arsenic, as well as the constraints of time and budget, field tests were the only practical option for this purpose[73].

The first widely used field kit in service during these initial programmes was that developed by German pharmaceutical and chemical company Merck. By 2002, it was estimated that these kits alone had been used to test one million tube wells across Bangladesh[131], and to classify them as containing above or below $50\,\mu g l^{-1}$ of arsenic. These original kits had several issues, not least that their published minimum detection threshold was $100\,\mu g l^{-1}$, double the threshold they were being used for.

All of the early test kits were based on the mercuric bromide stain method[131], also known as the Gutzeit test, as are most of the modern kits. The method is similar to the earlier Marsh test. First, any pentavalent arsenate is reduced to arsenite, $As^{III}$, typically by means of potassium iodide and stannous chloride. Reaction with zinc under acidic conditions then causes the release of arsenic in the form of arsine gas, $AsH_3$. This gas then reacts with a mercury bromide impregnated paper test strip, forming a variety of arsenic salts with mercury, the exact combination of which depends on the concentration of arsine gas, itself dependent

on the original arsenic concentration. The colour ranges from yellow [$H(HgBr)_2As$], to brown [$(HgBr)_3As$], and finally to black [$Hg_3As_2$][123].

Arsine gas is one of the most toxic known substances, and is also highly flammable. The Materials Safety Data Sheet (MSDS) for arsine[127] lists the gas as

- Fatal if inhaled
- Suspected of causing cancer
- May cause damage to organs (liver) through prolonged or repeated exposure
- Very toxic to aquatic life with long lasting effects
- May form explosive mixtures with air

The MSDS states that release of arsine into the environment is to be avoided, and that "protective gloves, protective clothing, eye protection, respiratory protection, and/or face protection" is to be used. Any exposure to inhaled arsine is to be considered a potentially toxic dose, as symptoms may be delayed. As a precaution, the MSDS recommends that "Emergency eye wash fountains and safety showers should be available in the immediate vicinity of any potential exposure". Although arsine has a garlic-like odor, it has "poor warning properties" at low concentrations.

The original Merck field test kit for arsenic was later found to allow up to 50% of the arsine developed to escape the testing chamber into the environment, with the estimated arsine concentration in close proximity to the test becoming as high as 35 times the Threshold Limiting Value (TLV) [1] for arsine exposure[70]. This exposed those nearby to significant short and long-term danger, particularly for the workers administering the tests. As late as 2007, scientific publications were advocating for conducting the field tests "in the presence of an enthusiastic crowd"[73] in order to raise awareness of arsenic contamination, thus potentially exposing a far wider group to arsine gas.

Modern tests based on arsine quantification have improved their ability to contain the gas[99, 139], however the possibility for human error or malfunction remains and it is doubtful that field testers will ever have access to the emergency eye wash fountains and safety showers which would be a health a safety requirement in the developed world. Safe storage and disposal of the flammable arsine gas and other toxic kit reagents remains an issue, with the environmental contamination caused by millions of used field kits becoming a real concern[131].

[1] the level to which it is believed a worker can be exposed to a chemical substance day after day for a working lifetime without adverse effects, known as Workplace Exposure Limit (WEL) in the UK

Accuracy is another major issue, particularly with the earlier tests. Correct classification of wells depends on the operator correctly determining whether the exact shade of light yellow presented was above or below the threshold value. As is to be expected, accuracy using this visual identification varied from person to person, with accuracy particularly poor when in new hands[131]. Early studies of the classification accuracy – carried out by comparing field results with laboratory controls – recorded false negatives, where contaminated wells are painted green, as high as 68%, and false positives as high as 35%[131]. While the early kits could determine whether a well had very high arsenic or not, it was quickly realised that in the target range of $50\,\mu g l^{-1}$, the tests could not reasonably be described as quantitative[131].

This early poor performance spurred on a series of improvements, and soon newer test kits became available. In 2001 the U.S. company Hach released their arsenic field kit, while Merck released a 'doubling kit', both with a stated $10\,\mu g l^{-1}$ sensitivity[131]. The Hach kit saw considerable use, and in 2003 was shown to have correctly classified 88% of wells in a district of Bangladesh relative to the $50\,\mu g l^{-1}$ standard[158]. However, incorrect use appears to have persisted during this time, with reports of field workers feeling pressured into testing as many sample as possible during a day and cutting down the recommended twenty minute incubation time as only one sample can be processed at a time[158].

In 2003, UNICEF began using a new test kit, the Arsenator kit developed by Wagtech. This kit showed good accuracy, with a later version – the Wagtech Digital Arsenator – including a small battery operated colorimeter to decrease the variability of colour measurements from the mercury strips[48, 139, 149]. The Digital Arsenator is still used today[13] as a convenient field test for inorganic arsenic detection.

However, since 2006 government testing regimes have declined[43, 68]. Despite the accuracy of the field tests having increased and their cost having decreased – refills of the Arsenator kit cost around $0.7USD in 2008[139] – the up-front costs of the testing kit and training required for their use make them prohibitive for local communities. In 2009 a joint survey by UNICEF and the Bangladeshi Bureau of Statistics found that 50% of wells in Bangladesh remained untested[48, 125]. In reference to this report, George et al. [48] noted "an urgent need for expanding the availability of well testing at the village level", to combat the need to both ensure complete testing and repeated testing of wells, something which has proven impossible with the current approach[40]. While classical arsenic tests based on arsine production remain infeasible for roll-out at village level, another sensing technology might provide a solution.

Stocker et al. [153] showed that engineered *E. coli* strains were able to detect arsenic concentrations of $4\,\mu g l^{-1}$, significantly lower than both the Bangladeshi and WHO arsenic thresholds. Some strains were also capable of behaving quantitatively in the range of 8 and

$80\,\mu g l^{-1}$. A similar study in 2012[149] used lyophilized *E. coli* – thus avoiding the need to maintain a cold chain – which became bioluminescent in response to arsenic. While the accuracy of these first-generation tests remains modest, the technology shows significant promise. An accurate and cheap biological test would generate significantly less hazardous waste than current options, while being comparatively simple to use. And although each individual test may take longer than the $\sim 20$ minutes taken by current tests, the tests would be highly paralellisable leading to a far greater overall throughput.

A cheap, reliable and easy to use arsenic biosensor would be a vital technology for putting the ability to test water points into the hands of those who use them.

## 1.2 Biosensors

A biosensor is any sensing device which uses a biological element to detect or quantify the presence of a target element. Biosensors are typically employed when it would be technically difficult or not economical to sense the target directly using traditional methods such as those of analytical chemistry. The biological element acts to convert the target modality into one which can be more easily transduced, either electrically or by eye.

A common example is the home glucose sensor. This cheap inexpensive sensor allows diabetes patients to monitor their glucose levels at home without the need to see a professional. This lowers the pressure on healthcare, and allows patients to more effectively manage their condition, improving the effectiveness of their treatment as well as their quality of life.

Blood glucose sensors are based on a single enzyme – typically a glucose oxidase or glucose dehydrogenase – which specifically catalyses the oxidation of glucose. The progress of this oxidation can be easily measured by measuring changes is resistivity of the sample. This method has many advantages, but chief among them is that the specificity of the enzyme used means that no sample preparation is required, facilitating the development of cheap and simple-to-use in-home devices.

Traditional biosensors of this sort are relatively common – the home pregnancy test is another example – but they may not be easy to apply directly to arsenic detection. The bacterial response to arsenic is controlled by the behaviour of one protein, ArsR, discussed in detail in Section 3.1.1. Briefly, this protein binds to its promoter region, repressing its own expression as well as the downstream genes in the operon. When $As^{III}$ is present, ArsR releases the DNA – probably because of some conformational change brought about by interaction with arsenite – allowing transcription of the downstream genes. The exact combination of downstream genes present depends on the organism and operon, but they typically involve a transmembrane protein which facilitates $As^{III}$ egress from the cell, reducing its toxicity, and

a reductase which converts $As^{V}$ into $As^{III}$ which is then also removed from the cell. These functions are discussed in detail in Section 3.1.

There is no single biological molecule which is specifically sensitive to both tri- and pentavalent arsenic, meaning a traditional style biosensor would require either some initial sample preparation to reduce $As^{V}$ to $As^{III}$, or for both the regulatory protein (ArsR) and the reductase to be present and functional. Together with the fact that affinity for a specific DNA sequence is not a particularly convenient modality to detect within the requirements of the sensor – particularly that it be cheap and easy to use – this makes a single molecule approach to building an arsenic biosensor unattractive.

These traditional biosensors are not the only option, however, as whole-cell biosensors have also seen commercial success. Examples of such sensors are the MicroTox (Strategic Diagnostics, SDIX) and BioTox (Aboatox) systems, which employ naturally bioluminescent bacteria such as *Vibrio harveyi*, *Vibrio (Photobacterium) fischeri*, and *Photobacterium phosphoreum* as their sensing element[45]. These sensors are thus very broad spectrum, and will detect the presence of any analyte which interferes with the metabolism of the bacteria – detected via a reduction in bioluminescence output. While further tests are needed in order to identify the specific contaminant present, the absence of a large number of potentially harmful agents can be confirmed with a single test.

While bacteria are exquisitely sensitive to arsenic[24, 113, 141], and are able to activate a very specific set of genes to mitigate the harmful effects of arsenic, there is no visible change in phenotype that is specific to arsenic. However, synthetic biology allows us to engineer an new strain of bacteria that do exhibit a change, which is easy to measure. Previous efforts at creating an arsenic biosensor have used the luciferase system from *Vibrio harveyi*, a bioluminescent marine bacterium, to generate light as the output of the system[153]. These efforts were able to demonstrate some level of differentiation between arsenic concentrations, and could be used as a crude sensor, although the accuracy and dynamic range could be improved upon.

A sensor of this kind already has several advantages over a traditional field kit, however. Firstly, the sensor doesn't produce any harmful agents such as the arsine produced by standard kits. It also doesn't require the use of any chemical treatments – the water sample is simply added to the growth media – and is thus far simpler to use as well as requiring less equipment. Furthermore, since the sensing technology is based on living cells it should be cheap to fabricate as they are of course self-replicating and suitable growth media are thus the only purified chemicals that are required.

However, several challenges remain before the technologies discussed in Melamed [99], Siegfried et al. [149], Stocker et al. [153] can be used in the field. Firstly, while

these sensors represent a leap forward, they are not yet suitable for use by a completely untrained individual, at least in order to get a highly accurate reading. Secondly, ethical issues exist surrounding the use of biosensors based on Genetically Modified Organisms (GMOs) in developing countries. Countries such as Bangladesh often do not have sophisticated legislation on the use of GMOs, and previous studies have been able to conduct field trials using their synthetic strains[149]. However, due to the possibility for accidental release, significant regulatory oversight would be required in order to use such a device legally in countries such as the U.K. or the U.S. Widespread use of such a technology in a developing country without first receiving regulatory approval of this kind might not be illegal, but would be highly reminiscent of the attitudes that caused the arsenic disaster, when unsafe tube wells were installed without the level of oversight that would be expected in a developed country.

In the longer term, it is possible that a third technology might prove more successful in overcoming some of the regulatory issues encountered by whole-cell biosensors – the cell-free biosensor. Instead of using cells, such a sensor uses cell extract supplemented with sources of energy and amino acids[64, 152] to perform sensing in a manner similar to a whole-cell sensor. Since the extracts do not contain cells, they are not self-reproducing and are thus not GMOs, reducing regulatory overhead. Cell free extracts can also be freeze dried[124], extending their shelf life and removing the need for a cold chain. However, such systems would have other drawbacks, such as the relatively high cost of generating efficient cell extracts in sufficient quantity for the output to be cheaply detected, although this cost may fall as the technology becomes more established. Cell free extracts also lack cell walls, meaning they are less well isolated from their environment than whole-cells. It seems likely therefore that cell extracts will be more sensitive to deleterious effects caused by contaminants present in water samples than whole-cell alternatives. Without proper controls, safe levels of other contaminants could interfere with the detection of arsenic more strongly than whole-cell sensors.

For the time being, roll-out of a cell extract based arsenic biosensor is limited by the technology's higher cost and sensitivity to non-specific factors. There also remains a significant gap in our knowledge of the natural arsenic systems present in bacteria. While a good qualitative understanding has been gained over the last several decades – reviewed in Section 3.1 – there has been little to no quantitative data or rigorous mathematical modelling of the arsenic cycle in bacteria, and indeed the underlying mechanism behind the sensing of arsenic remains poorly understood. This work represents an attempt to redress this imbalance, by first developing a mathematical understanding of the response to arsenic in our target host, *Bacillus subtilis*, establishing a system for reliable ratiometric characterisation in that system, and then performing a thorough characterisation of the arsenic response in the organism.

It is hoped that this work will help speed the development of an arsenic biosensor – whether whole-cell or cell-free – as well as laying some of the groundwork necessary for bringing rational system design to *B. subtilis*.

## 1.3 *Bacillus subtilis*

*Bacillus subtilis* is a Gram-positive, rod shaped bacterium that has been commonly studied in laboaratories for the past six decades[179]. The bacteria were first described by Ehrenberg [36] in 1835, and are typically around 4 to 10 µm long and between $\frac{1}{4}$ and 1 µm in diameter. *Bacillus subtilis* is commonly found in many different environments across the globe, both terrestrial and aquatic[35]. By far the most commonly studied strain, as well as the most industrially relevant, is strain 168 which was isolated after two Yale University botanists exposed the wild-type *B. subtilis* Marburg to X-rays in the 1940s[16]. This strain has proven to be highly transformable – easily taking up foreign DNA and inserting it into its single, circular chromosome – a trait which makes the bacteria particularly suited to research and industry[179].

*B. subtilis* is also capable of forming highly resilient spores, and typically does so when faced with nutrient depravation or other environmental stresses[86, 179]. The spores are almost completely inactive and contain very few high energy compounds, and are protected from their environment by a number of mechanisms, reviewed by Setlow [144]. A thick spore coat and two membrane barriers protect the spore from many harmful chemical agents, and proteins on the surface reduce the toxicity of several toxins. Spores are protected from heat by their low water content, as well as by Small Acid Soluble Proteins (SASPs) which bind tightly to the spore's DNA. SASPs also protect against UV radiation damage, and spores also appear more resistant to $\gamma$-radiation than vegetative cells, although the mechanism for this is not clear. Upon germination, several DNA repair mechanisms are activated in an attempt to repair DNA damage accrued during potentially millennia of dormancy.

While *B. subtilis* is often referred to as a "soil bacterium", it is uncertain that this environment is ideal for vegetative growth or simply where spores of the bacterium collect. Early experiments suggest that vegetative growth is more commonly associated with decaying organic material, and the bacterium is also often found in close proximity to plant roots, and is thought to enhance plant growth. The bacterium is commonly ingested by animals and humans – soybeans fermented with *B. subtilis natto* are a popular food in Japan, believed to confer pro-biotic benefits[35].

The behaviour of *B. subtilis* in the gastrointestinal (GI) tract is not fully understood. *B. subtilis* was long believed to be an obligate aerobe, and as such passed through the GI

tract in spore form, however in 1998 it was revealed that the organism is able to respire anaerobically[108], and can in fact complete its entire life-cycle within the GI tract[35].

*Bacillus subtilis* are a promising chassis for a whole-cell arsenic biosensor. The are non-pathogenic and have a long history of safe use as part of the human diet and within the laboratory, and strains that are commonly studied – such as 168 – are out-competed by wild strains. The bacterium has a natural defence from arsenic in the form of the *ars* and *ase* operons, introduced in detail in Section 3.1.1, which are highly sensitive to low levels of environmental arsenic. Finally, the ability of the bacterium to form highly efficient spores removes the need for a cold chain or an inefficient freeze-drying step required in *Escherichia coli* based sensors, as simply produced spores can be shipped and held indefinitely in almost any conditions until needed.

## 1.4 Summary and Aims

There is a clear need for technological innovation to help combat arsenic contamination of groundwater in the developing world. Widespread and continuous monitoring of tube-wells has proven to be the single most effective strategy for mitigating the effects of arsenic ingestion, particularly when combined with interventions such as the installation of deeper wells in the worst affected areas. However, blanket testing campaigns funded by international aid organisations in cooperation with local governments have lost their way. A safe, inexpensive, and accurate test for arsenic, distributed to and conducted by those who need it could help millions avoid the debilitating and deadly effects of arsenicosis, as well as help governments and NGOs to better target their resources to those in need.

Despite initial enthusiasm, many technical, regulatory, and social challenges exist before a workable arsenic biosensor can be rolled out. As synthetic biology comes of age, it is hoped that the regulatory process will become more receptive, with the development of a smooth and well defined pathway towards approval together with increased public engagement and understanding. Increased education about the dangers of arsenic contaminated water, as well as proper control of sensitive well contamination data, may also help alleviate some of the social challenges.

Central to this work is the question of how we can build a genetic circuit capable of accurately and reliably detecting very low levels of arsenic. By definition, we are interested in detecting arsenic at levels far lower than those that cause harm – at least to humans and wildstock – making reliable detection by microbes more difficult.

It seems, therefore, that a central challenge will be the faithful amplification of a weak, noisy and stochastic biological signal, however we are also hampered by our poor quantitative

understanding of the mechanisms by which several bacterial species are able to protect themselves from arsenic toxicity. In light of this, the primary goal of this work is to develop a better and more quantitative understanding of the response of our target organism, *Bacillus subtilis*, to arsenic, and to demonstrate how synthetic biology might use this native system to design a more robust sensor.

# Chapter 2

# Materials and Methods

## 2.1 Introduction

Standard protocols were used where possible throughout this project, which largely involved straightforward molecular cloning techniques. This chapter gives an overview of the techniques used, referencing more detailed protocols where appropriate and noting any modifications that were made to the standard protocols. Section 2.5 details the various strains, and media which were generated and used during the project and which are referred to during the text.

## 2.2 DNA Manipulation and Assembly

All PCR reactions were carried out using Q5 High-Fidelity DNA Polymerase[110] from NEB, 1% agarose was used for gel electrophoresis unless otherwise stated and gel extractions and gel slice purifications were performed using the QIAquick Gel Extraction kit from Qiagen[130]. DNA sequence verification was by Sanger sequencing, performed by Source BioScience[117], which was used for both linear PCR products and assembled plasmids.

Plasmids were mostly assembled using Golden Gate assembly, following the standard procedure described in Engler and Marillonnet [39], with the modification that 45 °C was used for restriction when using BsmBI as opposed to the 37 °C recommended for the BsaI restriction enzyme, as discussed in Section 4.2.2.

In some circumstances Gibson assembly was used instead, following standard procedures described in Gibson et al. [49]. Gibson assembly was mainly employed in situations where Golden Gate assemblies were not feasible, such as in introducing silent mutations to remove

the BsmBI restriction sites which were present in the backbone of the original template plasmids used during the development of the reporter system described in Chapter 4.

## 2.3 Bacterial Cloning

*Escherichia coli* was used throughout to amplify assembled plasmids. Plasmids were transformed into chemically competent *E. coli* and grown under selection on solid media. Chemically competent *E. coli* were generated in house using the standard 'E. cloni' strain and standard methods from OpenWetWare[118]. Briefly, cells are grown to an Optical Density at 600 nm (OD-600) of approximately 0.3, before being centrifuged at 6000 RCF for 10 min and resuspended in ice cold CCMB80 buffer (see Section 2.5.5) and incubated on ice for 20 minutes. The cells are again centrifuged at 4 °C at the same speed and time then resuspended in ice cold CCMB80 such that a mixture of 200 µl of growth media and 50 µl of cells has an OD-600 of between 1.0 and 1.5. This is then divided into 50 µl aliquots into pre-chilled micro-centrifuge tubes and stored at −80 °C.

Transformation was carried out by thawing on ice before adding circular DNA – usually the result of an assembly reaction – pipetting gently to mix thoroughly and then holding on ice for 30 minutes. The cells were then heat shocked at 42 °C for 45 s and then returned to ice for 2 minutes. 250 µl of LB growth media was then added – many protocols recommend glucose supplemented media such as SOC, but it did not appear to be any more effective than plain LB – and grown with shaking at 90 RPM at 37 °C for 1 hour to allow expression of the newly transformed resistance marker. 100 µl was was then plated on solid media with antibiotic marker and grown overnight. In the case of low-efficiency assembly reactions, the entire mixture were first centrifuged at low speed before being resuspended in 100 µl of growth media and plated.

Positive colonies were used to inoculate overnight cultures with antibiotic and harvested using the QIAprep Spin Miniprep[129] kit from Qiagen. Harvested plasmids were then sequence verified an linearised by restriction digestion with ScaI-HF before being transformed into competent *Bacillus subtilis* 168 *trpC2* auxotroph by homologous recombination.

Competent *B. subtilis* were generated also using a standard method[116], where cells are first grown in 10 ml of minimal growth medium for 18 h at 37 °C with shaking at 90 RPM. 1.4 ml are then sub-cultured into a further 10 ml of pre-warmed minimal growth medium and grown with shaking at 90 RPM. After 3 h, 11 ml of starvation medium is added and growth continued for 2 h and 45 min. Sterile glycerol is then added to 10% v/v and then 300 µl aliquots are made and stored at −80 °C.

Thawed cells were transformed by the addition of linear DNA, and incubated with shaking at 90 RPM at 37 °C for one hour. 700 µl of LB medium were then added, and the cells grown for 1 h to 2 h at 37 °C depending on the antibiotic marker present, with a longer incubation given for bactericidal antibiotic markers. The cells were then plated on solid media with antibiotic and grown overnight.

### 2.3.1 Selectable Marker Excision in *B. subtilis*

As several points during this project, the Xer recombinase system was used to remove an antibiotic marker after successful transformants had been selected for. This allowed for the same antibiotic to be re-used in further transformations, but also meant the gene deletions which were made using this technique were less likely to have a wider, non-specific effect as the antibiotic is no longer produced once it has been excised.

The *cat* gene, which confers resistance to chloramphenicol, a bacteriostatic antibiotic which prevents protein elongation[143], was used as the selectable marker to be removed by Xer recombinase throughout. To this end a variant of the *cat* resistance cassette dubbed *dcat* which is flanked by the 28 bp diff regions defined in Bloor and Cranenburgh [11] was built using primer extension. This linear fragment included ends compatible with Golden Gate using the BsaI enzyme, allowing swift insertion into a plasmid.

Such a plasmid could be transformed as normal, however it was found that a greater efficiency was achieved by reducing the outgrowth period – where cells are grown in antibiotic-free media – to between 30 and 60 min. This is likely due to the possibility of the antibiotic marker being cured out before selection takes place, which increases with a longer outgrowth phase. Furthermore, it was found that reducing the concentration of chloramphenicol to $5\,\mu g\,l^{-1}$ increased the number of positive colonies. The growth of these colonies was typically far lower than one would normally expect and on occasion is was necessary to incubate transformed cells on solid media for 36 h before visible colonies appeared.

Once colonies we present, they were grown in plain LB media for at least 6 h, but typically overnight, before again being plated but this time onto plain LB agar. These colonies were then verified by colony PCR for deletion of the target gene and excision of the *cat* resistance.

Although Bloor and Cranenburgh [11] recommend demonstrating the removal of the *cat* resistance by again plating onto chloramphenicol to demonstrate that the resistance has been lost, colony PCR was preferred as it generates faster results. A wild-type *B. subtilis* strain as well as the plasmid with which the strain was transformed were used as controls, the first to demonstrate transformation and control against potential miss-priming elsewhere in the genome, and the second to demonstrate the excision of the *cat* gene.

## 2.4 Flow Cytometry

Flow cytometry was carried out using a Cytek FACScan with DXP 8 colour upgrade. mVenus fluorescence – used for the query channel – was collected from the BluFL1 channel, corresponding to an excitation wavelength of 488 nm and an emission wavelength of 530 nm. mScarlet-I, which was used as the reference reporter was collected from the YelFL2 channel, corresponding to an excitation wavelength of 561 nm and collected in the far-red region up to 740 nm.

Logarithmic data collection was enabled for all channels, as is recommended for bacterial samples.

Ratiometric values were calculated on this data after transformation back into the linear domain. Measured autofluorescence was first subtracted separately from each channel before taking the ratio of query to reference signal for each sample. For each experimental run, the analogue gain of each channel was set such that the ratio of the reference strain – which contains both query and reference reporter under strong constitutive expression – was equal to one, and that the signal for each channel was approximately 85% of the saturating value of the device in order to give good resolution.

In order to correct for differences between different sessions, the ratiometric output was normalised against the output from the reference strain, with the output from the reference strain defined to be unity such that small differences in gain were corrected for.

## 2.5 List of Materials

### 2.5.1 Plasmids

The key plasmids generated during the project were the query and reference template plasmids, which each allow for easy insertion of a promoter and RBS combination. The original source of the backbones sequences for these plasmids was an in-house plasmid available at the start of the project, however the backbones were modified to remove two BsmBI cut-sites using Golden Gate assembly. The final plasmids were thoroughly verified using Sanger sequencing.

Variants of the template plasmids for which a promoter and RBS sequence has been inserted are labelled in the format $<TemplateID>.<PromoterNumber>.<RBSNumber>$, where the promoter number and RBS number are as defined in tables 4.4 and 4.5 respectively.

In the case of the *ars* operon based plasmids, the operon and around 1kb either side was cloned by colony PCR from *Bacillus subtilis* sp. 168 and inserted into a high copy plasmid backbone and was they verified with Sanger Sequencing.

| Name | Description | Sequence link |
|---|---|---|
| pHK025v2 [a] | mCherry Reference Template in amyE locus | https://benchling.com/s/YVhf20eG |
| pHK026 [a] | comGA-mVenus Query Template in amyE' locus | https://benchling.com/s/SDtt5Wgg |
| pHK027 | comGA-mVenus Query template in amyE' locus, reversed | https://benchling.com/s/seq-AFaO9GR4e02PE7QB3yB9 |
| pHK028 | mCherry Reference Template in amyE' locus | https://benchling.com/s/seq-8YioKG17G1KmcRpoNn6z |
| pHK029 | comGA-mVenus Query Template in amyE locus | https://benchling.com/s/seq-vKZ4cep1bdbkwQji3HvS |
| pHK030 | mVenus Query template in amyE' locus | https://benchling.com/s/seq-WTBBECN6lOXnRxnFRGKl |
| pHK031 | mKate2 Reference Template in amyE locus | https://benchling.com/s/seq-Ca3sCn7xlF1VT9cLoNwR |
| pHK032 | comGA-mKate2 Reference Template in amyE locus | https://benchling.com/s/seq-xR2itEn5OXJ4avHtNGSt |
| pHK033 | mScarlet Reference Template in amyE locus | https://benchling.com/s/seq-UuVmm9L9K7SqSAdV660Y |
| pHK034 | comGA-mScarlet Reference Template in amyE locus | https://benchling.com/s/seq-rfXUv6usEglKdIVWk4NB |
| *ars* Operon | pHK001v2 | https://benchling.com/s/NJc2Jzv2 |
| *ars* Operon Δ*arsR* | pHK001_arsR | https://benchling.com/s/Hogwaufh |

Table 2.1 Key plasmids developed during the project. Name refers to the name written on the tubes containing plasmid and glycerol stocks in the cold storage boxed labelled 'plasmids' and 'glycerol stocks', respectively. (a) Sequences also available in Appendix D

### 2.5.2 Oligos

Single stranded oligos were synthesized externally by either IDT or NEB. Short double stranded sequences such as the promoters and RBSs described in tables 4.4 and 4.5 were synthesized as single stranded oligos and annealed by heating to 98 °C and slow cooling to room temperature.

All oligos synthesized for the project can be found in the cold storage boxes at −20 °C labelled 'Haydn Oligos'. The mapping between labels and sequence information can be found in appendix E.

### 2.5.3 Fluorescent Reporters

Some fluorescent reporters were already available in-house (mVenus, mCherry, and mTurquoise) while codon optimised versions of all others were synthesized using IDT. Sequences for each fluorescent reporter used in the project can be found in the plasmids described in table 2.1.

### 2.5.4 Strains

The strains developed during the project are enumerated in Table 2.2. The strain description use the convention $P_{\langle promoter \rangle}$-⟨RBS1⟩-⟨gene1⟩-... to describe transcription units.

### 2.5.5 Media

Pre-made powdered Luria Broth (LB) was used throughout.

#### Minimal Salts Solution

Minimal salts solution was prepared with

- 2 g Ammonium sulphate
- 14.8 g Potassium Hydrogen Phosphate
- 5.4 g Potassium Dihydrogen Phosphate
- 1.9 g Sodium Citrate
- 0.2 g Magnesium Sulphate Heptahydrate

which was dissolved in 150 ml deionized water and adjusted to pH 7.0 with HCl or NaOH. The final volume was adjusted to 200 ml and autoclaved.

| | Designation | Description |
|---|---|---|
| Chapter 4 | R | $P_{pen}$-*gsiB-mCherry* inserted into AmyE' locus |
| | R* | $P_{pen}$-*gsiB-mScarlet-I* inserted into AmyE' locus |
| | cG-mS cG-mV | $P_{pen}$-*gsiB-comGA-mScarlet-I* in AmyE' and $P_{pen}$-*gsiB-comGA-mVenus* in AmyE |
| | mS cG-mV | $P_{pen}$-*gsiB-mScarlet-I* in AmyE' and $P_{pen}$-*gsiB-comGA-mVenus* in AmyE |
| | cG-mS mV | $P_{pen}$-*gsiB-comGA-mScarlet-I* in AmyE' and $P_{pen}$-*gsiB-mVenus* in AmyE |
| | mS mV | $P_{pen}$-*gsiB-mScarlet-I* in AmyE' and $P_{pen}$-*gsiB-mVenus* in AmyE |
| | R* Q$\langle x\rangle$.$\langle y\rangle$ | R* + $P_{\langle x\rangle}$-$\langle y\rangle$-*comGA-mVenus* in AmyE ($P_{\langle x\rangle}$ and $\langle y\rangle$ defined by Tables 4.4 and 4.5 respectively) |
| | R* A$\langle x\rangle$ | Equivalent to R* Q$\langle x\rangle$.1 |
| | R* B$\langle x\rangle$ | Equivalent to R* Q$\langle x\rangle$.1 but with mVenus inserted into LacA locus |
| | R* C$\langle x\rangle$ | Equivalent to R* Q$\langle x\rangle$.1 but with mVenus inserted into CotVWX locus |
| Chapter 5 | Δase | deletion of ase operon using *dcat* system |
| | Δase R* | Δase + R* |
| | Δase R* Q8.1 | Δase R* + $P_{ars}$-*gsiB-comGA-mVenus* in AmyE locus |
| | Δase R* Q9.1 | Δase R* + $P_{ars}$-*arsR-arsR-gsiB-comGA-mVenus* in AmyE locus |
| | Δase R* Q8.1 ΔarsR | Δase R* Q8.1 + deletion of native *arsR* gene |

Table 2.2 List of relevant strains developed during the project

**Minimal Growth Media**

Minimal growth media was made with

- 10 ml Minimal Salts Solution
- 0.5 ml Glucose (50% (w/v))
- 0.5 ml Casamino Acids (2% (w/v))
- 0.1 ml Tryptophan (10 mg ml$^{-1}$)
- 0.05 ml Iron Ammonium Citrate (2.2 mg ml$^{-1}$)
- 39 ml Deionised Water

per 50 ml, and filter sterilised.

**Starvation Media**

Starvation media was made with

- 10 ml Minimal Salts Solution
- 0.5 ml Glucose (50% (w/v))
- 39.5 ml Deionised Water

per 50 ml and filter sterilised.

### 2.5.6 CCMB80

CCMB80 contains

- 10 mM KOAc pH 7.0
- 80 mM $CaCl_2 \cdot 2H_2O$
- 20 mM $MnCl_2 \cdot 4H_2O$
- 10 mM $MgCl_2 \cdot 6H_2O$
- 10% Glycerol

and was pH adjusted to 6.4 using HCl, filter sterilised and stored at 4 °C.

# Chapter 3

# Modelling of the native arsenic operon of *B. subtilis*

## 3.1 The Bacterial Response to Arsenic

All living organisms have mechanisms for mitigating the toxic effects of arsenic. The fundamental mechanism by which bacteria protect themselves appears highly similar, although the implementation details vary across species and are reviewed in Rosen [135]. Several of the proteins involved have homologues in other organisms[106, 175], allowing gene function to be inferred. This chapter focusses on the response in *B. subtilis*, although knowledge of the behaviour in other organisms is often relied upon due to the lack of direct studies in the bacterium.

### 3.1.1 In *Bacillus subtilis*

*Bacillus subtilis* contains two arsenic related operons, the *ase* operon[60] and the *ars* operon[141]. While functionally similar, the two operons share little sequence homology and appear to act independently. The *ase* operon is the simpler of the two operons, containing only two genes, and is thus likely to be the older based on theories of the development of the arsenic operon[135].

The *ars* operon is almost exactly at the opposite side of the chromosome from the *ase* operon, some 2 Mb away. This operon is within the sigma-K interrupting ('skin') element, a $\sim$48 kb region which is excised during late stage sporulation. This excision creates a new composite gene, *sigK*, as a fusion of the *spoIIIC* and *spoIVCB* genes which flank the skin element, a sigma-factor associated with late stage sporulation genes.

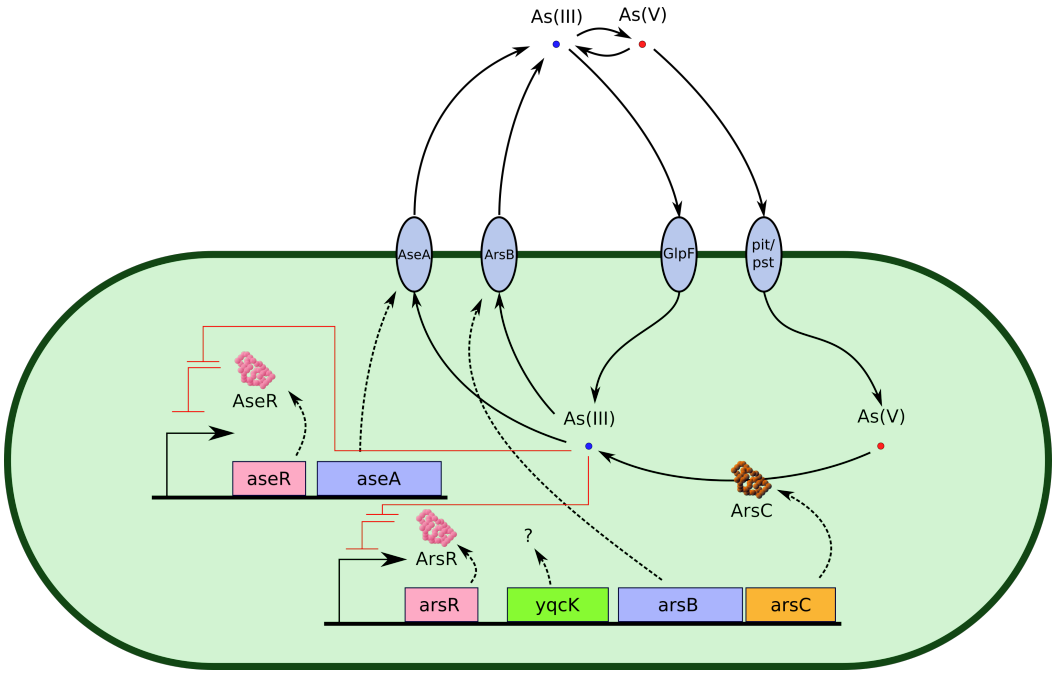



Fig. 3.1 Schematic of arsenic detoxification in wild-type *Bacillus subtilis* showing both operons and the action of all known gene products.

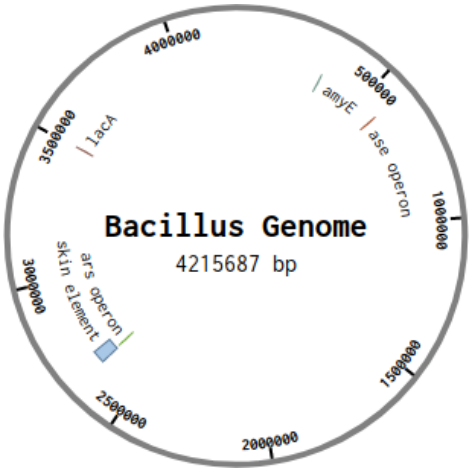



Fig. 3.2 Genome map of *B. subtilis*, showing locations of the *ars* and *ase* operons, the skin element, and two common insertion loci, *amyE* and *lacA*

As well as the four genes of the ars operon - *arsR*, *yqcK*, *arsB*, and *arsC* - several other open reading frames (ORFs) exist in the skin element. Several of these are believed to be of phage origin, suggesting that the skin element may be the relic of an ancient phage integration.

Whatever its true origin, the *ars* operon has been shown to be more sensitive to induction by arsenic than the *ase* operon, as well as conferring a higher tolerance to arsenic[102]. Because of this it has been researched in considerably more detail than the *ase* operon, and although the mode of action of the two operons is similar, this makes the *ars* operon considerably more attractive as the basis of a biosensor.

The *ase* operon should not be ignored altogether though; were the *ars* operon considerably more fit for arsenic detoxification than it in all conditions, then one would expect the *ase* operon to have been lost entirely. The fact that it has been retained suggests that it may play an important role in some as yet unknown function, and it certainly cannot be ignored when building a biosensor.

The remainder of this section introduces what is known about arsenic detoxification in *B. subtilis*, both from direct experimental observation and through inferred homology with other organisms, in the context of the *ars* operon. The simpler known behaviour of the *ase* operon is then discussed, and finally arsenic detoxification in *B. subtilis* is compared to the systems which exist in *E. coli*, the only organism to have been previously targeted for arsenic biosensor development.

### Uptake of Arsenate and Arsenite

Arsenate enters the cell via phosphate transporters[8, 58, 136, 167, 168] as the two share a pentavalent state. In *E. coli*, two phosphate uptake systems exists, the phosphate-specific transport (pst) system and the phosphate inorganic transport (pit) system. Arsenate is able to enter via both systems, although the pit system appears to be the dominant method of entry[167]. A homologue of the pst system has been confirmed in *B. subtilis*, and the presence of a pit homologue is strongly inferred[128], making it highly likely that arsenate entry into *B. subtilis* is via a similar mechanism.

The more toxic arsenite on the other hand was relatively recently discovered to enter via glycerol channels[91]. It was first observed that antimonite [$Sb^{III}$] enters through the GlpF membrane protein in *E. coli*. Since antimonite and arsenite share oxidation states it was suspected that arsenite also enters via the same channel, and this was later confirmed in *Saccharomyces cerevisiae*[172]. Similarly to the case with phosphate, a homologous glycerol uptake facilitator, *glpF*, has been identified in *B. subtilis*[6], although arsenite uptake by this channel has not been explicitly tested in *B. subtilis*, and no quantitative data is available.

Despite clear structural differences between the cell structure of the gram-positive *B. subtilis* and gram-negative *E. coli*, or the yeast *S. cerevisiae*, it appears that uptake of arsenite and arsenate is similar in each. Similar behaviour has also been demonstrated in the gram-positive *Corynebacterium glutamicum*, which was demonstrated to be more prone to arsenate uptake when grown in glycerol-rich media[97].

### Regulation of the *ars* Operon

As introduced above, the *ars* operon contains four genes, the first of which, *arsR*, represses transcription of the operon in the absence of arsenite by binding immediately upstream of the promoter region. In the presence of arsenite or antimonite, the strength of the ArsR – DNA interaction is reduced, allowing unbinding of ArsR and transcription of the downstream genes[174]. In this way, expression of the arsenic detoxification machinery is limited to the case where arsenite is present in the cytoplasm. All known arsenic operons are negatively regulated in this way – although in rare cases the regulatory protein is expressed independently such as in *Thiobacillus ferrooxidans*[21] – suggesting that expression of the remaining proteins incur a significant metabolic penalty.

ArsR is unable to bind to DNA as a monomer, and instead forms a homodimer before it is able to bind DNA[173]. Detailed experiments were carried out by Wu and Rosen [169] to identify the ArsR binding site in *E. coli*, summarised in Figure 3.3. First, DNAseI foot-printing experiments found a region protected from degradation which spans the 24 bases between the $-64$ and $-40$ loci relative to the transcription start site. An imperfect dyad symmetry containing 5 bp exists within this region, which are commonly associated with protein-DNA interactions[30, 169]. Although this does not overlap the predicted sigma factor binding sites it is expected to prevent transcription as RNA polymerase occupies promoter regions up until approximately the $50^{\text{th}}$ base upstream of the transcription start site[55]. Following this success, Wu and Rosen [169] continued using hydroxyl radical footprinting in an attempt to gain a higher resolution snapshot of ArsR – DNA interactions. This revealed only two small 4 bp protected regions separated by 10 bp, suggesting that the $ArsR_2$ complex binds from one side of the DNA helix only.

A similar structure exists within the *ars* promoter region in *B. subtilis* where two perfectly symmetrical 8-mer dyads of sequence *AAATAAAT* are separated by 2 bp, shown in comparison to the *E. coli* promoter in Figure 3.3. This structure is a mere two base pairs upstream of the predicted location of the $-35$ sigma factor binding site, thus any protein bound to the DNA at this location seems almost certain to physically prevent binding of the sigma factor and loading of the DNA polymerase. While there is no direct evidence that this

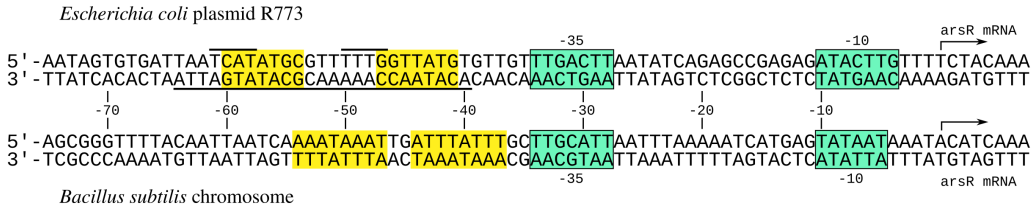



Fig. 3.3 Comparison of the *ars* operon promoters from *E. coli* plasmid R773 (upper) and the *B. subtilis* 168 chromosome (lower). cyan highlighting shows predicted sigma-factor binding sites, yellow highlighting shows locations of symmetric dyad repeats – imperfect in *E.coli*. Underlined and overlined show locations of *DNAseI* protected and hydroxyl radical protected sequences elucidated in Wu and Rosen [169] respectively.

sequence is the ArsR binding site in *B. subtilis*, the similarities with the *E. coli* sequence make this overwhelmingly likely.

The exact mechanism by which $As^{III}$ and $Sb^{III}$ trigger the release of DNA by ArsR is not known, though it is likely that some conformational change is caused by the interactions which significantly reduces binding. In *E. coli*, three cysteine residues of ArsR are known ligands of arsenite, $Cys-32$, $Cys-34$ and $Cys-37$, with the presence of any two being enough to bind arsenite[148]. However, while binding of arsenite persists when either $Cys-32$ or $Cys-34$ are mutated, arsenite binding no longer causes unbinding of DNA, suggesting that the conformational change is triggered by binding to these two residues[148].

This regulatory mechanism is not uncommon in systems for sensing metal ions, and several other systems in the wider family of genes – called the *SmtB/ArsR* family – have been better studied. As well as As(III) and $Sb^{III}$, other members of this family are able to detect $Zn^{II}$, $Co^{II}$, $Cd^{II}$, $Pb^{II}$, $Bi^{III}$, and $Ni^{II}$[19], summarised in Table 3.1. All members of this family are homodimers with large association constants, with the presence of metal ions having no significant effect on the strength of the dimer-dimer interaction.

In the absence of the target metal ions, all dimers in the *SmtB*/*ArsR* family bind strongly to palindromic or semi-palindromic DNA operators, although the exact mode of binding varies considerably. For example, in the case of SmtB, a zinc regulatory protein from *Synechococcus* PCC 7942, four $SmtB_2$ dimers bind to a single 40bp region containing one 12bp inverted repeat. Association constants for these interactions vary considerable from around $2.9 \times 10^9$ $M^{-1}$ for the first two homodimers to $3.4 \times 10^8$ $M^{-1}$ and $8.6 \times 10^6$ $M^{-1}$ for the third and fourth respectively. The exact nature of the cooperativity between these four proteins is unknown[160, 161].

Several of the members of this protein family contain two metal binding domains per monomer[19], although in some cases metal binding is exclusive between the two sites, and

| Repressor | Organism of Study | Metals | $K_{dimer}{}^{a}$ /$M^{-1}$ | $K_{metal}{}^{b}$ /$M^{-1}$ | $K_{DNA}{}^{c}$ /$M^{-1}$ | metal-induced fold reduction in $K_{DNA}{}^{d}$ | Reference |
|---|---|---|---|---|---|---|---|
| SmtB | *Synechococcus* PCC7942 | $Zn^{II}$, $Co^{II}$ | $3 \times 10^5$ | $\geq 1.7 \times 10^9$ | $29 \times 10^8$ | 1000 | [77, 160, 161] |
| ZiaR | *Synechocystis* PCC6803 | $Zn^{II}$ | – | – | – | – | [155] |
| CzrA | *Staphylococcus aureus* | $Co^{II}$, $Zn^{II}$ | $1.7 \times 10^5$ | $3 \times 10^9$ | – | – | [37] |
| NmtR | *Mycobacterium tuberculosis* | $Ni^{II}$, $Co^{II}$ | $1.9 \times 10^5$ | $\geq 1 \times 10^8$ | $6.8 \times 10^6$ | 1000 | [23, 126] |
| CadC | *Staphylococcus aureus* pI258 | $Cd^{II}$, $Pb^{II}$, $Bi^{III}$, $Zn^{II}$, $Co^{II}$ | $3 \times 10^6$ | $4.3 \times 10^{12}$ | $1.1 \times 10^9$ | 300 | [17, 18, 20] |
| ArsR | *Escherichia coli* | $As^{III}$, $Sb^{III}$ | N/A | N/A | $1 \times 10^6$ | – | [173] |

Table 3.1 Summary of members of the *SmtB/ArsR* family of metal ion sensors and their thermodynamic properties. Shown are association constants of (a) the dimers themselves (b) the dimer-metal complex, and (c) the dimer-DNA complex. (d) shows the fold reduction in dimer-DNA association constant observed when metal is bound to the dimer.

different metal ions can favour different binding sites[160]. However all known instances of ArsR have only one metal binding site per monomer[19], though it is possible for two arsenite ions to bind to each homodimer.

### Reduction of Arsenate

Unlike arsenite, arsenate does not interact with the regulatory protein ArsR and thus is unable to directly de-repress the *ars* operon[141]. Instead, arsenate is reduced to the more toxic arsenite by ArsC, which is present at low levels in the cell due to leaky expression of the *ars* operon. Once arsenite is produced, it then induces expression of the operon, further catalysing the reduction of arsenate to arsenite.

It is rather curious that the less toxic arsenate is converted into its more toxic form before extrusion from the cell, rather than being extruded directly. One hypothesis which would explain this was put forward by Rosen [135] is that this reduction to the more toxic form is an accident of evolution. Since arsenite is more toxic, it stands to reason that primordial cells would have evolved mechanisms to mitigate its effects before a strategy had evolved to deal with the less toxic forms. This hypothesis is upheld by the observation that several arsenic-related operons do not have an arsenate reductase gene and are thus unable to process arsenate – indeed, the *B. subtilis ase* operon is an example of this class. It is believed that at some point a single arsenic reductase developed[72], which then spread across several organisms, and was incorporated into the existing arsenite response. In *Bacillus subtilis*, ArsC is most structurally similar to low molecular weight tyrosine phosphatases[7], suggesting a possible common origin.

A fair amount is known about the mechanism by which arsenate is reduced in *E. coli*, with published structures having been made available in 2001 for *E. coli* ArsC and ArsC in complex with both arsenate and arsenite[96]. In *E. coli*, the reduction of arsenate by ArsC was originally discovered to be linked to glutathione (GSH) by Oden et al. [114]. Further kinematic analysis[50] determined Michaelis-Menten parameters for the reaction with $K_m$ for arsenate of $8mM$ and $V_{max}$ in the range of 0.8 to $1.5\frac{\mu mol}{min \cdot mg}$. Several inhibitors of arsenate reduction were also investigated, with arsenite, phosphate, and sulphate determined to have $k_i$ values of $\{0.1, 30, 10\}mM$ respectively. It is unsurprising that both phosphate and sulphate are inhibitors of the reduction as both share arsenate's pentavalent oxidation state, suggesting that they bind competitively to the active site of ArsC. No reduction of either phosphate or sulphate has been observed, however. The fact that arsenite is a stronger inhibitor of the reduction than either phosphate or sulphate is interesting, as it suggests an element of negative regulation in the reduction of arsenate, reducing the rate at which arsenite is produced when levels of the toxic ion are elevated.

A full reaction mechanism for the reduction of arsenate in *E. coli* was suggested by Shi et al. [147], shown below:

$$\mathrm{ArsCS^- + As^V \rightleftharpoons ArsCS^- \cdot As^V} \tag{3.1}$$

$$\mathrm{ArsCS^- \cdot As^V + GS^- \rightleftharpoons ArsCS{-}SG + As^{III}} \tag{3.2}$$

$$\mathrm{ArsCS{-}SG + Grx{-}S^- \rightleftharpoons ArsCS^- + GS{-}SGrx} \tag{3.3}$$

$$\mathrm{GS{-}SGrx + GS^- \rightleftharpoons Grx{-}S^- + GSSG} \tag{3.4}$$

$$\mathrm{GSSG + NADPH + H^+ \rightleftharpoons 2\,GSH + NADP^+} \tag{3.5}$$

where ArsCS represents the active $Cys^{12}$ residue of ArsC, which first binds with $As^V$ in Reaction 3.1. The actual reduction of $As^V$ occurs in Reaction 3.2, with one electron donated from the thiolate bond of $Cys^{12}$ and one from GSH, represented as $GS^-$. Functional ArsC is then restored in Reaction 3.3 as the attached GSH is removed by a glutaredoxin (Grx) enzyme. A further GSH is required in Reaction 3.4, freeing the Grx and leaving oxidised glutathione disulphide (GSSG), which is finally reduced by NADPH in Equation 3.5.

*E. coli* contains three glutaredoxin enzymes, $Grx_1$, $Grx_2$, and $Grx_3$, which share little sequence similarity[163] but show similar activity for reduction of general mixed disulphide bonds such as those found in ArsCS – SG[65]. However, $Grx_2$ was found to be between 1 and two orders of magnitude more active in reducing ArsCS – SG (shown in Reaction 3.3) than $Grx_1$ or $Grx_2$[147], suggesting that some form of protein-protein interaction may occur between $Grx_2$ and ArsR increasing the efficiency of the reaction.

No glutaredoxin system has been characterised in *B. subtilis* as of yet, although a thioredoxin system which is also known to be involved in the reduction of disulphide bonds[93] is present. Unlike the thioredoxin gene in *E. coli*, the thioredoxin gene TrxA in *B. subtilis* is essential[142], suggesting that it plays a more central role in cellular processes than the homologous system in *E. coli*. TrxA is thus a likely candidate for reduction of the ArsCS – SG bond in *B. subtilis*.

### Extrusion of Arsenite

The underlying mechanism of arsenite resistance in prokaryotes is through the extrusion of arsenite, keeping arsenite levels within the cell to sub-lethal levels[135]. In the *ars* operon of *B. subtilis*, this is mediated by ArsB, a membrane protein which allows for $As^{III}$ transport out of the cell.

Again, most major studies of ArsB have focussed on the version found in *E. coli* plasmid R773, which contains 12 membrane spanning regions[171]. In *E. coli* – and commonly

in other gram-negative species – ArsB is not the only component of the arsenite extrusion system, another *ars* operon gene product, ArsA is able to couple with membrane bound ArsB to form an anion transporting ATPase and thus drive arsenite export via ATP[34]. Remarkably however, *E. coli* R773 ArsB is also capable of ATP-independent arsenite transport when ArsA is not bound to it, with the transport instead driven by the electrochemical gradient formed by NADH respiration[84].

Significant work in characterising the ATP independent, ArsB mediated efflux of arsenite and other metals in *E. coli* was carried out by Kuroda et al. [84] and later by Meng et al. [100]. Both of these studies assayed arsenite extrusion by quantifying arsenite uptake by everted membrane vesicles formed from cells expressing ArsB. Metalloid uptake was quantified by first filtering whole vesicles and washing followed by either liquid scintillation counting using radioactive isotopes of each metalloid (as favoured in the earlier paper) or by inductively coupled plasma mass spectrometry (ICP-MS). The total mass of recovered protein was also assayed using a commercial BCA protein assay kit (Pierce), with bovine serum albumin as a standard. Uptake of metalloid by everted vesicles (i.e. efflux from cells) was then reported as moles of metalloid per nano-gramme of total protein.

Kuroda et al. [84] demonstrated that arsenite transport into everted vesicles by ArsB in the absence of ArsA is dependent on NADH oxidation, and that ArsB does not catalyse non-specific anion transport by testing a range of sodium and potassium salts. Under these conditions, the $K_m$ for transport with respect to arsenite concentration – i.e. the concentration at which half-maximal transport rate is achieved – was estimated to be approximately $0.14mM$, very similar to the $0.1mM$ observed for ATP-dependent transport[33], suggesting that the rate-limiting mechanism for arsenite extrusion by ArsB is independent of the source of energy. Since only total protein present in membrane vesicles was used to normalise metal ion content rather than vesicle volume or ArsB content, no estimate can be made of $k_{cat}$, the maximal rate at which unit concentration of ArsB can transport $As^{III}$.

Kuroda et al. [84] also made the observation that $As^{III}$ uptake by vesicles was catalysed by $Sb^{III}$ ions, increasing the rate at which arsenite was taken up by up to 5 fold. This effect was found to be an effect of ArsB mediated transport rather than some non-specific effect of antimonite, and was observed to be strongest when arsenite and antimonite are present in equimolar amounts. It is not surprising that antimonite is also transported by ArsB, since it shares oxidation state with arsenite and is also known to de-repress the *ars* operon. This led Kuroda et al. [84] to hypothesise that antimonite might in fact be a more efficient substrate for ArsB mediated transport, causing arsenic to be extruded faster as a mixed salt of arsenic and antimonite than when present on its own. This was confirmed by Meng et al. [100] who found that $^{125}Sb^{III}$ uptake by everted vesicles was approximately 35 times higher than

uptake of $^{73}As^{III}$, and that uptake of arsenite was up to 10-fold higher in the presence of antimonite. Conversely, antimonite extrusion was found to be inhibited by arsenite, with the rate of uptake of arsenite and antimonite becoming equal at equimolar concentrations.

This behaviour cannot be explained by a co-transport model (where arsenite would be expected to stimulate antimonite transport just as antimonite stimulates arsenite), or by a competitive model (where both arsenite and antimonite would be expected to inhibit transport of the other), leading Meng et al. [100] to further speculate that the true substrate for ArsB is in fact some salt of arsenite, antimonite or a combination, but not the anion itself.

Prior to Meng et al. [100] ArsB was hypothesised to be an anion uniporter[84], due to the fact that export was dependent on ion gradient. However, confirmation that ArsB is capable of antimonite transport challenged this, as the $pK_a$ of antimonite is 11.8, meaning that virtually all antimonite and arsenite present at physiological pH levels are neutral compounds. The dependence on NADH indicates that ArsB clearly performs more than assisted diffusion based transport, and must rely on some mechanism by which extrusion of a neutral compound is driven by the ion gradient. Since ArsB transport was found to be independent of any specific cation[100], it seems likely that exchange with $H^+$ is a likely mechanism for extrusion, making ArsB a possible metalloid-proton antiporter.

### YqcK

The functions of ArsR, ArsB, and ArsC of regulation of the operon, extrusion of arsenite and reduction on arsenate have been well established as discussed above. A fourth Open Reading Frame (ORF) exists between the *arsR* and *arsB* genes, named *yqcK*, whose function has not been demonstrated and is often not present in other arsenic operons such as the homologous *ars* operon in *Staphylococcus*[141]. A year after the *ars* operon was first reported in the skin element of *B. subtilis*, Rosen [134] noted that both the hypothetical gene product YqcK and ArsD protein from *E. coli* R773 contain 3 pairs of vicinal cysteine residues. In ArsD these vicinal pairs have been shown to have a role in metalloid binding[88], suggesting that they may perform a similar function in YqcK.

ArsD is believed to have two functions in *E. coli*. The first is as a secondary regulator of the *ars* operon, as ArsD binds to the same DNA binding site as ArsR, but with a lower affinity[170]. As with ArsR, the interaction between ArsD and DNA is relieved by arsenite, but more arsenite is required to give the same magnitude of reduction in bound fraction than with ArsR – DNA, implying that ArsD is less sensitive to arsenic[28]. These data suggest that ArsD is responsible for regulating the response of the operon at high levels of arsenite. At low levels of arsenite, ArsR out-competes ArsD for binding to the operator region, and thus control of the operon is dominated by the interactions of ArsR, DNA, and metalloid. At

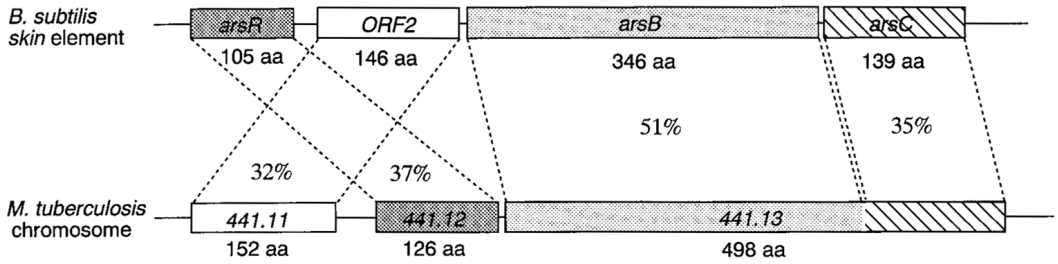



Fig. 3.4 Comparison of the *B. subtilis ars* operon and the *ars* operon of *M. tuberculosis*, reproduced from Sato and Kobayashi [141]. *ORF2* is now *yqcK* and 441.11, 441.12, and 441.13 have since been renamed *cadI*, *arsR*, and *arsC* respectively.

higher levels of arsenite, ArsR is less able to bind DNA due to allosteric inhibition and so the interaction between ArsD and arsenite becomes important. A possible explanation for this dual level control is that ArsD prevents over-expression of ArsB in high arsenite conditions, as ArsB itself can be toxic at high concentrations[28, 169].

A second function of ArsD was found much later by Lin et al. [89], where the protein was shown to have metallochaperone activity. ArsD binds tightly with arsenite in the cytosol, presumably reducing the toxic effect of the metalloid on the rest of the cell. The ArsD – $As^{III}$ complex then interacts with the membrane bound ArsAB, transferring arsenite into the ATP driven pump and out of the cell. A follow-up study[90] found that of the three vicinal cysteine pairs ($Cys^{12-13}$, $Cys^{112-113}$, and $Cys^{119-120}$) only the first together with another cysteine residue ($Cys^{18}$) were required for metallochaperone activity, and it is these three residues that were conserved across ArsD variants in other organisms.

It is possible that YqcK functions similarly to ArsD as a secondary repressor, however global alignment by Needleman-Wunsch algorithm of the protein sequences of the two show less than 15% amino acid identity, and the three essential cysteine pairs do not align to each other.

Another theory is possible which would explain the presence of Cysteine pairs in YqcK as well as the low sequence identity between YqcK and ArsD. The original paper describing the *B. subtilis ars* operon noted the similarity between the operon and a region of the chromosome of *Mycobacterium tuberculosis*, see Figure 5 of Sato and Kobayashi [141], reproduced in Figure 3.4. This region contains two ORFs with 37% and 51% amino acid similarity to $ArsR_{BS}$ and $ArsB_{BS}$ respectively, suggesting that the region encodes an operon that confers arsenic resistance in a similar manner to other *ars* operons. No ORF exists between the putative *arsR* and *arsB* homologs in *M. tuberculosis*, however an ORF exists immediately upstream of the putative arsR homologue which shows 32% amino acid homology with *B. subtilis* YqcK. This potential *M. tuberculosis* homologue was later characterised in Hotter et al. [67], where it was found that expression of the gene is induced by cadmium, specifically

$Cd^{II}$, and thus the ORF was renamed *cadI*. While the exact function of this gene remains unknown, it was recently shown[87] that its expression is also regulated by the downstream ArsR variant, although this ArsR variant's response to arsenic has not been demonstrated.

Further weight is added to the theory that *B. subtilis* YqcK is a homologue of *M. tuberculosis* cadI comes from the observation that while deletion of the *ars* operon in wild type *B. subtilis* strains has no effect on $Cd^{II}$ resistance, deletion of *ars* in Δ*cadA* strains reduces growth in the presence of $Cd^{II}$[102]. A possible explanation for this observation is that YqcK is indeed a homologue for *M. tuberculosis* cadI, which provides some form of secondary protection from $Cd^{II}$ which is independent of the previously characterised CadA mediated membrane transport (described in Tsai et al. [156]).

### The *ase* operon

As noted above, *B. subtilis* contains two arsenic related operons, the *ars* operon discussed above and the simpler *ase* operon. The *ase* operon contains only two genes, *aseR*, an *SmtB/ArsR* family regulatory gene responsive to arsenite and *aseA* (previously known as *ydfA*), a putative membrane protein which exports arsenite in a manner similar to ArsB.

The most detailed study of the *ase* operon to date is Moore et al. [102], which confirmed that aseR regulates expression of itself and AseA in response to $As^{III}$, and that the operon is capable of conferring some level of arsenite resistance independently of the *ars* operon, but that the lack of a reductase means that the *ase* confers no resistance to arsenate. Moore et al. [102] found allosteric regulation of AseR to be more specific for $As^{III}$ than ArsR (which is also induced by $Sb^{III}$), but that overall expression of the operon was weaker than for *ars*.

The existence of the *ase* operon leaves us with a choice when designing a biosensor in *B. subtilis*, as such a sensor could be based on either the *ars* or *ase* operon. While the *ase* operon appears to be more specific for $As^{III}$ than the *ars* operon based on available data, it has been the subject of significantly less study. The strength of the *ase* operon when fully induced is also less than that of the *ars* operon, reducing the maximum achievable dynamic range between an high and low signal for an *ase* sensor. While a hybrid promoter system may be possible – where a stronger promoter is combined with the AseR operator region is used in place of the natural *ase* promoter – the lack of published data on the *ase* operon makes such an approach more challenging.

Conversely, cross-talk between antimony and arsenic by the *ars* operon may be less problematic for a practical sensor than first thought as antimony is typically present at much lower concentrations in groundwater sources than arsenic[98, 166]. For these reasons, it was decided that prelimiary work on a *B. subtilis* based arsenic biosensor should leverage the *ars* operon, with the *ase* operon born in mind should antimonite interference become an issue.

## 3.2 Developing a Mathematical Model

### 3.2.1 Motivation

Since the beginning of the arsenic contamination crises, the mechanisms through which prokaryotes interact with arsenic has been the topic of considerable research. Much of this research has focussed on characterising the individual components of this response, such as how arsenic enters the cell, the mechanism by which $As^{V}$ is reduced to $As^{III}$, the interactions between the regulatory protein, DNA, and arsenite, and the varying mechanisms by which arsenite is extruded from the cell. From these isolated experiments, a clear qualitative picture has emerged of how many organisms mitigate arsenic toxicity.

This qualitative picture is necessary but not sufficient for engineering purposes. To reliably and intelligently engineer an arsenic biosensor, we need to first understand how the individual processes of the *ars* operon come together to produce the observed response.

Questions that modelling of the *ars* operon might reasonable seek to answer include

- what properties of the system determine the rate of leaky expression from the *ars* promoter in the absence of arsenic?
- which properties determine the maximum rate of expression under saturating arsenic conditions?
- what is the predicted form of the response at concentrations between these extrema?
- can the form of the response be changed by making changes to the model parameters alone?
- what are the likely effects of making simple perturbations to the system such as changing the promoter strength or copy number of the genes?
- what type perturbations are most likely to increase desirable sensor properties while keeping undesirable properties to an acceptable level?

It is these questions that this chapter seeks to answer in the hope that those answers and their interpretation will help inform the rational design of a functional sensor.

### 3.2.2 Modelling Strategy

A three step process was used in this chapter to model the behaviour of a system.

1. Express the system as a series of chemical reactions

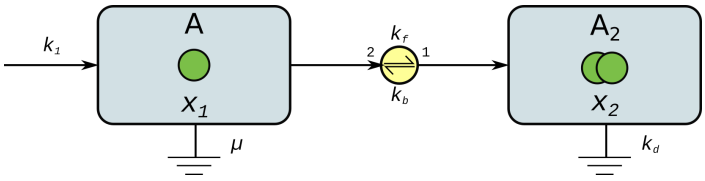



Fig. 3.5 A simple model of reversible dimerisation. $x_1$ is produced spontaneously at rate $k_1$, before binding reversibly with itself to form $x_2$. Both $x_2$ and $x_2$ degrade at rate $\mu$.

2. Convert those reactions into a set of Ordinary Differential Equations (ODEs), one for each species

3. Solve the resultant simultaneous ODEs

For simple systems, it is often possible to arrive at an analytical solution for the steady state of the system, but this rapidly becomes intangible when networks get large with several reversible reactions or exhibit cooperativity. In such cases, we need to arrive at a solution numerically.

### A Simple Example

An example system is shown in Figure 3.5. This reaction has only two species and four reactions, which can be represented as chemical equations as

$$\xrightarrow{\mathrm{k_1}} \mathrm{A}$$
$$\mathrm{A} \xrightarrow{\mu}$$
$$2\,\mathrm{A} \underset{\mathrm{k_b}}{\overset{\mathrm{k_f}}{\rightleftharpoons}} \mathrm{A_2}$$
$$\mathrm{A_2} \xrightarrow{\mu}$$

In this exploratory example it is simple enough to write down the equations governing the concentrations of each of the components by summing the edges entering or leaving each node in Figure 3.5. Defining the vector $\boldsymbol{x}(t) = [x_1(t), x_2(t)]$ as the concentrations of the species A and $\mathrm{A_2}$ respectively, we get

$$\dot{\boldsymbol{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} k_1 - \mu x_1 - 2k_f x_1^2 + 2k_b x_2 \\ k_f x_1^2 - k_b x_2 - \mu x_2 \end{bmatrix}$$

The steady state solution for this system can be arrived at analytically by setting $\dot{\boldsymbol{x}} = 0$ and substituting the equation for $\dot{x}_2$ into the equation for $\dot{x}_1$ such as to eliminate $x_2$, then solving

the resulting quadratic in $x_1$ to give

$$x_1 = \frac{-\mu + \sqrt{\mu^2 + 8k_1k_f\left(1 - \frac{1}{1+\frac{\mu}{k_b}}\right)}}{4k_f\left(1 - \frac{1}{1+\frac{\mu}{k_b}}\right)}$$

Even with this simple system, this equation for $x_1$ is not particularly convenient or informative, although we can tell that since our rate constants $k_1, k_f, k_b, \mu \geq 0$ there is exactly one positive solution for $\mu \neq 0$ and none otherwise. If we assume that $A_2$ is a relatively stable dimer, i.e. that $\mu >> k_b$, the ratio $\frac{\mu}{k_b}$ becomes very large allowing us to simplify our expression for $x_1$:

$$x_1 = -\mu + \sqrt{\mu^2 + 8k_1k_f}$$

### A Generalised Approach

Having seen a simple example, let us consider a solution for the general case of a system with $N$ species $\{X_1, \ldots, X_N\}$ and $M$ reactions, where the $m^{\text{th}}$ reaction is given by

$$s_{m,1}X_1 + s_{m,2}X_2 + \cdots + s_{m,N}X_N \xrightarrow{k_m} p_{m,1}X_1 + p_{m,2}X_2 + \cdots + p_{m,N}X_N$$

where $s_{m,n}$ and $p_{m,n}$ represent the stoichiometry of the $n^{\text{th}}$ substrate and product of reaction $m$, respectively, and $k_m$ is the $m^{\text{th}}$ rate constant. Reversible reactions are modelled as the linear combination of the forward and reverse reactions, which are computed separately. Defining the vector $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ as the concentrations of the species $\{X_1, \ldots, X_N\}$ we can deduce the rate of the $m^{\text{th}}$ reaction, $r_m$, by applying the laws of mass-action as

$$r_m(\boldsymbol{x}) = k_m \prod_{n=1}^{N} x_n^{s_{m,n}} \tag{3.6}$$

To calculate the rate of change in concentration of a component caused by a reaction, we multiply the rate of the reaction by the net change in stoichiometry due to that reaction. This can be accomplished efficiently by matrix multiplication, and by defining the stoichiometry matrix $\boldsymbol{S}$ and rate vector $\boldsymbol{r}$ we can express the governing equation of a general reaction network in the standard form

$$\boldsymbol{f}(\boldsymbol{x}) = \dot{\boldsymbol{x}} = \boldsymbol{S}\boldsymbol{r}(\boldsymbol{x}) \tag{3.7}$$

Where

$$\boldsymbol{S} = \begin{bmatrix} p_{1,1} - s_{1,1} & p_{2,1} - s_{2,1} & \cdots & p_{m,1} - s_{m,1} \\ p_{1,2} - s_{1,2} & p_{2,2} - s_{2,2} & \cdots & p_{m,2} - s_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1,n} - s_{1,n} & p_{2,n} - s_{2,n} & \cdots & p_{m,n} - s_{m,n} \end{bmatrix} \qquad \boldsymbol{r} = \begin{bmatrix} r_1 \\ \vdots \\ r_M \end{bmatrix}$$

This governing equation can now be used to numerically simulate the behaviour of any arbitrary system given a set of initial conditions $\boldsymbol{x}_0$ and fixed rate parameters $\boldsymbol{k}$. Details of how this was implemented in software are given in Section 3.2.2 below.

### Extending to Michaelis Menten Kinetics

Enzymatic reactions, where a species catalyses a particular reaction without itself being consumed in that reaction, are a common feature in biology, with the reduction of arsenate to arsenite by *ArsC* or the extrusion of arsenite by *ArsB* being obvious examples in the system we wish to model. A typical reaction of this kind could be represented as

$$\mathrm{E} + \mathrm{S} \underset{\mathrm{k_{rv}}}{\overset{\mathrm{k_{fw}}}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{\mathrm{k_{cat}}} \mathrm{E} + \mathrm{P}$$

where enzyme E binds reversibly to substrate S producing the complex ES which then produces product P and free enzyme. This interaction is commonly simplified according to the well known Michaelis-Menten relationship, where the rate of the reaction is equal to

$$r_{MM} = \frac{k_{cat}[E][S]}{K_M + [S]} \tag{3.8}$$

where $k_{cat}$ and $K_M$ are the maximum catalytic rate and the Michaelis constant, the substrate constant at which a rate of half the catalytic rate is observed, respectively.

We could simulate reactions of this type using three general irreversible reactions (as described in Section 3.2.2), however numerical values for $k_{fw}$ and $k_{rv}$ are rarely available, while estimates of $k_{cat}$ and in particular $K_M$ are more commonly available as they are easier to quantify experimentally. Simulating Michaelis-Menten dynamics thus reduces the number of unknown variables in the model, at the expense of a few model assumptions, notably that the concentration of the enzyme is much lower than the substrate, that reactions are diffusion limited, and that the formation of the product P is irreversible.

We thus define the general Michaelis-Menten reaction as

$$s_1X_1 + s_2X_2 + \cdots + s_NX_N \xrightarrow{\mathrm{X_E},\, \mathrm{k_{cat}},\, \mathrm{K_M}} p_1X_1 + p_2X_2 + \cdots + p_NX_N$$

with all symbols as defined previously and $X_E$ as the species responsible for enzymatic activity. The rate of equations of this type, $r_{MM}$ is given by substituting into Equation 3.8 to give

$$r_{MM} = \frac{k_{cat}x_E \prod_n^N x_n^{s_n}}{K_M + \prod_n^N x_n^{s_n}} \tag{3.9}$$

when a Michaelis-Menten reaction is encountered, all that is necessary is to calculate the relevant value in the rate vector $\boldsymbol{r}$ according to this equation rather than Equation 3.6 and the governing equation can then be found as before.

### Finding the Jacobian

Numerical methods exist for finding the solution to the general steady state problem

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$$

with only a set of initial conditions. However, most of these methods involve making many repeated calls to $\boldsymbol{f}$ in order to estimate the Jacobian matrix at each point. The Jacobian is a matrix of all the first order partial derivatives of the system at a given point, in other words it represents the rate of change of each output with respect to each input at a particular position. The information encoded in this matrix is useful in optimisation problems as it allows the solution to be found using algorithms based on gradient descent, where an initial estimate is refined by travelling down the gradient until a minimum is found.

We can reduce the computational work and increase the numerical accuracy by calculating the Jacobian directly. The Jacobian of $\boldsymbol{f}$ is given by

$$\boldsymbol{J_f} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_N}{\partial x_1} & \cdots & \frac{\partial f_N}{\partial x_N} \end{bmatrix} \qquad \text{where} \quad \boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \\ \vdots \\ f_N(\boldsymbol{x}) \end{bmatrix}$$

Applying the Jacobian to Equation 3.7 and making use of linearity as

$$\boldsymbol{J_f} = \boldsymbol{J_{Sr}} = \boldsymbol{SJ_r}$$

shows that we can find the Jacobian of $\boldsymbol{f}$ by first calculating the Jacobian of $\boldsymbol{r}$ and premultiplying by the stoichiometry matrix $\boldsymbol{S}$.

For a general irreversible reaction, $m$, we can calculate each element of $\boldsymbol{J_r}$ by taking partial derivatives of Equation 3.6 leading to

$$\frac{\partial r_m}{\partial x_p} = s_{m,p} x_p^{s_{m,p}-1} k_m \prod_{\substack{n \\ n \neq p}} x_n^{s_{m,n}}$$

which is relatively simple to compute. When the $m^{\text{th}}$ reaction is a Michaelis-Menten reaction, however, we instead take partial derivatives of Equation 3.9 which is slightly more involved. Writing Equation 3.9 as

$$r_m = \frac{k_{cat} x_E \prod_n^N x_n^{s_{m,n}}}{K_M + \prod_n^N x_n^{s_{m,n}}} = \frac{g_m}{h_m}$$

allows us to use the product rule to compute the Jacobian

$$\frac{\partial r_m}{\partial x_p} = \frac{\frac{\partial g_m}{\partial x_p} h_m - g_m \frac{\partial h_m}{\partial x_p}}{h_m^2}$$

where

$$\frac{\partial h_m}{\partial x_p} = s_{m,p} x_p^{s_{m,p}-1} \prod_{\substack{n \\ n \neq p}} x_n^{s_{m,n}}$$

and, through use of the product rule,

$$\frac{\partial h_m}{\partial x_p} \overset{p \neq E}{=} s_{m,p} x_p^{s_{m,p}-1} x_E k_{cat} \prod_{\substack{n \\ n \neq p}} x_n^{s_{m,n}}$$

$$\frac{\partial h_m}{\partial x_p} \overset{p = E}{=} (s_{m,p} + 1) k_{cat} \prod_n x_n^{s_{m,n}}$$

which allows us to find the Jacobian for any general reaction network.

### Constraining the Optimisation

The concentration of any physical species must be non-negative. Mathematically, this translates as $N$ inequality constraints that $x_i \geq 0 \forall i$ and while one could use the generalised Lagrange multiplier method to apply these constraints, a far easier method exists in this special case. For each $x_i$ we define $z_i$ such that $x_i = z_i^2$, and find the roots of $\dot{\boldsymbol{z}} = \boldsymbol{0}$. To find

the $\dot{z}_i$, we can use the chain rule to show

$$\frac{dz_i}{dt} = \frac{dx_i}{dt} \cdot \frac{dz_i}{dx_i} = f_i(\boldsymbol{x}) \cdot \left[ \frac{d}{dz_i} \left( z_i^2 \right) \right]^{-1} = \frac{f_i(\boldsymbol{z}^2)}{2z_i}$$

where $\boldsymbol{z}^2 = [z_1^2, z_2^2, \cdots, z_N^2]$. A similar calculation can be made to update the Jacobian matrix. One negative consequence of this, is that any state where any variable is zero becomes numerically unstable. The simplest solution to this is to numerically estimate the limit at $z_i = 0$ by setting $z_i$ to some very small value.

In some situations, we may need to constrain the solution space of the optimisation to avoid trivial solutions. For example, consider the reaction network

$$\begin{aligned} p &\xrightarrow{\mathrm{k_r}} p + \mathrm{r} & \mathrm{r} &\xrightarrow{\mu} \\ p + \mathrm{r} &\xrightarrow{\mathrm{k_d}} \mathrm{q} & \mathrm{q} &\xrightarrow{\mu} p \end{aligned}$$

where promoter $p$ produces the negative regulator r which binds to $p$ to produce q and also decays at rate $\mu$. If we attempt to solve this system in the usual way, we define our state vector $\boldsymbol{x} = [p, r, q]$ and generate the governing equation

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} \dot{p} \\ \dot{r} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} \mu q - k_r p r \\ k_r p - \mu r - k_d p r \\ k_d p r - \mu q \end{bmatrix} = \boldsymbol{0}$$

which is defective as equations for $\dot{p}$ and $\dot{q}$ are not independent, leading to an infinite number of solutions. Most numerical solvers will in fact converge on the trivial solution $\boldsymbol{x} = \boldsymbol{0}$, regardless of the initial conditions.

What this system is missing is the constraint that the total bound and unbound promoter is constant, i.e. $p + q = C$ where $C$ is a constant equal to the copy number of the promoter. Since constrained numerical optimisation algorithms are typically more complex than unconstrained ones, it is easier to reform the constrained system as an unconstrained one than numerically solving the constrained problem directly. There are two simple ways of achieving this. One method involves re-arranging the constraint and substituting $q = C - p$ in the governing equation and removing $q$ from the state vector $\boldsymbol{x}$. This has the benefit of reducing the dimensionality of the problem, but introduces a new inequality constraint, namely that in order to guarantee $q \geq 0$ we must have $p \leq C$, which is harder to apply than the original constraint.

Instead, we can make use of Lagrange multipliers to convert our constrained problem into an unconstrained one by adding one extra dimension. Rearranging our constraint as

$p+q-C=0$ we can define the unconstrained state vector $\boldsymbol{x}_u$ as

$$\boldsymbol{x}_u = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \\ \mu_1 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \\ p+q-C \end{bmatrix}$$

such that a solution of $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$ is necessarily a solution to our constraint.

### Limitations of ODE Modelling

The most obvious limitation of ODE modelling is that ODE models are inherently deterministic. For any given state vector $\boldsymbol{x}$ and set of model parameters $\boldsymbol{K}$, the full evolution of the system can be calculated. Because cellular processes such as transcription and translation rely on molecular interactions between small number of particles, these processes are essentially randomised. Real cells therefore exhibit random stochastic behaviour which is not captured by ODE models.

In many cases the random variability of individual cells is "averaged out" when bulk average behaviour is observed, and the average behaviour of a large population of cells closely follows the predicted ODE model behaviour. In other cases, nature has exploited this inherent noise in order to confer an evolutionary benefit. A classic example of this in *B. subtilis* is the induction of competence, when cells in stress conditions attempt to uptake DNA from their environment in the hope that this DNA will improve the cells' resistance to the stress.

When a single *B. subtilis* cell enters the competence pathway, its ability to grow and divide is reduced. If every cell in a population were to simultaneously become competent until DNA had been taken up, the growth of the overall population would be significantly diminished, potentially allowing another species of non-competent bacteria to come to dominate. Instead, members of the population become competent stochastically, and for varying amounts of time.

This behaviour maximises the possibility of taking up beneficial DNA while keeping the impact on growth rate to acceptable levels. Since the rate of uptake of DNA depends on the external DNA concentration, there is for any given DNA concentration an optimal duration of competence which achieves this. Individual bacteria are unable to rapidly sense external DNA, and so instead of attempting to calculate this optimal window, the transition from competence to normal growth is also stochastic, leading to a broad range of competence

duration. The population as a whole benefits from this "bet hedging" behaviour as some individuals will successfully take up DNA with minimal detriment to their growth rate, allowing the species to overcome the stress conditions in the shortest possible time. In this instance, the stochastic behaviour exhibited by individuals is not completely "averaged out", and while a deterministic model may accurately reproduce the average behaviour, it fails to fully capture the complex behaviour of the system.

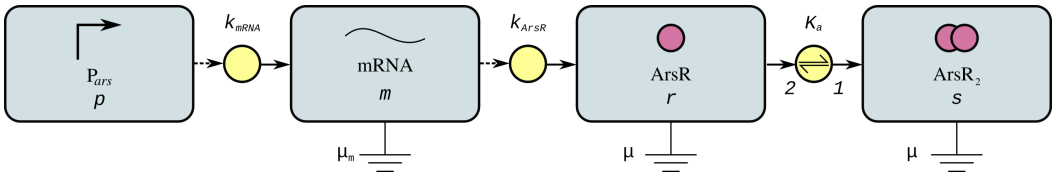
Is seems quite possible that this bet hedging behaviour is applied to arsenic contamination as well. While population averages confirm that expression of the *ars* operon is very low under anarsenic condidions, it seems possible that some memebers of this population may express arsenic responsive genes at high levels. While the energy expended expressing the extra genes would be an impediment to the indivudial cells, the population oveall would be better protected against a sudden increase in arsenic contamination as some of its members would already have the machinery in place for dealing with the threat and thus be minimally affected by the sudden condition change.

While bet-hedging behaviour has been observed in some antibiotic resistance strategies, it has not been observed in the case of arsenic, and it is unclear what the regulatory mechanism for such behaviour would be. However, the possibiliy cannot be discounted out of hand until this behaviour has been confirmed or otherwise by experimental evidence.

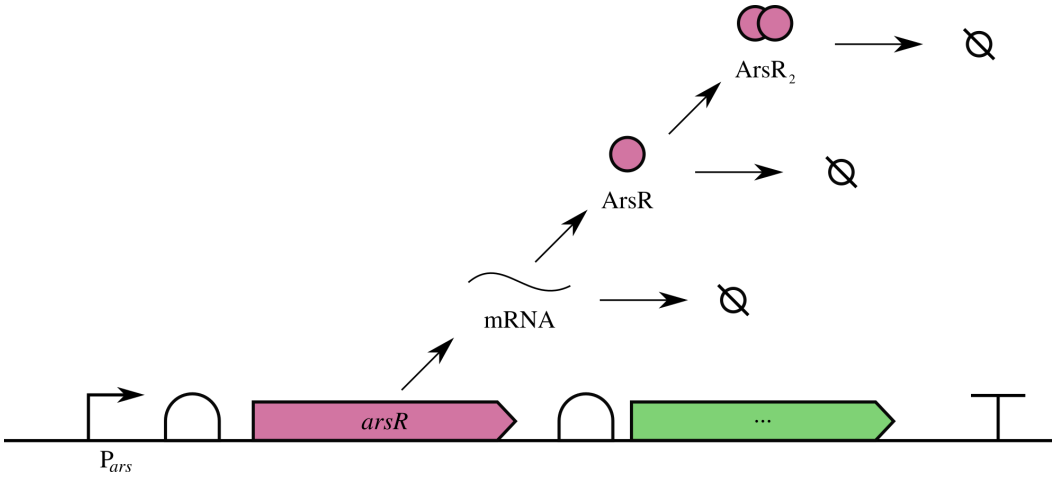
### Software Implementation

Commercial packages exist which are able to perform some of the calculations above, such as MATLAB's SimBiology. However, they are not very efficient in finding steady state equilibria, and have poor programmatic interfaces, as well as being highly proprietary. Instead, the above code was implemented as a python package called *pysim* which is available from github at https://github.com/haydnKing/pysim.

A model file is used to specify all the parameters, species, constraints and reactions of the model. From this file, *pysim* is able to automatically generate the governing equation and Jacobian of the system, and supports changes being made to parameters at run-time. The syntax for this file is introduced in appendix C. The governing equation and Jacobian are 'compiled' as closures to run as efficiently as possible, and the steady state is then solved using the *fsolve* function from the *scipy.optimize* optimisation toolkit, which leverages the MINPACK[104] library.




(a) Visual representation of the mathematical model. Species are shown in rectangles, circles represent reactions and a double headed arrow represents a reversible reaction. A dotted line indicates that a species is not consumed by the reaction, numbers next to reactions indicate the stoichiometry of the reaction. The electrical earth symbol represents a dilution/degradation reaction with no products.




(b) SBOL representation of the system

Fig. 3.6 Open-loop production of the $ArsR_2$ homodimer in the absence of arsenic and with interaction between $ArsR_2$ and $P_{ars}$ removed.

### 3.2.3 Open-loop production of $ArsR_2$

Figure 3.6 shows a model of the open-loop production of the homodimer $ArsR_2$ – referred to as open-loop because the regulatory interaction between $ArsR_2$ and the promoter $P_{ars}$ has been ignored for the time being. The promoter $P_{ars}$ produces mRNA which in turn produces protein ArsR, which binds reversibly with itself to form the homodimer. Neither the promoter nor mRNA are consumed by these reactions, but the mRNA is diluted and degraded at rate $\mu_m$ while both proteins are diluted at rate $\mu$. Protein degradation is ignored, as it is assumed to be considerably slower than growth dilution at high growth rates such as during exponential growth. mRNA degradation is not ignored, with the term $\mu_m$ corresponding to $\mu + \phi$, growth dilution plus degradation. The concentration of the promoter $P_{ars}$ is not diluted, as DNA replication holds this approximately constant as cell mass increases.

In this simple case, we can quite easily find an expression for the total ArsR in the steady state, which can be found by solving a quadratic in $s$. However, we can use our knowledge of the real-world system to simplify this model. Firstly, we know that $ArsR_2$ is a very stable dimer, with very little monomeric ArsR being found in cell extracts even in conditions with high arsenic. This implies that the rate of the dissociation reaction is small compared with the rate of the association reaction, leading to a large association constant $K_a$. This has been confirmed as values for $K_a$ for other members of the *SmtB/ArsR* family (shown in Table 3.1) are in the range of $3 \times 10^5\,\text{M}^{-1}$ to $3 \times 10^6\,\text{M}^{-1}$. Therefore, if we assume that $K_a$ is large, specifically such that the forwards rate of the reaction is much larger than $\mu$, we can neglect the steady state concentration of monomeric ArsR or $r$, we can write the approximate steady state solution for $s$ as

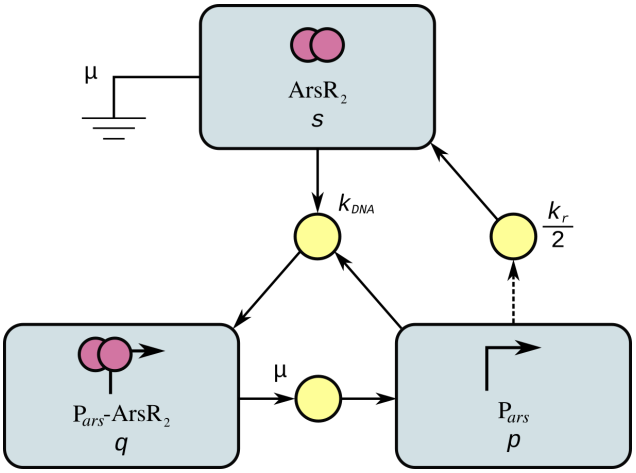
$$s \approx \frac{k_r}{2\mu} p$$

where $k_r := \frac{k_{mRNA} k_{ArsR}}{\mu_m}$.

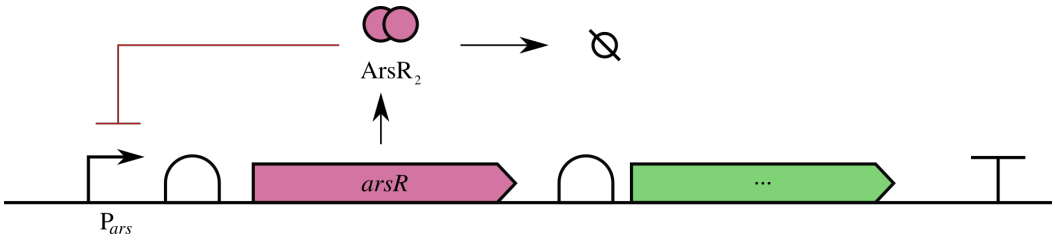
Such a simplification allows us to greatly reduce the dimensionality of our models, as this is equivalent to the promoter $P_{ars}$ producing $ArsR_2$ directly with rate constant $\frac{k_r}{2}$ diluted at rate $\mu$. This assumption was tested on all later models described in this chapter, and found to have minimal effect on the outcome of the model for all physiologically plausible values of $K_a > 10^2\,\text{M}^{-1}$.

### 3.2.4 A Simple Model Without Arsenic

A simple model showing repression of the system in the absence of arsenic is shown in Figure 3.7. In this simple model, unbound promoter $P_{ars}$ directly produces the dimer $ArsR_2$ using the assumptions outlined above. The control loop is now closed, with the dimer binding




(a) Visual representation of the mathematical model with three species, unbound promoter $p$, homodimeric $ArsR_2$ $s$, and bound promoter $q$. See figure 3.6 for key




(b) SBOL representation of the system

Fig. 3.7 A model of the ars operon in the absence of arsenic. Here, $ArsR_2$ is produced directly and inhibits the promoter $P_{ars}$, leading to a reduction in the rate at which $ArsR_2$ is produced due to negative feedback.

with the promoter at rate $k_{DNA}$, preventing its own expression. Using similar logic as in Section 3.2.3 when treating the dimerisation reaction, this model only considers the binding of $ArsR_2$ with DNA and neglects the dissociation rate. Since DNA replication does not also replicate the bound $ArsR_2$ repressor, bound promoter is also diluted at rate $\mu$, producing unbound promoter.

After such simplification, experiment showed that the behaviour of the system depends on two non-dimensionalised model parameters – $\frac{k_r}{2\mu}$ and $\frac{k_{DNA}}{\mu}$. If we normalise our unit of concentration such that the total promoter concentration is $p+q=1$, then the first parameter relates to the basal, unrepressed strength of the promoter and is equal to the total number of copies of $ArsR_2$ per cell under the open-loop system. While this has not been quantified, we can place upper and lower bounds on this value by observing that Maass et al. [94] found the range of protein copy number in *B. subtilis* to range from around $10^1$ to around $10^5$ with a mean of $4.5\times10^3$ copies.

Figure 3.8 shows a heat-map of the unbound promoter fraction, $p$, over this range of $\frac{k_r}{2\mu}$ and for a chosen range of $\frac{k^+_{DNA}}{\mu}$. Proper interpretation of this plot can tell us about two interesting states of the operon – leaky expression without arsenite, and the fully induced state under saturating arsenite conditions.

In the case where arsenite is absent, we can investigate the model parameters which determine the level of leaky expression. Furthermore, we can see that for a fixed growth rate, $\mu$, and DNA binding rate, $k_{DNA}$, increasing the basal strength of the promoter will decrease the level of leaky expression, equivalent to moving to the right in Figure 3.8.

We can also predict an upper limit of the behaviour at very high arsenic concentrations by considering the system's behaviour in saturating arsenic conditions. When the concentration of $As^{III}$ is very large, the rate at which arsenite binds with $ArsR_2$ is much greater than the rate at which $ArsR_2$ is produced, leaving very little $ArsR_2$ which has not formed a complex with arsenite. Thus we can estimate an upper bound for maximal induction of the system by simply replacing $ArsR_2$ for $ArsR_2-As^{III}$, and thus $K^+_{DNA}$ for $\frac{k^+_{DNA}}{\alpha}$, where $\alpha$ is the fold decrease in DNA binding rate of $ArsR_2$ caused by interaction with $As^{III}$.

Interestingly, Figure 3.8 shows that a fold reduction in $\alpha$ of between 100 and $1,000$ in $k^+_{DNA}$ is enough to transition from the operon being tightly repressed to the promoter being almost completely unbound. This is closely matched by the change in $K_{DNA}$ that has been observed for other members of the SmtB/ArsR family of between 300 and $1,000$, see Table 3.1. While we would expect this change in $K_{DNA}$ to be caused by both a combination of a reduction in binding rate, $k^+_{DNA}$, and an increase in unbinding rate, $k^-_{DNA}$, modelling suggests that a decrease in $k^+_{DNA}$ alone is enough to trigger induction, at least in growth phases where $\mu$ is large.

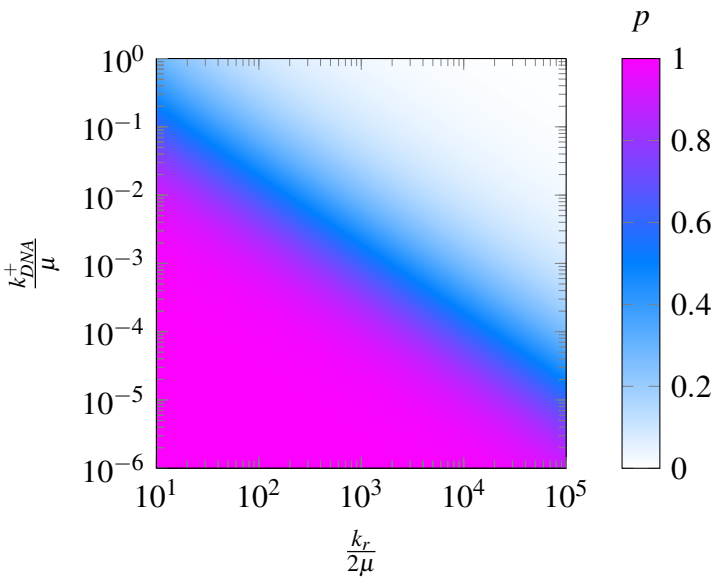



Fig. 3.8 Parameter variations of the arsenic-free model from Figure 3.7 for variations in the non-dimensional quantities $\frac{k_r}{2\mu}$ and $\frac{k^+_{DNA}}{\mu}$.

The range between a high and low signal is critical for biosensor function, and since we can now estimate the maximum possible range of the system, we can say something about the effect of variations in promoter strength on that range. From the observed behaviour of the system, it is reasonable to assume that the value of $k^+_{DNA}$ places the uninduced system somewhere above the transition shown in Figure 3.8, while induction by arsenite moves the stable point somewhere below it. Since increasing the strength of the promoter is equivalent to moving to the right in figure, leaky expression is expected to decrease, but as we move further to the right, the upper bound of maximal expression will also begin to decrease. This implies that for a given $k^+_{DNA}$, $\alpha$, and $\mu$, there is a particular basal promoter strength, $k^*_r$, which maximises the range of the response, as is demonstrated Figure 3.9.

This model has shown that a physiologically plausible value of $\alpha$ is enough to cause complete de-repression under saturating arsenic conditions. Further models will therefore neglect the binding of $ArsR_2-As^{III}$ complexes with DNA, essentially taking the limit of $\alpha \to \infty$. This is likely to be valid for physiological model parameters, however, care should be taken when varying $k_r$, the strength of the promoter, without varying $k^+_{DNA}$, as the true values of $k^+_{DNA}$ and $\frac{k^+_{DNA}}{\alpha}$ may no longer straddle the transition shown in Figure 3.8, causing a reduction in dynamic range.

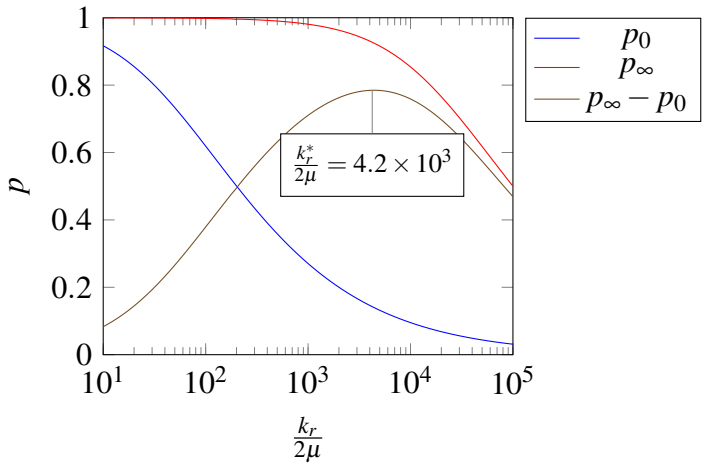



Fig. 3.9 Effect of relative promoter strength on the range of the response. $K^+_{DNA}$, $\alpha$, and $\mu$ were held constant at $3.46 \times 10^{-4}$, 500, and $3.46 \times 10^{-2}$ respectively. The value for $\mu$ was chosen to correspond to a doubling time of approximately 20 minutes and $\alpha$ was chosen based on literature values from Table 3.1. $K^+_{DNA}$ was set arbitrarily such that $K^*_r$ would fall within the expected range for $K_r$.

### 3.2.5 De-repression by Arsenic

The model shown in Figure 3.10 represents a stepping stone on the way to developing a full model of the system in which the internal concentration of unbound $As^{III}$ is held constant. Again, it is assumed that the values of $K^-_{DNA}$ are much less than the growth rate $\mu$, and $\alpha$ is assumed to be large enough that the rate at which $ArsR_2 - As^{III}$ binds with DNA, $\frac{K^+_{DNA}}{\alpha}$ is much less than growth dilution, $\mu$, and is thus ignored.

In this model, only one $As^{III}$ ion binds to $ArsR_2$, with association constant $K_{As}$. In reality, two $As^{III}$ can bind, one to each member of the dimer, but this model assumes that the second binding event has no further effect on DNA-$ArsR_2$ interactions. To justify this assumption, let us consider the case where an arsenite ion has bound to the $\alpha$ domain of one member of the homodimer, but that the other remains unbound. Since the presence of arsenite has no observable effect on the stability of the dimer, it seems unlikely that an arsenite induced conformational change in one member of the complex is transmitted to the second, as this would change the nature of the bound surfaces between the proteins, which would be expected to have some effect on the stability of the complex. If we assume this to be true, then one member of the homodimer remains able to bind DNA with high affinity, while the other cannot. Experiments have shown that mutant ArsR proteins which were unable to form homodimers were unable to bind DNA, implying that either the homodimeric interaction itself causes some other conformational change to enable DNA binding, or, more

probably, the combined and cooperative interactions between both members of the dimer and DNA are required for the complex to bind. On balance, it therefore seems unlikely that the binding of a second $As^{III}$ ion has a large effect on the DNA binding of $ArsR_2$, certainly when compared with the effect of the first ion.

Figure 3.11 shows the effect of parameter variations on the response curve of this model from low to high arsenite concentration. In each case, the response increases monotonically from a basal, leaky response to a higher maximal response, while the gradient of the curve decreases smoothly towards zero. The form of this function is very similar to exponential decay such as $a(1-e^{-b\cdot t})+c$, ratios such as $\frac{a}{(b+x)^c}$ or sigmoidal functions such as the Gompertz function as well as hyperbolic functions such as $tanh$. Unfortunately, while such functions tend to fit well for certain model parameter values, each one fits particularly poorly at other parameter values. As such, there appears to no convenient general form for the response of the model.

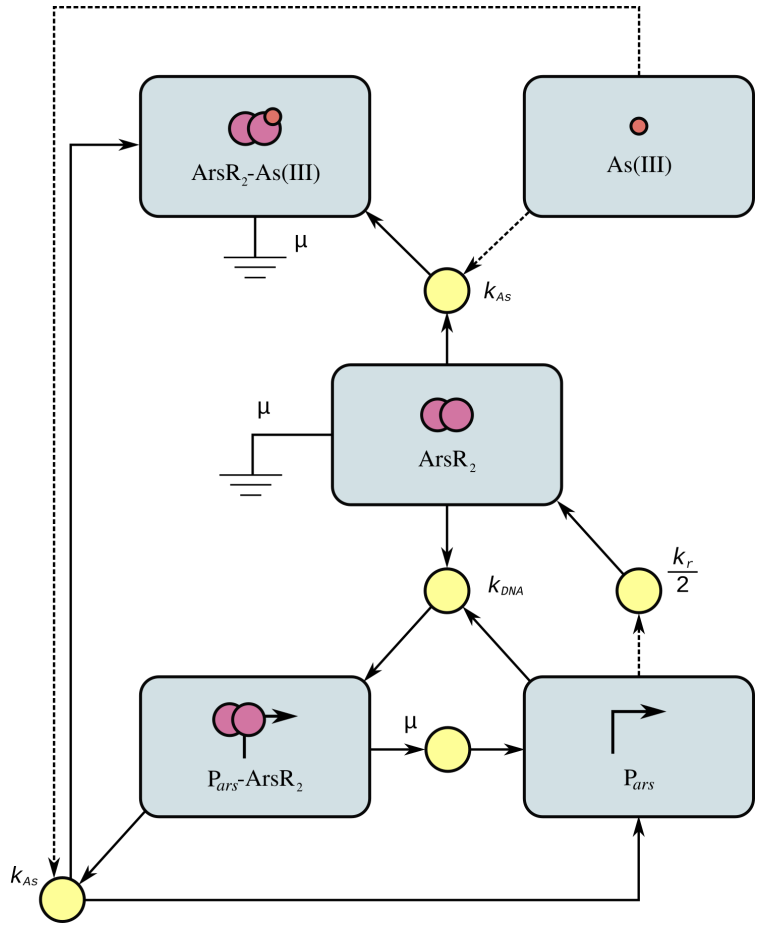
Nevertheless, we can learn several things from the parameter variations shown in Figure 3.11 beyond the basic shape of the curve. The leftmost figure confirms that increasing the strength of the promoter decreases the response both at zero and saturating arsenic levels, but has little effect on the shape of the response. The central plot shows that changes in $K^+_{DNA}$ have a similar effect, with tighter binding of the promoter reducing expression of the operon at all arsenite concentrations, under the assumption that $\alpha$ is large enough to effectively completely relieve DNA binding.

Increasing the rate at which arsenite binds with the dimer, $k^+_{As}$, relative to the rate at which $ArsR_2$ binds with DNA, $K^+_{DNA}$, has no effect on the leaky expression of the system, but significantly increases the response at high arsenite concentrations, as well as reducing the amount of arsenite required to attain close to the saturated response.

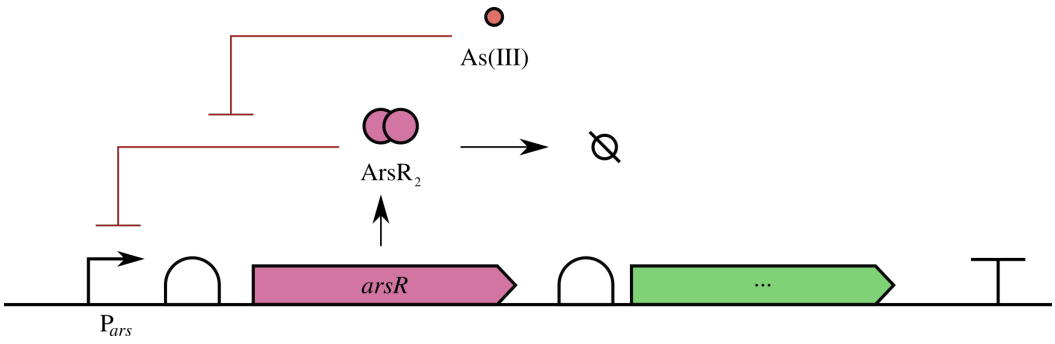
### 3.2.6 Arsenic and the Cell

Having modelled the response of the system to internal arsenite, we now turn our attention to the effect of the cell membrane. The entire purpose of the *ars* operon is to protect the cell from the harmful effects of arsenite by reducing the concentration in intracellular arsenic by pumping arsenite out of the cell. Thus, for a given external arsenite concentration, we expect the steady-state internal arsenite concentration as experienced by the operon to be significantly lower.

The model shown in Figure 3.12 models the uptake of arsenite as a Michaelis-Menten reaction with maximum velocity $k^{uptake}_{max}$ and concentration required for half-maximal velocity $K^{uptake}_M$. The catalyst for this reaction is assumed to be the GlpF transmembrane protein




(a) Visual representation of the mathematical model, now showing arsenite interacting with both $ArsR_2$ and $P_{pen}$-$ArsR_2$ to prevent binding of the promoter




(b) SBOL representation of the system

Fig. 3.10 The *ars* operon in the presence of arsenic. $As^{III}$ binds to the $\alpha$-binding domains of the homodimeric $ArsR_2$ at rate $K_{As}$, which induces a conformational change in $ArsR_2$ completely removing the dimer's ability to bind DNA. In this preliminary model, the concentration of unbound $As^{III}$ is fixed. Other species are as described in Figure 3.7

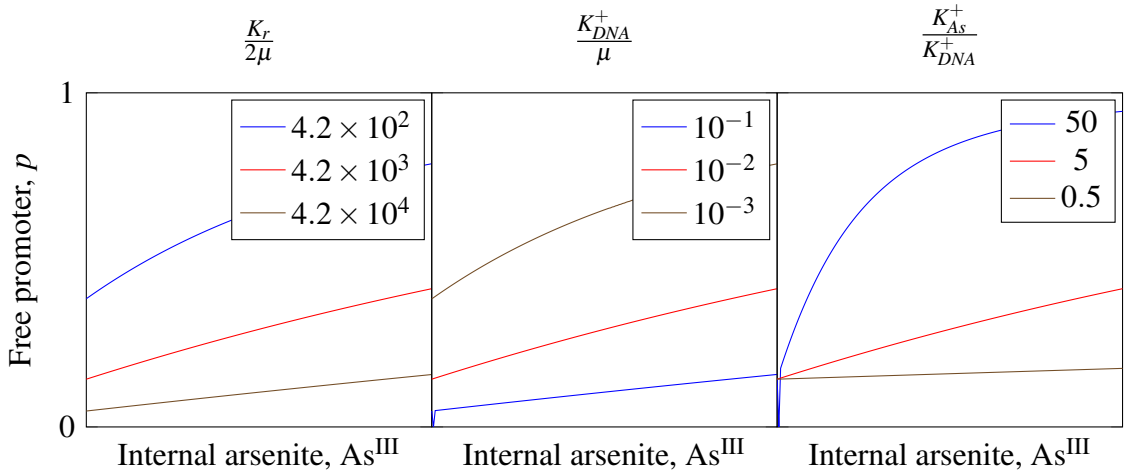



Fig. 3.11 Response to arsenite under variations of the dimensionless parameters shown above each plot for the model shown in Figure 3.10. The red curves are the same for each plot while the blue and brown curves show a ten-fold increase or decrease in the value respectively. In each case the denominator was held fixed while the numerator was varied.

as discussed in Section 3.1.1, the concentration of which is taken to be constant in these experiments. No reliable estimates of $k_{max}^{uptake}$ and $K_M^{uptake}$ are available in literature.

Extrusion of $As^{III}$ is catalysed by ArsB, as discussed in Section 3.1.1, which is also modelled as a Michaelis-Menten reaction with maximal velocity $k_{cat}^{extrude} \cdot [\text{ArsB}]$ and half maximal internal arsenic concentration of $k_{cat}^{extrude}$. In the absence of antimonite, $k_{cat}^{extrude}$ has been measured as $0.14mM$ by Kuroda et al. [84], while $k_{cat}^{extrude}$ has not been measured.

One further data point is available to us from literature. Singh et al. [150] found that the internal arsenite concentration caused by an external arsenite concentration of $10mM$ was $0.2mM$ in *E. coli*. This factor of 50 difference is quite remarkable given that $10mM$ is the WHO's threshold for safe arsenic levels, and thus represents the lower end of what the cells can tolerate.

We are left with a model with ten parameters, of which we have good direct empirical data for just one, $k_{cat}^{extrude}$. We can, however make a reasonable estimate of $\mu$, the dilution rate, and have derived ballpark estimates for the likely ranges of $K_r$, and $K^+_{DNA}$, which are compatible with the expected system behaviour and within the realms of biological plausibility. While there is no reason to believe that these values correspond closely to the real-world values, qualitative interpretations of the model behaviour should still be possible.

As expected, variations in $k_b$ and $k_{cat}^{extrude}$ have exactly the same effect on the model behaviour – doubling $k_b$ while halving $K_{cat}$ will double the steady state concentration of ArsB, but won't affect anything else in the model. Increasing (or reducing) the value of the product of the two increases (decreases) the rate at which arsenite is expelled from the

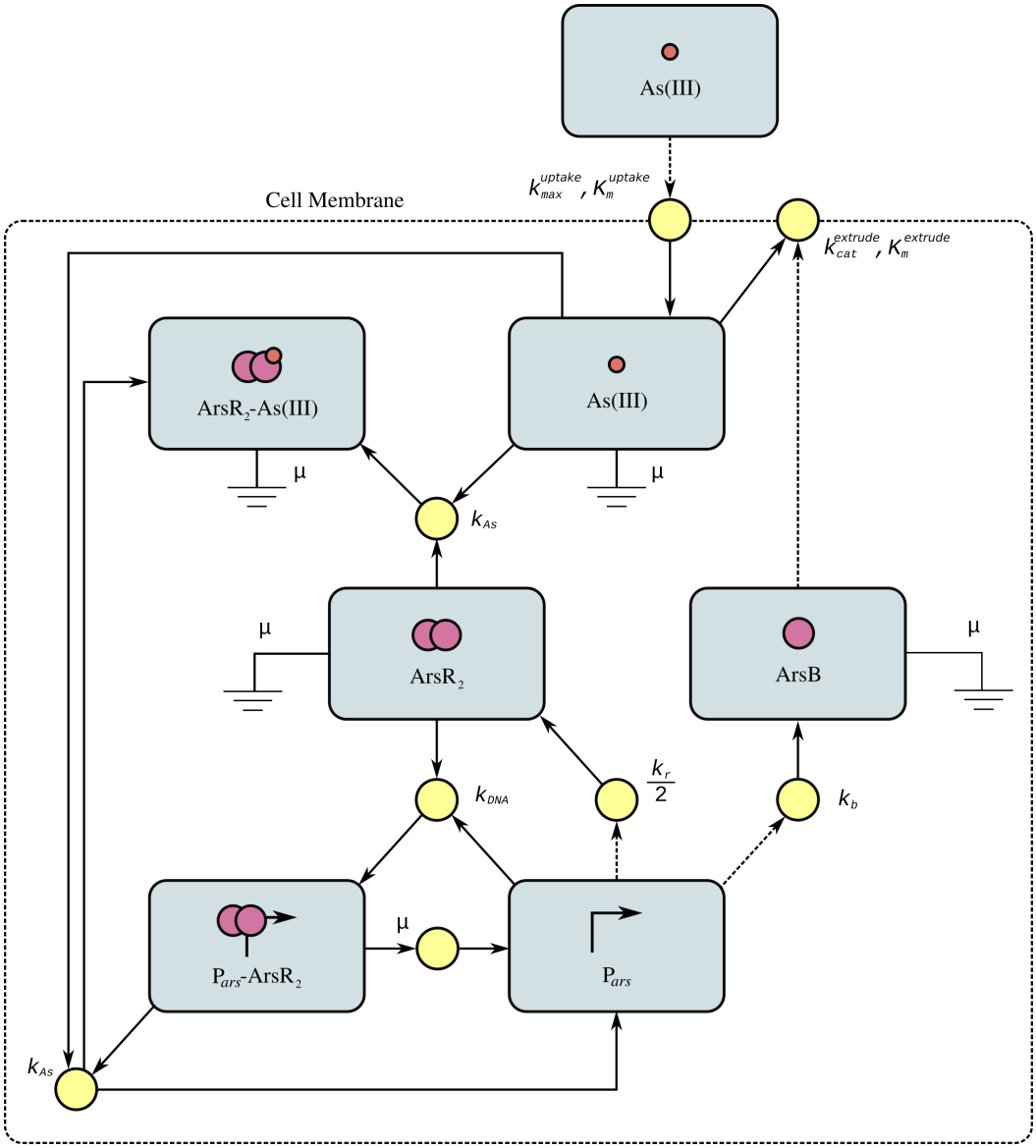



Fig. 3.12 The interactions of the *ars* operon, arsenic, and the cell membrane. Arsenite is taken up into the cell with Michaelis-Menten parameters $k_{max}^{uptake}$ and $K_M^{uptake}$ with the concentration of the membrane proteins through which arsenite enters assumed to be constant. In addition to ArsR, the promoter $P_{ars}$ produces ArsB which catalyses the extrusion of cellular arsenite with Michaelis-Menten parameters $k_{cat}^{extrude}$ and $K_M^{extrude}$. The extracellular space is assumed to be much larger than intracellular volume, and so the extracellular arsenite concentration is held constant. Other species are as described in Figure 3.10

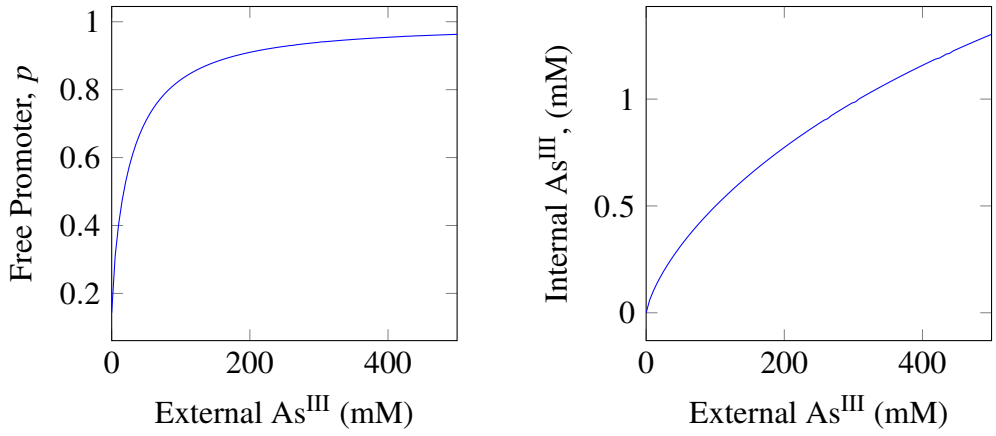



Fig. 3.13 Example response to arsenic for the cell model shown in Figure 3.12. Parameters were chosen for the model such that $10mM$ of external arsenite leads to $0.2mM$ of internal arsenite, as observed by Singh et al. [150].

cell, thus reducing (increasing) the induction of the operon for any given non-zero arsenite concentration. The inverse is true for $k_{max}^{uptake}$, where increasing (or reducing) $k_{max}^{uptake}$ increases (reduces) the amount of arsenite entering the cell, increasing (reducing) the induction of the operon for a given non-zero arsenite concentration.

This is as expected, however the value of $K_M^{uptake}$, the external arsenite concentration required for half-maximal uptake of arsenite, becomes highly important in determining the steepness of the induction curve of the operon. As noted in Section 3.2.5 above, $K_{As}^{+}$ has some effect on the amount of internal arsenite required to reach half-full induction of the system, but this full model shows that the value of $K_M^{uptake}$ has a much greater effect in determining this point than any other model parameter.

We can reduce the number of unknown parameters by considering meaningful combinations of them, such as the maximum steady state amount of $ArsR_2$, $\frac{K_r}{2\mu}$, as discussed previously. Another is $\frac{k_b \cdot k_{cat}^{extrude}}{\mu \cdot k_{max}^{uptake}}$, which corresponds to the ratio maximum rate at which arsenite can possibly be extruded from the cell when the promoter is fully unbound to the maximum rate at which arsenite can be taken into the cell. For an effective operon, this ratio must be grater than one, otherwise there exists a threshold external arsenic concentration above which internal arsenic will begin to increase dramatically, possibly becoming larger than the external arsenic concentration if $k_{max}^{uptake}$ is greater than $\mu$. It may seem reasonable to assume, then, that this ratio must be greater than one in the real system, however there are legitimate reasons why a ratio of less than one might lead to optimal performance, particularly if $K_M^{uptake}$ is large, requiring a large amount of external arsenite to cause rapid entry of arsenite into the cell. Since ArsB is toxic at high concentrations, and as a membrane protein likely to

have other negative effects, it is likely that optimal behaviour minimises the production of ArsB – indeed if this were not the case, ArsB would be produced constitutively, as no arsenite responsive control would be needed. Reducing this ratio below one might be a way to further reduce unnecessary expression of ArsB without compromising the operon for naturally occurring arsenite concentrations, at the expense of poor performance at exceptionally high arsenite concentrations.

## 3.3 Conclusions and a Strategy for *in vivo* Characterisation

This modelling effort has taught us several important things about the behaviour of the *ars* operon with regards induction by arsenite.

Firstly, Section 3.2.4 investigated the asymtotic behaviour of the system at extremes of zero and large concentrations of arsenite. It was found that increasing the strength of the promoter (for example, by constructing a hybrid promoter including the $ArsR_2$ binding domain, or by including an extra copy of the gene) would reduce the level of expression at both low and high arsenite concentrations. The model also predicted that, while holding all other parameters constant, there would be an optimal promoter strength, $K_r^*$, which maximises the difference between the upper limit of induction and the level of leaky induction – equivalent to maximising the difference between a high and low sensor response. Increasing the strength of the promoter beyond this point would begin to decrease the maximal induction level faster than the leaky induction level, reducing the performance of the sensor.

This section also found that for any given promoter, there was only a band of around 3 orders of magnitude for which values of $K_{DNA}^+$, the rate at which $ArsR_2$ binds to DNA, had a significant effect on the leaky expression, see Figure 3.8. From knowledge of system behaviour, it was therefore assumed that the true value of $K_{DNA}^+$ is somewhere above this transition in the absence of arsenite, but a conformational change causes this value to drop below this transition once arsenite has bound to the dimer.

Section 3.2.5 began to look at how induction levels varied as arsenite concentrations varied between these extrema. A new parameter, $K_{As}^+$, the rate at which arsenite binds to $ArsR_2$ plays a significant role in the level of induction caused by arsenite, as the level of induction depends on the proportion of $ArsR_2$ which has bound to arsenite. In this cell-free model, if $K_{As}^+ \cdot [As^{III}]$ is very small compared with the rate at which $ArsR_2$ is produced, $\frac{K_r}{2}$, very little induction is observed. Increasing $K_{As}^+$ therefore increases induction for all non-zero

concentrations of arsenite, and also reduces the concentration of arsenite required for half-full induction.

Finally, this model was placed in a cellular context in Section 3.2.6. Here it was shown that the rate at which arsenite is able to cross the cell membrane is crucial to operon behaviour, and that in particular the ratio of the maximal rate of arsenite extrusion, $k_{cat}^{extrude} \cdot \frac{k_b}{\mu}$, to the maximum rate at which arsenite can enter the cell, $k_{max}^{uptake}$. This ratio essentially controls the tolerable concentration of internal arsenite – increasing it reduces the internal concentration of arsenite. Modelling shows that if this ratio were to fall below one – for example by weakening the promoter or ribosome binding site associated with ArsB – the operon will fail to protect the cell from arsenite above a certain threshold concentration, above which point the internal arsenite concentration will grow quickly with the addition of more external arsenite. Furthermore, the value of $K_M^{uptake}$, the external arsenite concentration which causes half maximal rate of arsenite entry into the cell, is crucial to the shape of the operon's response, having a far greater effect than $K_{Aa}^{+}$ in the cellular context, assuming the latter is not very small.

This work has aided the understanding of the *ars* operon in *B. subtilis* – although the results could easily be generalised to other members of the *SmtB/ArsR* family – and clarified the roles of the various components within that system. The likely effect of parameter variations has also been studied, and in some cases, ranges of likely parameters have been arrived at. However, many parameters remain unknown, and considerably more robust empirical characterisation of the operon is required in order to validate this model.

# Chapter 4

# Establishing Ratiometric Characterisation in *Bacillus subtilis*

## 4.1 Introduction

### 4.1.1 Characterisation in Synthetic Biology

Robust part characterisation is central to the core mission of synthetic biology – to be able to reliably engineer novel and useful biological machines from synthetic parts. This goal cannot be achieved without a secure understanding of the function of each part, what its operating limits are, and how the part will interact with other biological parts within the synthetic context and the wider synthetic organism.

As synthetic biology becomes more established as an engineering discipline, evermore projects will be organised on the basis of a rationalised design-build-test cycle. Thorough, systematic part characterisation allows us to build libraries of synthetic parts with a range of biological properties. We can leverage this knowledge at design-time, when, given an abstract system design we can select from our library the various biological components – promoters, RBSs, genes and so forth – to give rise to the desired behaviour. We cannot guarantee that our design-time prediction will be accurate due to the very high interaction-space within any biological system, but with a large enough library of parts and the rise of automated methods for the compilation of concrete DNA sequences from abstract system designs, as well as the increase in lab automation and high-throughput DNA assembly, transformation, and testing techniques, we can test several such designs, and perform several iterations through the design-build-test cycle before arriving at a final, well characterised machine.

In this chapter, we focus on the characterisation of *expression systems*, i.e. the combination of promoter region and Ribosome Binding Site (RBS) which controls the expression

of a protein from a gene. Clearly, these factors are not the only ones which determine final protein concentration, or even the rate of production of a protein, as factors such as RNA turnover rate, RNA secondary structure, RNA-RNA and RNA-protein interactions, and gene codon bias all have an effect on protein production rate. In some cases – such as in the case of Fluorescent Reporter Proteins (FRPs) – post-translational modifications might be required for the production of mature protein, the rate of which may further modulate production of mature protein.

Each of these regulatory processes make for interesting potential control inputs for future synthetic systems, however synthetic biology has tended to focus mainly on implementing control at promoter and sometimes RBS level. Primarily, this is due to the relative simplicity of the physical interactions behind their behaviour. The (basal) strength of a promoter is determined by how tightly the relevant $\sigma$-factor binds to it, with tighter, faster binding causing the loading of more RNA polymerase per unit time and thus the production of more mRNA. In an inducible promoter, the $\sigma$-factor binding rates are modulated by other DNA binding transcription factors, which bind or unbind in response to the target modality. While these behaviours can become quite complex, they largely rely on simple association and dissociation reactions, which can be characterised and well modelled as seen in Chapter 3.

Interactions at the RBS can be more complex, as the rate at which ribosomes bind and successfully begin translating protein depends not only on how strongly the ribosome binds with the RBS sequence, but also on the secondary structure of the mRNA and the proximity of the RBS to the start codon of the gene. Since we can easily manipulate the sequence of the mRNA, we can to some extent control how tightly the ribosome will bind, as well as the distance between the RBS and start codon. However, reliably controlling mRNA structure is significantly more challenging. While algorithms exist for predicting mRNA structure, their accuracy is far from perfect, and those with reasonable algorithmic complexity specifically ignore more complex structures such as pseudo-knots. Even when armed with perfect RNA structure prediction, each mRNA is free to interact with every other RNA in the cell and binding of other RNAs can stabilise structures that weren't predicted or destabilise structures which were strongly predicted, leading to unreliable behaviour.

This problem becomes even more extreme when considering some of the other potential control mechanisms mentioned above, and so promoter and RBS level control remain most important in synthetic biology. We need, therefore, to develop characterisation methods which are aware of other causes of variation and either predict and account for them directly or find some other way of mitigating their effects.

### 4.1.2 Dual-Channel Characterisation

Biological measurements are inherently noisy, as cells do not behave in a deterministic way but are instead made up of a vast number of highly stochastic processes. In characterising a biological element such as a promoter, we would really like to dial down the effects of this wider stochastic context and access the promoter's underlying ability to promote RNA polymerase activity, independent of its specific cellular context. Dual channel or ratiometric reporter systems are a method of doing just this, by measuring the ratio between two expression systems rather than the absolute output, individual stochastic variation can be rejected while the true value is recorded.

Noise in expression systems can be classed as intrinsic or extrinsic[38, 132], each of which are generated by different mechanisms. Consider a population of genetically identical cells growing in identical media. We would expect some natural variation in the concentrations of cellular machinery within each cell – subtle differences in the concentrations of each $\sigma$-factor, RNA polymerase, and others. Some individuals within the population are likely to stochastically enter radically different phenotypes as a result of *bet-hedging*, introduced briefly in Section 3.2.2 in the context of the *B. subtilis* competence pathway. These sources of variation will necessarily have an effect on the apparent strength of a promoter being characterised, and this variation is known as extrinsic noise, as it is external to the system under test.

Intrinsic noise on the other hand refers to the variation one would expect if the population were not only genetically identical but if the copy number of every cellular component was exactly equal. In such a (hypothetical) scenario, the apparent strength of the promoter under test would also vary between individuals, as the processes by which genes are expressed are inherently randomised by the fact that cellular components essentially move around at random, bouncing off each other until a stable conformation with low free energy is reached such as a RNA polymerase binding with a $\sigma$-factor.

These two mechanisms by which noise is introduced require two different tactics to solve them. Intrinsic noise can be defeated simply through scale. Since each of the many stochastic molecular level interactions involved in expression are not heavily correlated, it is reasonable to expect the central limit theorem to apply to the observed rate of any given cellular process, and thus that the distribution of individual samples from this system are normally distributed. From the properties of the normal distribution, as the number of observations grows, the mean of those observations tends towards the true value. Often this is achieved using a bulk assay such as with a plate reader, where the fluorescent output of a large population of cells is assayed simultaneously and normalised using the optical density (OD).

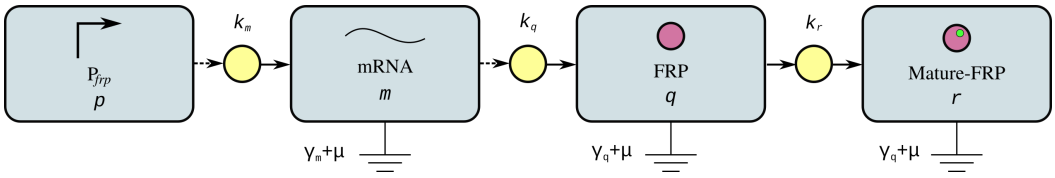



Fig. 4.1 Model of Fluorescent Reporter Protein (FRP) production adapted from de Jong et al. [32]. Species are shown in rectangles, circles represent reactions and a double headed arrow represents a reversible reaction. A dotted line indicates that a species is not consumed by the reaction. The electrical earth symbol represents a dilution/degradation reaction with no products. $k_m$, $k_q$, and $k_r$ represent the transcription, translation, and maturation rate respectively. $\gamma_m$ and $\gamma_q$ are respectively the mRNA and protein degradation rates, and $\mu$ is the growth dilution rate. Protein maturation is assumed to have no effect on the rate of degradation.

Extrinsic noise does not behave quite so well however. Consider an expression system with parameters $\boldsymbol{\phi}$ in a population containing two sub-populations, for example competent and non-competent from our earlier example. Let the average output from the first system equal $A(\boldsymbol{\phi})$ and the second equal $B(\boldsymbol{\phi})$, with population sizes of $N_a$ and $N_b$ respectively. The normalised observed value is

$$\frac{N_a \cdot A(\boldsymbol{\phi}) + N_b \cdot B(\boldsymbol{\phi})}{N_a + N_b}$$

which is highly dependent on the functions $A$ and $B$, as well as the proportion of cells in each, which is likely to be highly dependent on the particular conditions in which the population was grown, and thus not a very useful characterisation.

In ratiometric characterisation, the ratio of the system's output is taken as a ratio of the output of a reference system[69] with parameters $\boldsymbol{\phi_r}$. In the simple two-state example the output becomes

$$\frac{N_a \cdot \frac{A(\boldsymbol{\phi})}{A(\boldsymbol{\phi_r})} + N_b \cdot \frac{B(\boldsymbol{\phi})}{B(\boldsymbol{\phi_r})}}{N_a + N_b}$$

If the two functions $A$ and $B$ are linear or approximately linear in $\boldsymbol{\phi}$, then both ratios become some function, say $r$, of $\boldsymbol{\phi}$ and $\boldsymbol{\phi_r}$ such that the observed value simplifies to $r(\boldsymbol{\phi}, \boldsymbol{\phi_r})$ and no longer depends on the conditions $A$ and $B$ or the proportion of the population in each state. This value is therefore more robust to changes in these extrinsic parameters, and represents a more specific characterisation of the particular system under test, which should be more context independent than a single channel measurement.

### 4.1.3 A Mathematical Framework for Dual-Channel Characterisation

A straightforward ODE model of the production of a Fluorescent Reporter Protein (FRP) was established and verified by de Jong et al. [32], outlined in Figure 4.1. In this model, mRNA $m(t)$ is produced at rate $k_m p(t)$, where $k_m$ represents the rate at which mRNA is produced by the promoter under basal conditions, while $p(t)$ represents the level of induction ($p > 1$) or repression ($p < 1$) of the promoter. The mRNA catalyses the production of inactive protein with concentration $q(t)$ at rate $k_q$, which then matures spontaneously at rate $k_r$ to produce mature, observable protein $r(t)$.

Crucially, this model assumes that all steps in producing mature protein are linear – doubling $p(t)$ doubles the rate of production of mRNA, doubling the concentration of immature protein doubles the rate at which mature protein is produced, and so on. de Jong et al. [32] successfully validated these assumptions from empirical data, but it is important to understand the limitations that these assumptions create, as no real physical system is truly linear in all conditions. In the case of transcription and translation, a very strong promoter or RBS might begin to saturate the cell's capacity for expression such that a proportional increase in promoter or RBS strength causes a sub-proportional increase in transcription or translation. This effect is likely to be more pronounced under stress conditions such as nutrient starvation when fewer metabolic resources are available.

The protein maturation step does not directly place a burden on the cell's expression machinery, but rather FRP maturation is known to require oxygen[4, 32, 145]. Thus, if a very large amount of immature protein is produced, or if cells are grown in sufficiently anoxic conditions, the lack of oxygen is likely to cause a non-linearity in this final step, and the real system output will deviate from model behaviour.

Making these assumptions, we can write down the governing differential equations for $m$, $q$, and $r$ as

$$\begin{aligned} \dot{m} &= k_m p - (\gamma_m + \mu)m \\ \dot{q} &= k_q m - (\gamma_q + \mu + k_r)q \\ \dot{r} &= k_r q - (\gamma_q + \mu)r \end{aligned} \tag{4.1}$$

In the case where the promoter is constitutive (i.e. $p(t) = 1$), we can find the steady state of $r(t)$ this system of equations by setting $\dot{m} = \dot{q} = \dot{r} = 0$ and solving for $r$, which gives

$$\lim_{t\to\infty} r(t) = \frac{k_r k_q k_m}{(\gamma_m + \mu)(\gamma_q + \mu + k_r)(\gamma_q + \mu)} \tag{4.2}$$

which is clearly linear in $k_m$ and $k_q$. This linearity justifies the assumption of linearity made at the end of Section 4.1.2, that the functions $A$ and $B$ are linear functions of the intrinsic properties of the expression system (previously denoted as $\boldsymbol{\phi}$), in this case the promoter strength $k_m$ and translation rate $k_q$. This result proves the efficacy of the dual channel method in reducing the effect of extrinsic variation of the system under test.

The system of governing equations (4.1) can also be solved in the case of a spike of promoter activity – also known as the impulse response of the system – by setting $p(t) = \delta(t)$ where $\delta$ is the Dirac delta function. The impulse response is a convenient property of a linear system, as it enables us to calculate the response of the system for any input by simply convolving the input with the impulse response. We can therefore think of the impulse response as a filter which is applied to the parameter we wish to measure. Using Laplace transforms, the impulse response can be found as

$$r(t) = \begin{cases} 0 & t < 0 \\ k_q k_m \left( \frac{e^{-(\gamma_q+\mu)t}}{(\gamma_m-\gamma_q)} - \frac{e^{-(\gamma_q+\mu+k_r)t}}{(\gamma_m-\gamma_q-k_r)} + \frac{k_r e^{-(\gamma_m+\mu)t}}{(\gamma_m-\gamma_q)(\gamma_m-\gamma_q-k_r)} \right) & \text{otherwise} \end{cases} \tag{4.3}$$

and is shown plotted using values derived for GFP from de Jong et al. [32] in Figure 4.2. As expected for a physical system which obeys causality, the instantaneous response at $t = 0$ is zero, but this quickly grows to a maximum after around twenty minutes, before decaying away exponentially to zero. Figure 4.2 shows that as $t$ becomes large, the first term comes to dominate, and so the limiting decay rate has a half-life of $\gamma_q + \mu$ due to degradation and dilution of the protein. A FRP system therefore 'smears out' the fine detail of the promoter activity $f(t)$, with a slower degradation and dilution rate causing greater smearing of the output reading.

It is therefore advisable that in cases where good resolution in the time domain is important, such as the characterisation of the response to a step change in inducer molecule, the dilution and degradation rate ($\gamma_q + \mu$) should be large, which can be achieved either by making sure cells are growing quickly during the assay or by using a FRP with low stability.

### 4.1.4 Dual Channel Reporters in *B. subtilis*

Dual channel reporters similar to this have been commonly used in *E. coli* (e.g. Yordanov et al. [177]), but not so in *B. subtilis*. Partly this is due to the fact that recombinant genes are commonly introduced via a plasmid vector in *E. coli* as this is simpler and more efficient than chromosomal integration. However, measurements of the activity of plasmid-borne expression systems must contend with an extra source of variation, the plasmid copy number, which can vary quite considerably from cell-to-cell.

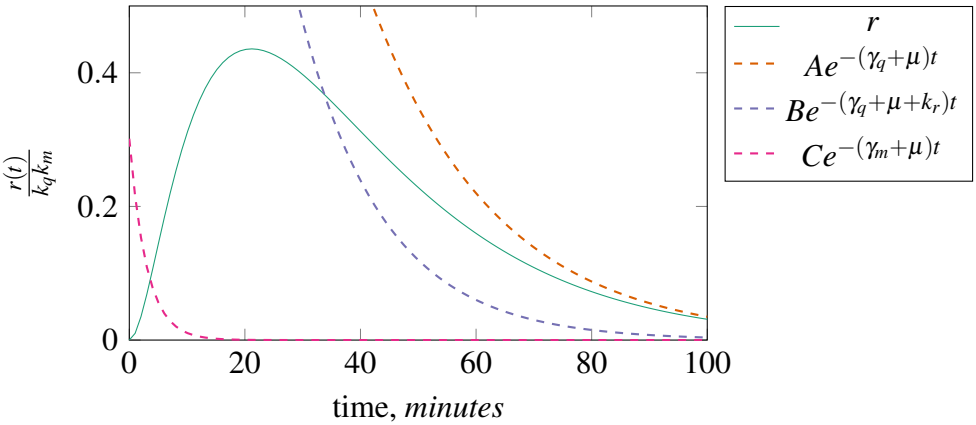



Fig. 4.2 Impulse response of the fluorescent reporter model described in Equation 4.1, showing the contribution of each term. Values for $k_r$, $\gamma_q$, and $\gamma_m$ were 0.023, 0.012, and 0.3min$^{-1}$ respectively taken from de Jong et al. [32] and $\mu = 0.034\text{min}^{-1}$. See Equation 4.3.

Unlike *E. coli*, Plasmid vectors are typically not stable in *B. subtilis*, and so recombinant DNA must be inserted into the chromosome in order to be stable. Fortunately, as discussed in Section 1.3, *B. subtilis 168* is a naturally highly competent strain and can easily be made chemically competent by which it will insert DNA into the chromosome by homologous recombination. Since the normalisation of plasmid copy number has often been seen as the key benefit to dual channel reporter systems in *E. coli*, such reporter systems have been underused in hosts where chromosomal integration is more common, as genome copy number is considerably more tightly controlled.

The need for chromosomal insertion brings with it its own design challenges, however, and a dual channel reporter system has not previously been established in *B. subtilis*. The remainder of this chapter deals with the design, construction, and testing of such a reporter system.

## 4.2 Design

### 4.2.1 Requirements

The most basic requirement of the dual reporter system is that the chosen fluorescent reporters have good properties for a ratiometric application. The chosen reporters must be easily expressed by the host and have good brightness such that they can be easily detected by whichever sensing technology is applied and must be easily distinguishable from wild-type auto-fluorescence even at low expression rates. The two reporters must also be orthogonal,

meaning that overlap in the absorption or emission spectra between the two reporters must be minimal to prevent cross talk between the two signals.

For a characterisation platform to be useful, it must also be possible to insert different test systems quickly and reliably, ideally in a way suitable for automation for high-throughput methods. Such a system should also remain as flexible as possible, placing only the minimum of requirements on the sequences which can be inserted reliably. In particular, such a system must be able to deal with very short inserts well, as 5' UnTranslated Regions (UTRs) are often fewer than 10bp in length.

*B. subtilis* has ten native sigma factors, reviewed in Haldenwang [57] and enumerated in Table 4.1. Each sigma factor controls the expression of a large number of genes responsible for a particular cellular function. For example $\sigma^A$ is responsible for the expression of most genes under normal growth conditions, while $\sigma^B$ is expressed in response to a wide range of stress conditions and in turn induces the expression of a wide range of stress response proteins which help shield the cell from a range of potential hazards, known collectively as the general stress response. Locke et al. [92] found that the level of stress is not encoded by level of $\sigma^B$ expression, but that actually expression of $\sigma^A$ and $\sigma^B$ is exclusive with one or other being fully expressed at all times. Expression of $\sigma^A$ and $\sigma^B$ is therefore pulsatile with the level of stress encoded by the number and duration of pulses of $\sigma^B$ expression.

Correct ratiometric characterisation of an expression system requires that both the query and reference promoter interact with the same sigma factor, as otherwise the output would be dependent on the proportion of each sigma factor present, and thus on cell state. This could be useful when characterising the behaviour of sigma factors under different conditions, but this falls outside the scope of this work. It is therefore a design requirement that the reference expression system should be reasonably easy to change, although it need not be as simple as changing the query system which is expected to change more often.

### 4.2.2 Choice of Assembly Technology

Two categories of assembly technology were considered for the characterisation platform, the PCR/exonuclease based Gibson Assembly and the Type-IIS restriction enzyme based Golden Gate. Both were used at various stages of the project, and simple protocols and references are introduced in Section 2.2. Each technology has its own advantages and limitations for use in this context.

Gibson assembly is a scarless assembly technique which relies on each fragment having a region of overlap between them. So long as successive fragments agree, the sequence of this overlap is free to vary so long as the melting temperature of the overlap remains within an acceptable range (although sequences with long repeats or strong secondary structure should

| | Factor | Description |
|---|---|---|
| Vegetative | $\sigma^A$ | Housekeeping/early sporulation |
| | $\sigma^B$ | General stress response |
| | $\sigma^C$ | Postexponential gene expression |
| | $\sigma^D$ | Chemotaxis/autolysin/flagellar gene expression |
| | $\sigma^H$ | Postexponential gene expression; competence and early sporulation genes |
| | $\sigma^L$ | Degradative enzyme gene expression |
| Sporulation | $\sigma^E$ | Early mother cell gene expression |
| | $\sigma^F$ | Early forespore gene expression |
| | $\sigma^G$ | Late forespore gene expression |
| | $\sigma^K$ | Late mother cell gene expression |

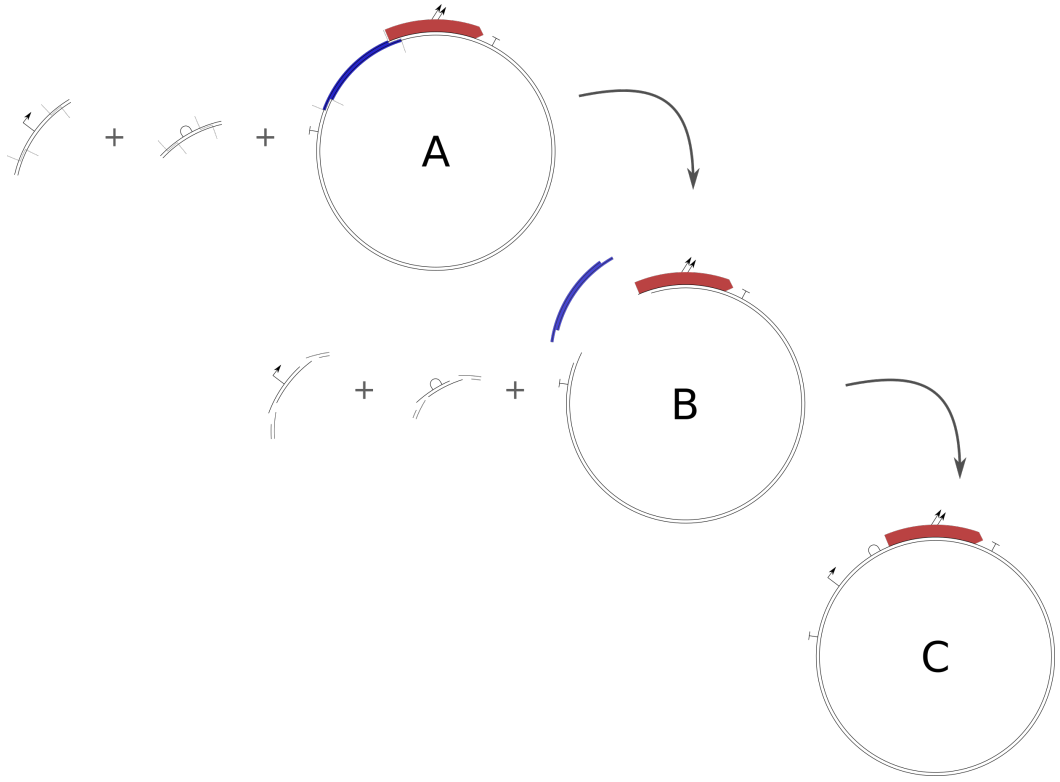
Table 4.1 The sigma factors of *Bacillus subtilis*. See Haldenwang [57] for a review.

be avoided). However, Gibson assembly is considerably less reliable with short sequences, as the assembly relies on the action of a 5'-exonuclease to expose the overlapping regions as single stranded sticky ends. Control of exactly how much DNA the exonuclease degrades is limited, as this reaction happens quickly as the reaction mixture is heated directly from 4 °C to 50 °C, at which temperature the enzyme will have a short half life. If the fragment is too short, it is entirely destroyed before this denaturation temperature can be reached, and so Gibson assembly is limited to fragment sizes of the order of 100 bases or more.

Golden Gate, on the other hand has far more permissive limitations on fragment size, with $\sim$ 5bp fragments having been successfully integrated. However, as a restriction enzyme based technology Golden Gate does place limits on the specific sequences being used, with sequences that include the chosen enzyme's recognition site being at best significantly less efficient to insert, and sometimes impossible. Furthermore, Golden Gate is not truly scarless as a 3-4bp overlap is required between fragments. Unlike traditional restriction enzymes, the overlap sequence is not specific to the enzyme used but can be any non-palindromic, non-repetitive sequence. Overlap regions in successive fragments must be equal, however, meaning that Golden Gate based technologies are often referred to as *quasi-scarless*, since the sequence of the scar can be specified within well defined limitations. We can turn this apparent limitation to our advantage, however, by standardising the overlaps used in our test expression systems. By specifying that promoters and RBSs within the testing library must be flanked by specific, well chosen overlaps, we can easily generate a library of promoter and RBSs which can be used interchangeable with the characterisation platform.

In order to test this, the type-IIS restriction enzyme BsmBI was selected, as none of the candidate FRPs contained the BsmBI cut site and the enzyme would not conflict with


(a) The landing pad sequence. Two outward-facing BsmBI cut-sites allow the landing pad region to be excised and replaced with any sequence with compatible overhangs. The right-hand overhang in the figure, *ATGG*, includes the start codon of the reporter protein plus one base pair, which means that the 5' UTR can be fully specified. The final '*G*' potentially places a limitation on chosen reporter protein, see Section 4.2.3




(b) Assembly schematic showing the use of the landing pad to insert an arbitrary promoter and RBS combination into a reporter plasmid. The assembly contains three fragments, a promoter, an RBS and a plasmid template, shown in step A, each with appropriate BsmBI cut sites and overhangs listed in table 4.2. Digestion then removes the recognition sites and exposes the matching sticky ends in step B, before the fragments are ligated together shown in C.

Fig. 4.3 Design and use of the landing pad assembly system, showing the sequence of the landing pad region and the process by which a promoter and RBS can be inserted into a template plasmid.

enzymes being used elsewhere in the assembly process. A short 'landing pad' region which would accept a promoter and RBS was designed and inserted immediately upstream of both the reference and query FRP, as shown in Figure 4.3. The landing pad design was tested by inserting a strong promoter and RBS which were previously known to be highly functional in *E. coli* and *B. subtilis* to make colony selection as simple as possible. The result of the Golden Gate assembly was transformed into chemically competent *E. coli* and plated on chloramphenicol plates. No false positives were observed, although the number of transformants was quite low, with only a handful of colonies present after plating all cells.

A control experiment demonstrated that the competent cells were not at fault, and so suspicion fell on the efficiency of the assembly reaction. The most likely cause of low efficiency was suspected to be the restriction enzyme – most type-IIS enzymes have an operating temperature of 37 °C, however BsmBI cuts most efficiently at 55 °C. Golden Gate assembly is typically carried out by cycling between two temperatures, a digestion temperature during which the fragments are cleaved (typically 37 °C) and a restriction temperature during which matching fragments are ligated, usually 16 °C. It was hypothesised that increasing the digestion temperature from the 37 °C specified by standard Golden Gate protocols while keeping the ligation temperature constant would improve the efficiency of digestion and therefore increase the efficiency of the overall reaction. The test assembly was a four-part assembly, including backbone, promoter $-35$ region, promoter $-10$ region, and RBS. The two promoter regions and RBS were roughly 50 bp each, with the total promoter and RBS insert measuring 101 bp in length. These lengths can be resolved on a 4% agarose gel, as shown in Figure 4.4, which shows the result of assembly at 37, 45, and 55 °C. Image analysis shows that the most efficient assembly occurs at the medium temperature of 45 °C, with no visible assembly taking place at the higher temperature of 55 °C, although more restriction has taken place at this temperature than at 37 °C. The most likely explanation is that exposure to temperatures as high as 55 °C causes significant denaturation of the T4 ligase which is required to assemble the fragments, whose published denaturation temperature is 65 °C. All further assemblies involving BsmBI were therefore carried out with a 45 °C digestion step, which tended to produce around two orders of magnitude more colonies than at 37 °C.

The specification of the Type IIS overhangs used throughout the characterisation effort is shown in Figure 4.2.

### 4.2.3 Fluorescent Reporter Selection

Two fluorescent proteins are required for the dual reporter system, one to report the output of the query system and one to report the output of the reference system. The chosen reporters

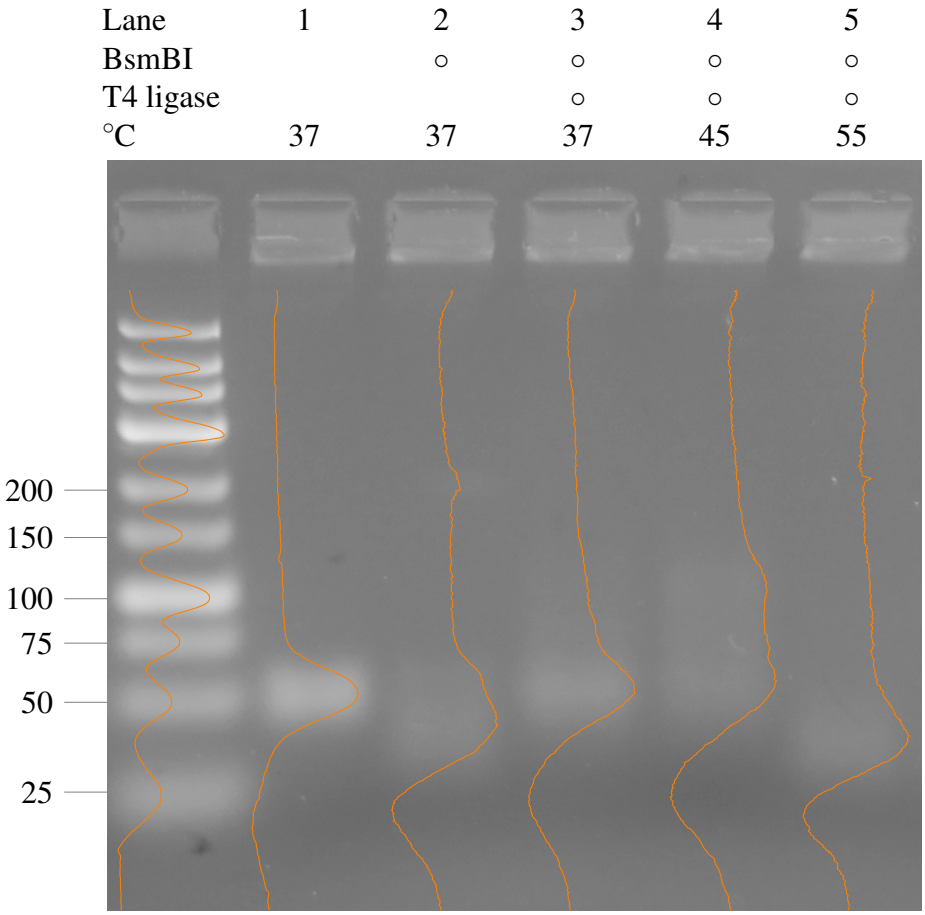



Fig. 4.4 Optimisation of BsmBI Golden Gate, showing most efficient assembly at the intermediate value of 45 °C. Orange overlays show average pixel intensity across each lane of the 4% agarose gel. Three fragments were assembled, with lengths of 58, 45, and 51 bp, seen combined without enzyme in lane one. Each fragment contains two inward facing BsmBI cut-sites; the effect of cleavage at these sites is shown in lane two where no ligase was present. Lanes 3-5 show the result of a full Golden Gate reaction after 50 cycles with an annealing temperature of 16 °C and digestion temperature of 37, 45 and 55 °C, respectively. Full assembly results in a fragment size of 101 bp, as can be seen most clearly in lane 4.

| Part Type | 5' Overhang | 3' Overhang |
|---|---|---|
| Promoter (-35 region) | CGTA | TAAC |
| Promoter (-10 region) | TAAC | AGAC |
| Promoter (whole) | CGTA | AGAC |
| 5' UTR | AGAC | ATGG |

Table 4.2 Overhang specifications for each part type in the Dual Reporter library. Promoters were originally split into two parts to allow the -35 and -10 regions to be varied independently, and also to allow promoters to be synthesised by short oligo synthesis, however this design was not used in the final characterisation effort except in the case of the reference promoter, $P_{pen}$. Parts are required to end with the sequence 5'-NNNCGTCTCN-3' at each end in order for restriction by BsmBI to reveal the expected sticky ends. The values of the Ns was chosen to minimise predicted secondary structure around the end of the sequence.

| Reporter | $\lambda_{max}^{abs}$ $(nm)$ | $\lambda_{max}^{em}$ $(nm)$ | $\varepsilon^a \times 10^3$ $(M^{-1} \cdot cm^{-1})$ | QY$^b$ | $\varepsilon \cdot$ QY$^c$ | $t_{\frac{1}{2}}^{mat}$ $^d$ (h) | Reference |
|---|---|---|---|---|---|---|---|
| mVenus | 515 | 528 | 92.2 | 0.57 | 52.6 | $\sim 0.28$[25] | [107] |
| mTurquoise2 | 434 | 474 | 30 | 0.93 | 27.9 | – | [51, 52] |
| eForRed | 589 | 609 | 111.3 | 0.16 | 17.8 | – | [4] |
| mRuby | 558 | 605 | 112 | 0.35 | 39.2 | 2.8 | [82] |
| mKate2 | 588 | 633 | 62.5 | 0.4 | 25.0 | 0.33 | [146] |
| mCherry | 587 | 610 | 72 | 0.22 | 15.8 | 0.25 | [145] |
| mScarlet | 569 | 593 | 100 | 0.7 | 70 | $\sim 2.9$ | [10] |
| mScarlet-I | 569 | 593 | 104 | 0.54 | 56.2 | $\sim 0.6$ | [10] |

Table 4.3 Summary of FRPs considered for the dual reporter. (a) extinction coefficient (b) quantum yield (c) theoretical brightness (d) maturation half-life

must be bright enough to be easily detected over the relevant range of system outputs, and should be easy to isolate from background autofluorescence. It must also be possible to distinguish the two reporters from each other, meaning their absorption and emission wavelengths must be sufficiently far apart for the output of the query system to be rejected by the filters recording the reference system and vice versa. Although steady state output of the reporter system is linear in maturation rate (see Equation 4.2), and so differences in maturation rate can be corrected by normalisation, it is wise to choose reporters which mature at approximately the same rate such that the effect of any non-linearities which do occur is minimised. In addition, choosing a reporter with a faster maturation rate will lead to a brighter strain as more of the expressed reporter will be fluorescent at any one time which will in turn reduce measurement error.

Initial rounds of experiments were conducted using mVenus as a query reporter and mTurquoise2 as the reference. mVenus is a bright, fast maturing and stable Yellow Fluorescent Protein (YFP) variant[107], which had been previously used successfully in both *E. coli* and *B. subtilis*. mVenus has a relatively high extinction coefficient of $92.2 \times 10^3\ \mathrm{M}^{-1}\,\mathrm{cm}^{-1}$ and a quantum yield of 57%, meaning it absorbs a large proportion of photons with appropriate energy and emits 57 photons for every 100 absorbed. Fittingly, these properties place mVenus among the brightest FRPs, as its namesake Venus is among the brightest object in the night sky.

The Cyan Fluorescence Protein (CFP) variant mTurquoise2 has an even higher quantum yield than mVenus, but its extinction coefficient is only a third of mVenus' and so its overall theoretical brightness is equal to around half that of mVenus. mTurquoise2 was initially chosen as a reference reporter as its absorption/emission spectra do not overlap with mVenus, and it had previously been used in bacteria. The lower brightness was not too problematic in itself, as the reference system does not vary as much as the query system, and so by fixing the reference system as a relatively strong promoter and RBS, the output is easily monitored.

After some effort described in Section 4.2.4, a test strain was constructed consisting of mVenus and mTurquoise2 both under identical strong promoters and RBSs. This strain was assessed by a plate reader growth assay, and while the mVenus signal was easily detectable, no signal could be detected from mTurquoise2. After some experimentation, it was discovered that replacing the standard Lysogeny broth (LB) growth medium for a minimal growth medium such as M9 led to some detected mTurquoise2 activity. The most likely explanation is that some component of the LB medium displays strong autofluorescence in the cyan range, making detection of the relatively weak signal due to mTurquoise2 difficult.

Since requiring the system to be assayed in only minimal media seems somewhat limiting, it seemed more sensible to search for a reference reporter protein in a different area of

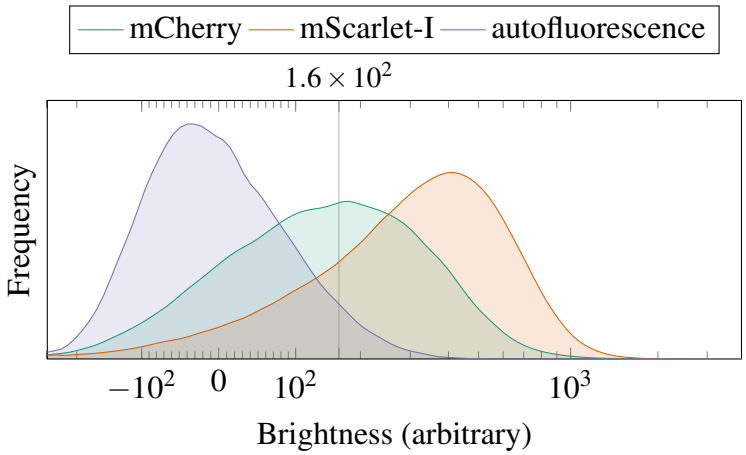



Fig. 4.5 Comparison of mCherry and mScarlet-I fluorescence against wild-type autofluorescence. The vertical line shows the 95$^{\text{th}}$ percentile of autofluorescence samples, with 45% and 77% of mCherry and mScarlet-I samples being above this threshold (and thus confidently not caused by autofluorescence), respectively. Cells were grown overnight in LB at 37 °C with shaking and used to inoculate preheated fresh LB medium. The fresh inoculates were grown for 6 hours, before being assayed with flow cytometry. 100,000 events were captured for each sample and no gating was applied to any channel.

the spectrum, and so a range of red fluorescent proteins were investigated, see Table 4.3. Of those shown in the table, three were investigated, namely eForRed, mCherry, and the recently reported mScarlet-I, while mRuby and mScarlet were discarded due to their long maturation times. In order to compare the suitability of these reporters, it is useful to define some threshold brightness below which samples are considered undetectable. Ratiometric measurements whose reference brightness is below this threshold are then discarded.

The 95$^{\text{th}}$ percentile of autofluorescence measurements is a simple choice of threshold value which has a simple statistical interpretation – samples which pass this threshold are at least 95% confident of belonging to a distribution with mean brightness larger than autofluorescence. Figure 4.5 shows the distribution of 100,000 far-red channel events captured using flow cytometry for mCherry and mScarlet-I reporters as well as the autofluorescence measurements. mScarlet-I is clearly the better reporter, with 77% of recorded events appearing above the 95% confidence interval threshold, while only 45% of mCherry events passed this threshold.

The *comGA* leader sequence is a short sequence taken from the first 8 aa of the *B. subtilis comGA* gene, equal to `MDSIEKVS`, with DNA sequence 5'-`ATGGATTCAATAGAAAAGGTAAGC`. When placed at the start of the FRP, the sequence causes an increase in brightness of the resulting reporter, as shown in Figures 4.6 and 4.7. The exact mechanism behind this increase in brightness is not known, however since ComGA is a membrane targeted protein[14, 15],

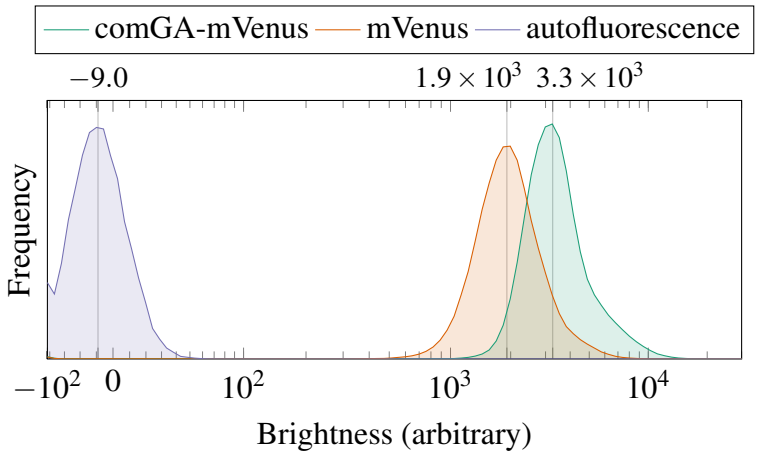



Fig. 4.6 Brightness of query FRP mVenus with and without *comGA* leader sequence. Vertical lines show the median brightness, showing a modest increase of around 1.7 fold for mVenus.

one possibility is that the leader sequence causes the expressed reporter protein to also be membrane targeted. If a significant proportion of the tagged reporter is embedded in the membrane in this way, one would expect it to be less subject to targeting by the cell's degradation machinery than when freely moving in the cytoplasm. In this way, the comGA tag may reduce the rate of protein turnover, increasing the amount of reporter protein present in the cell.

Figure 4.6 shows the effect of this leader sequence on the observed brightness of mVenus, which is already a very bright reporter. Here the increase is only a moderate 1.7 fold, however the effect on the weaker mScarlet-I reporter, shown in Figure 4.7, is only slightly larger at 2.9 fold. In the case of mScarlet-I, this increases the strength such that 99% of observed events are above the 95% percentile of autofluorescent events.

Since the mechanism by which comGA increases the observed brightness of the reporters isn't known, it's possible that the effect is influenced by any number of mechanisms outside our control. If the comGA sequence were present on one reporter but not the other, changes in these hypothetical factors would affect the effect of comGA on that reporter's output, changing the ratiometric output. However, if comGA is present on both reporters then fluctuations in the level of comGA-based induction should affect each reporter approximately equally and thus have little effect on the ratiometric output.

Given the obvious advantage of comGA in increasing the observed brightness of the reference reporter, it was decided to include the *comGA* leader sequence upstream of both query and reference reporter.

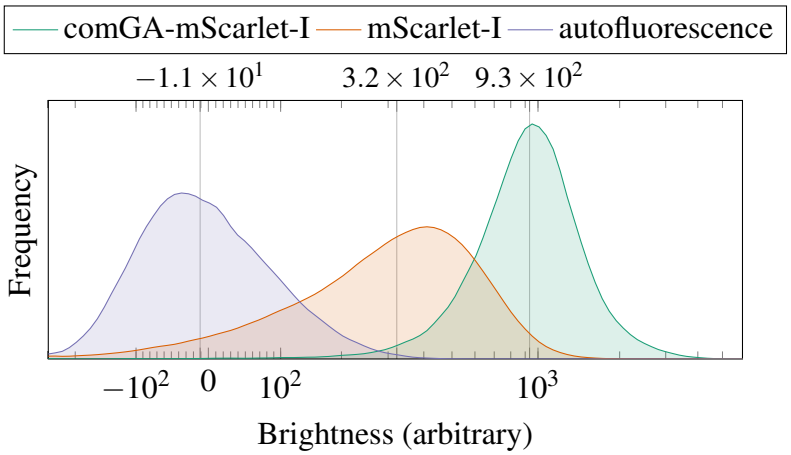



Fig. 4.7 Brightness of reference FRP mScarlet-I with and without *comGA* leader sequence. Vertical lines shown the median brightness, with a modest increase of around 2.9 fold for mScarlet-I. However, the proportion of samples above the 95$^{\text{th}}$ percentile of the autofluorescence samples increased from 77% without comGA to 99% with comGA, increasing the number of useable samples.

### 4.2.4 Architecture

Two construction architectures were considered for the dual reporter system, a parallel and a serial one. In the parallel, one template plasmid is constructed which contains the full reporter system (including promoter and RBS), the query reporter downstream of the landing pad region (introduced in Section 4.2.2), an antibiotic resistance marker, and appropriate homology regions for insertion into the commonly used AmyE region of the *B. subtilis* chromosome. The query promoter/RBS combination are then inserted using the BsmBI based reaction optimised in Section 4.2.2, and the entire region is transformed into competent *B. subtilis* 168.

Several variants of this plasmid were attempted, including versions with divergent and convergent reporters – though always insulated by terminators – and various choices for locating the antibiotic resistance. Both Gibson assembly and Golden Gate assembly methods were also attempted, but no assembly strategy yielded reliable results. In the case of Gibson assembly, all screened plasmids contained missing sections due to incorrect annealing of the overlapping ends. Golden Gate also yielded very low assembly efficiency and often gave no correct colonies at all.

Returning to the drawing board, the idea of a serial assembly architecture was explored, described in figures 4.8 and 4.9. In this design, two template plasmids are built, one containing the reference reporter and the other containing the query reporter. In the templates, the landing pad system described in figure 4.3 is upstream of each reporter, meaning that

the reference can be changed more easily in this system, however the homology regions of the query template plasmid are modified such that the chromosomal insert happens just upstream of the AmyE locus (dubbed AmyE'). The antibiotic resistance markers were also changed to chloramphenicol and spectinomycin resistance for the reference and query systems respectively.

Once the landing pad has been replaced by the reference system, see figure 4.3, the reference plasmid is transformed into competent *B. subtilis* 168, to create a strain referred to as *B. subtilis* R$\langle X\rangle \cdot \langle Y\rangle$, where $\langle X\rangle$ and $\langle Y\rangle$ identify the promoter and RBS present respectively. The chosen query system is then inserted into the query template plasmid, and this plasmid is transformed into the reference strain, to create a strain called *B. subtilis* R$\langle X\rangle \cdot \langle Y\rangle$-Q$\langle U\rangle \cdot \langle V\rangle$ where $\langle U\rangle$ and $\langle V\rangle$ identify the promoter and RBS used in the query system. Since the reference system was kept constant during this work, $\langle X\rangle$ and $\langle Y\rangle$ were typically omitted from strain descriptions.

However, this 'serial' insertion re-design has two downsides as compared with the initial 'parallel' design. The first is one of turnaround time – two transformations are required, with the second dependent on the first meaning that building a full reporter system from scratch takes longer with the second system. However, since the stages of inserting the reference and query system are independent, a large batch of the reference strain can be made ahead of time and so only the actual query system need be inserted to generate the final strain. A new reference strain will still be needed in some cases – for example when characterising a promoter which interacts with a sigma factor for which no reference strain has been generated – but in most cases using the same reference system is rather the point of dual channel characterisation.

The second disadvantage is that the second transformation introduces a second antibiotic resistance into the organism. While this approach works, each extra gene that is added has the potential to change the way the cell behaves, at the very least expression of the antibiotic resistance genes uses cell resources, increasing the chance of some form of starvation response being triggered (see the discussion about pulsatile expression of $\sigma^{\mathrm{B}}$ in Section 4.2.1). At worse, the antibiotic resistance gene might directly affect cell behaviour in some way – for example in changing the way a sugar is metabolised – which might, if we are unlucky, directly affect the system we're trying to interrogate.

A convenient mechanism for removing antibiotic markers in *B. subtilis* is Xer recombination, as described by Bloor and Cranenburgh [11]. In this system, two 28bp 'dif' sites are placed either side of the region to be removed. Once selective pressure is removed (by growing the cells on normal media), the two dif sites are resolved due to the action of the native *ripX* and *codV* genes. The chloramphenicol antibiotic resistance marker employed in

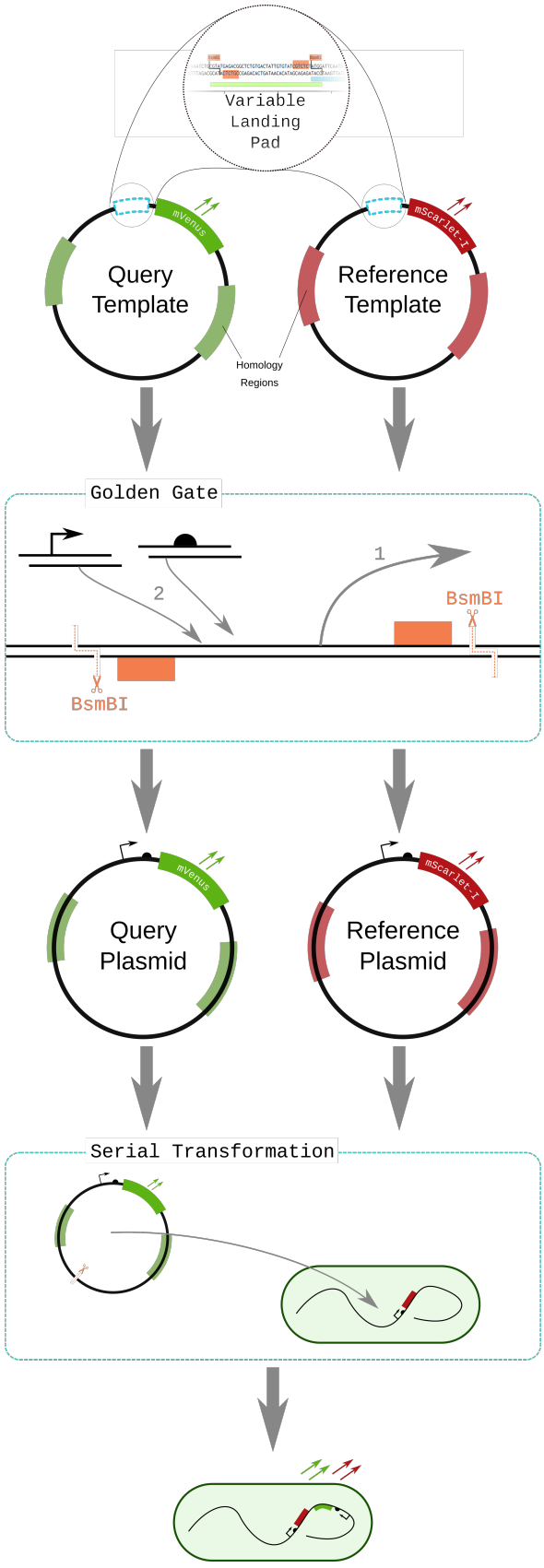



Fig. 4.8 Illustration of the serial assembly architecture. Two template plasmids are constructed, each containing the easily variable 'landing pad' region described in the text and illustrated in figure 4.3. The chosen promoter and RBS sequences are inserted then into the landing pad regions for each template. Typically only one reference plasmid is generated while an entire library of query plasmids can be constructed. The reference and query plasmids are then linearised and transformed into *B. subtilis* in serial into loci defined by the homologous sequences included in the template plasmid backbones. In this way, a library of query strains can be swiftly and reliably generated.

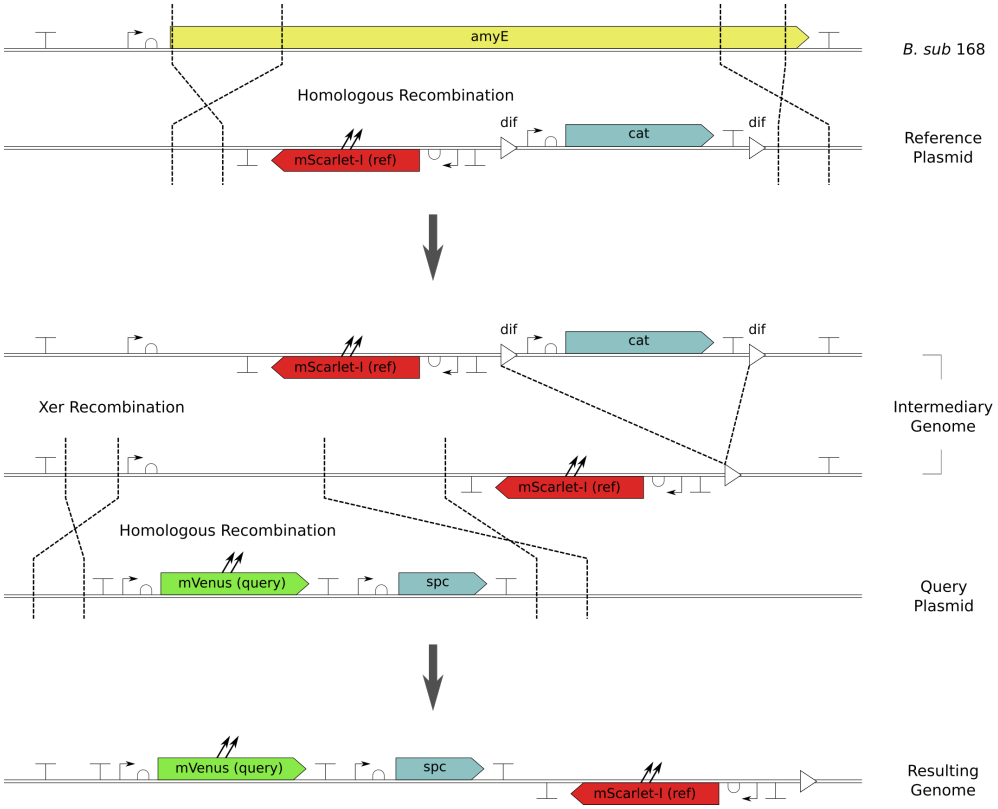



Fig. 4.9 Schematic of the serial transformations carried out during the serial assembly shown in figure 4.8. Having generated the query and reference plasmids seperately as described in figure 4.3, the reference plasmid is inserted into the genome using homologous recombination as shown in the first step. The antibiotic marker is the removed by Xer recombination, prior to the query system being inserted again via homologous recombination. The final system includes both query and reference system, separated by the remaining spectinomycin resistance cassette and insulated by terminators.

the reference plasmid was cured out in this way, with dif regions placed either side of the cat resistance cassette. Growth on antibiotic media was slowed, but once a positive colony was selected, growing for a couple of hours in plain media and replating onto plain media was typically enough for the resistance marker to be removed, the protocol used is described more fully in Section 2.3.1.

Removing the second antibiotic selection marker would prove more cumbersome, however. One practical consideration is that the resolution of the dif regions leaves one dif site as a scar, which would then be resolved with the scar remaining from any other nearby antibiotic resolution. Careful design would be needed to make sure that no such excisions cause any of the reporters to be excised, and that sufficient space remains between them.

The main difficulty in removing the other resistance marker was the choice of resistance marker – spectinomycin – which was undoubtedly a mistake. Even when using relatively high levels of the antibiotic, it appears to act more as a bacterio-suppressant than a bacteriostatic antibiotic in *B. subtilis*, with resistant 'colonies' appearing as small bulges of faster growing bacteria amid a lawn of negative growth. In most cases, it was necessary to streak out cells from these regions onto fresh antibiotic plates in order to isolate individual cells. Since including the dif sites decreases the efficacy of the resistance markers due to a lower proportion of the cells containing the marker, it seemed as though including the dif regions was unlikely to work. One improvement to this system would be to replace the spectinomycin resistance marker with something more effective, such as the chloramphenicol marker used in the other plasmid.

### 4.2.5 Quantification and Data Processing

Previous dual reporter systems have used bulk assay methods such as plate reader assays, and a particularly thorough method for doing so is described in Yordanov et al. [177]. A plate reader records the evolution of fluorescence and OD for a culture growing in a small ($\sim$150 µL) volume of media. The ratiometric output is calculated by extracting samples taken during exponential growth (estimated from blank corrected OD values), removing OD-related auto-fluorescence and taking the ratio of query to reference output.

This method has several advantages, most notably the simplicity of running experiments, although the experimental set-up can be somewhat convoluted due to limitations of the proprietary control software provided by plate reader vendors. In some cases (e.g. BMG Labtech), the control software is unable to natively collect OD and fluorescence data in a single run, even though there is no hardware impediment to doing so. The solution is a script which repeatedly runs three separate programs - one collecting fluorescence data, one collecting OD data, and the third controlling the heating and shaking between data-collection.

An external executable is then called to merge the data output by these programs, however the data stream will become corrupted if the plate layouts specified separately to each of the three programs do not match exactly. This convoluted set-up means that what should be a simple and reliable experimental protocol is actually rather tedious and error prone, as plate reader vendors don't expect their hardware to be used in this way.

Another downside of the technology is that no per-cell information is gathered. The plate reader simply measures the average fluorescence observed in the media, and can say nothing about the variance of that output. Furthermore, small volume exponential growth is the only condition which can be reliably investigated, and once stationary phase is reached cells will often begin to clump together at the bottom of the small wells, causing drastic variations in sensor output.

A technology which would allow for single cell measurements is quantitative fluorescence time-lapse microscopy (QFTM)[178], in which a single cell is seeded onto an agar pad in a climate controlled environment. The growth of this cell is then recorded with motion stabilised fluorescent microscopy, producing a time-lapse video of the formation of a colony which can then be analysed. QFTM has proven highly effective in studying the dynamics of systems as cells grow and divide, but the complexity of the experimental set-up as well as the limited conditions that can be studied – in particular the need for solid media which also makes accurate dosing more complex – make it unsuitable as a standard for ratiometric characterisation.

Flow cytometry offers a compromise between the simplicity of plate-reader methods and the single-cell resolution of QFTM. A sample of cells are suspended in media which passes through a microfluidic setup which ensures a very narrow stream of media, approximately one cell in diameter. Laser light is shone at this narrow stream, and deflected by objects in the stream. Each event is recorded with a time-stamp, as well as a range of other information about the event depending on the specific machine employed. This information is most often saved directly in one of a small number of commonly used formats, which can later be analysed using either a commercial suite or with some relatively straightforward scripting.

Unlike plate readers, flow cytometers capture per-cell information, but unlike plate readers and QFTM they cannot capture time-lapse information (at least, not while tracking individual cells). In most cases this is not a serious limitation unless the temporal dynamics of the output are of particular importance, which is unlikely to be the case when characterising constitutive expression systems or inducible expression systems under a constant level of induction.

Most flow cytometers are able to simultaneously record fluorescence data for a range of excitation and emission wavelengths as well as Forward SCatter (FSC) and Side SCatter

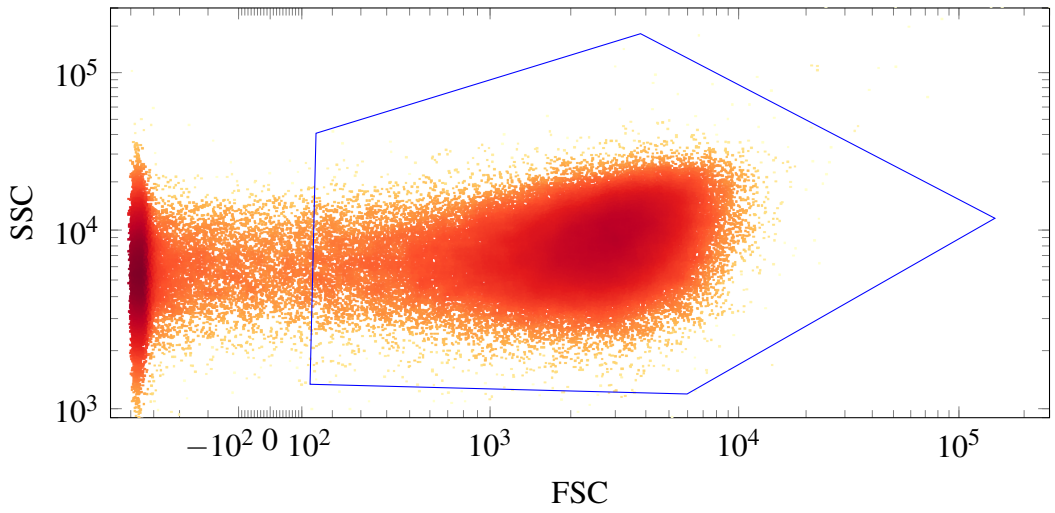



Fig. 4.10 Forward and Side scattering for flow cytometry of an autofluorescing example. The blue polygon shows the gate used throughout the analysis to select cells, in this example 71.3% of events pass the gate, which is lower than most samples collected. The concentration of events to the left of the gate are most likely spores or small debris from lysed cells, and so these events are removed. In some cases, another concentration of events is seen above and to the right of the main grouping, and these are most likely doublet events where multiple cells pass the sensor at once. These doublet events are also rejected by the gate.

(SSC) information. FSC refers to the brightness observed directly in the path of the laser beam once it has passed through the flow of media. When there is no deflection of the beam, a small bar prevents the laser beam from directly entering the sensor, however when a large object such as a cell is between the laser source and sensor, it causes scattering of the light around the bar which then enters the sensor. The amount of forward scattering depends on the size of the object, as the larger the object the more of it protrudes around the bar and can be seen from the FSC sensor. SSC refers to the brightness detected at a sensor placed perpendicular to both the stream and laser beam. Differences in index of refraction within the particle cause light to be internally scattered and emitted in all directions, and the degree to which light is scattered perpendicularly depends on the internal complexity of the object.

Analysis of this data was carried out using the *cytoflow* package in *python*. The first step is to screen the input data based on the FSC and SSC values, such that only particles which correspond to whole cells are passed on for further analysis. This process is often referred to as 'gating' within the flow cytometry community, and a representative example of the gating applied to *B. subtilis* is shown in Figure 4.10.

Once the samples have been gated by FSC and SSC, samples whose reference signal is below the $95^{\text{th}}$ percentile of the auto-fluorescence measurements are discarded. Ratiometric values from these samples would likely be very noisy as the denominator would be close to

zero, and so removing them increases the stability of the ratiometric output. In the case of a sufficiently bright reference signal, very few samples should be removed during this step.

Next, the average auto-fluorescent contribution to the output of each of the reference and query channels is calculated subtracted from those outputs. This can be done on either an experiment-wide basis, or individual autofluorescent controls can be used to independently control each growth condition. Having experimented with using both techniques, it was found that at least when using our particular flow-cytometry setup (described in detail in Section 2.4), autofluorescence from both channels was minimal, and its removal had very little effect on the ratiometric output.

Finally, the ratio of the query to the reference channel is taken to give the final ratiometric output which can be further analysed. Care should be taken when doing so as assuming that both query and reference channels are normally distributed, the ratiometric output will be distributed according to a Cauchy distribution. While the Cauchy distribution looks not dissimilar to a normal distribution, it can be rather inconvenient and is what mathematicians refer to as a 'pathological' distribution, as it has no mean or variance. Estimates of the mean made from samples of the distribution do not converge, and in fact such estimates themselves will be distributed identically to the Cauchy distribution. This issue can be avoided by instead calculating the median of the samples, as this is well defined and will converge with a large enough number of samples.

## 4.3 Construction, Verification, and Characterisation

The remainder of this section describes the construction of a dual reporter system in the manner described above and its use during an initial characterisation effort in *B. subtilis*.

### 4.3.1 Motivation

In addition to demonstrating the use of the dual reporter technology, the goal of this section is to establish a small set of well characterised expression systems for future use in *B. subtilis*. As discussed previously, this will simplify future engineering efforts in the organism as an expression system with the desired properties can simply be selected from the library at design time.

The main question when planning this characterisation effort was one of scale – since transcription and translation are coupled in bacteria, it is reasonable to suspect that the choice of promoter and the choice of RBS are not fully independent, and that the promoter sequence may in fact effect the effective strength of the RBS and vice-versa. Testing whether this is the

case requires testing all promoter-RBS combinations in the library, and drastically increases the number of strains required.

In order to keep this potential combinatorial explosion under control, it was decided to select only a handful of promoters and RBS for inclusion in the library. In order to construct a library which encompasses the full range of expression strength in *B. subtilis*, the choice of promoter and RBS are important, and must also represent the full range of expression potential in the organism.

Gene locus also effects expression strength in most organisms, with factors such as genome availability significantly affecting expression. This is important for engineered systems; if a part which has been characterised in one locus is inserted into an un-characterised locus, it is possible that the locus itself will affect the behaviour of the part in an unpredictable manner. To combat this, a method for ratiometric characterisation of different loci is introduced and tested for three different loci in *B. subtilis*.

### 4.3.2 Choice of Reference System and Construction

Since $\sigma^A$ is the major sigma factor in *B. subtilis* it was decided that this initial characterisation would focus only on $\sigma^A$ promoters. It was necessary therefore to select one well-behaved constitutive $\sigma^A$ promoter as a reference promoter. The promoter chosen for this task is the promoter of the *penP* gene, a precursor of beta-lactamase. This promoter, known as $P_{pen}$ (see sequence in Table 4.4), was chosen as it has been used previously in literature and in our lab and known to be is a strong, constitutive promoter active in both *B. subtilis* and *E. coli*. Despite its common use in *B. subtilis*, the version of $P_{pen}$ commonly used is from *B. lichenformis*, although the activity of the native $P_{pen}$ is shown in Figure 4.12 for completeness' sake.

The RBS chosen was that of *gsiB*, a glucose starvation-inducible protein, which was found to be highly active in both *B. subtilis* and *E. coli*, see Table 4.5 for the sequence information. Stress response proteins seemed to be a fertile place to search for highly active RBSs, as transcription only begins after the onset of stress and so it seems logical for the translational efficiency of stress genes to be quite high in order to produce stress response proteins as quickly as possible and maximise the potential for survival.

The reporter strains were then constructed in the manner as described in Section 4.2.4 and Figure 4.8. The reference strain – which contains only the reference reporter system – was verified thoroughly by Sanger sequencing before competence was induced as this strain affects all subsequent characterisation strains.

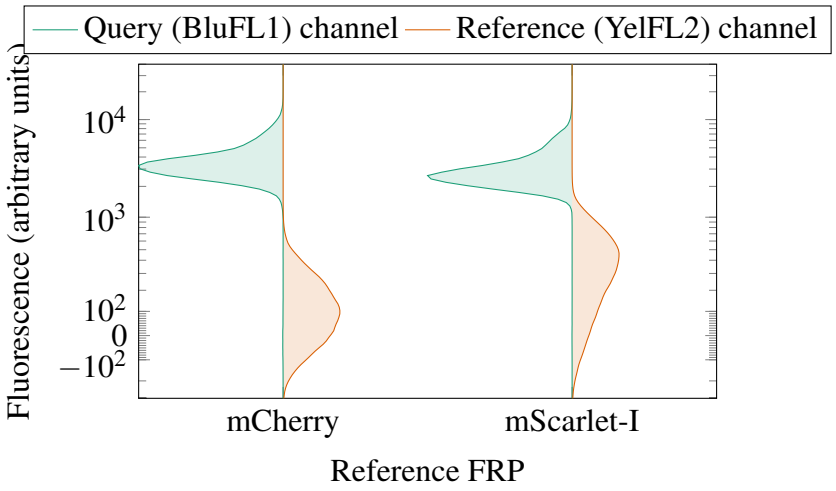



Fig. 4.11 Effect of changing reference FRP from mCherry to mScarlet-I shown on a logistic plot. Median value of reference increased by a fold change of 2.3 from 95 to 318, while the median of the query decreased by a fold change of about 0.3 from 3262 to 2660. The reference expression system was used in all cases, and so the uncorrected ratiometric output for the system changed from $\frac{3262}{95} = 64.3$ to $\frac{2660}{318} = 8.36$

### 4.3.3 Control Strain and Initial Experiments

Initial experiments with the dual reporter focussed on demonstrating a working strain with both reporters under control of the reference expression system. Since both reporters are expressed by equivalent expression systems with the same copy number, any differences in brightness are due to other factors such as the intrinsic brightness of the individual reporters or how closely the fluorescent outputs match the filter set in use. The ratiometric output of this strain therefore gives the correction factor by which other strains should be divided to give the true ratiometric output. In an ideal ratiometric system, this ratio would be equal to unity, signifying that the two reporter systems have identical gains from expression system output to fluorescent output.

The ratiometric output of two variants of this system were investigated, both used mVenus as the query FRP while one used mCherry as the reference and the second used mScarlet-I. The uncorrected output of both systems is shown in Figure 4.11, showing that this correction factor is closer to unity when mScarlet-I is used as the reference system. In this case, the flow-cytometer's gain for each channel were set to be equal to each other, however better resolution is possible by setting the gains such that the brightness observed for these strains is closer to unity.

### 4.3.4 Choice of Promoters

Approximately 3,242 transcriptional units are believed to exist in *B. subtilis*, with transcription of approximately 1,868 (58%) of those initiated by $\sigma^A$ dependent promoters[112]. There is therefore a vast library of native promoters from which to choose from when selecting a handful of promoters for characterisation, even after limiting ourselves to $\sigma^A$ promoters.

From this library, we wish to select a set of promoters which have:

- a range or transcriptional activity from the lowest to highest
- no known inducers or repressors such that their activity is constitutive

This selection can be made simpler by making use of the dataset provided by Nicolas et al. [112], in which tiling array data was used to detect transcriptional units, predict sigma factor binding sites and quantify the transcriptional strength of native promoters in *B. subtilis*.

The experimental method is described in detail in Nicolas et al. [111], however it is described briefly here. Firstly a tiling array is created, this is simply a chip containing an array of DNA probes which bind to cDNA samples derived from harvested RNA. The genome of many bacteria is small enough that chips can be designed which cover the entire genome, such that a single experiment can determine the concentration of cellular RNA present for all parts of the genome. By developing a model of RNA-probe interactions with overlapping probes, Nicolas et al. [111] were able to process the raw chip data (using the concentration of genomic DNA as a control), and produce traces of predicted RNA concentration for every base in the genome. Data was collected for 104 different conditions, including different growth phases, competence, sporulation, and germination, as well as a range of different stresses, carbon sources, and inducers (arsenite was, unfortunately, not one of them). Transcriptional Start Sites (TSS) and terminators can then be identified by looking for up and downshifts in this data. For each predicted transcriptional unit, the region from -60 to +40 surrounding the predicted TSS was extracted and used to construct a a Hidden Markov Model (HMM) of the sigma factor binding site. By creating several such models, and combining sequence information with the information on transcriptional correlation between conditions, the sigma factor responsible for initiating transcription was also predicted.

Needless to say, this dataset is a valuable resource for synthetic biology in *B. subtilis*, however, as we shall see, the native characterisation of promoter regions carried out in this paper is not a full replacement of proper characterisation as promoter sequences will often not behave as they appear to naturally when placed in a synthetic context. However, this data is highly useful in helping narrow down the search for promoters, and in particular to give a good chance of selecting promoters with a range of activity.

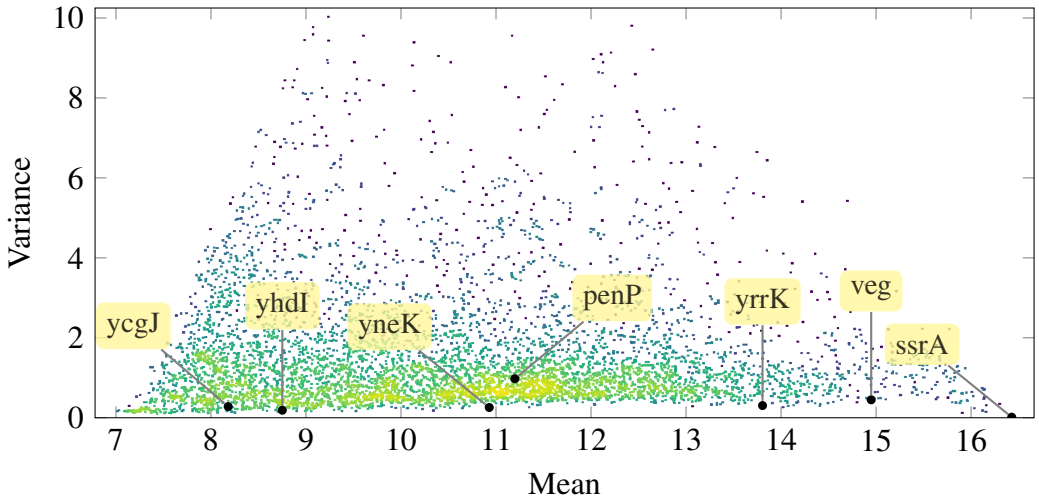



Fig. 4.12 Mean and Variance of Transcriptional Unit (TU) concentration across conditions from Nicolas et al. [112], annotations show promoters which were chosen for characterisation. The reference promoters, *penP*, is also included for completeness, although the *penP* promoter from *B. lichenformis* was actually used as a reference. The units are arbitrary, but are proportional to the average of the $\log_2$ of the corrected brightness observed for the TU.

Figure 4.12 shows the mean and variance of the normalised concentration of each $\sigma^A$ related transcriptional unit from the Nicolas dataset. The distribution forms a trapezoid with a greater variation in mean for promoters which exhibit lower variance. While the variance ranges from close to zero to the full range of expression strength, 50% of promoters exhibit a variance of around 10% or less of the full range.

Shown annotated on Figure 4.12 are the locations of the six native promoters chosen for inclusion in the study, plus the location of $P_{pen}$, the native homologue of the reference promoter. $P_{veg}$ was chosen for inclusion because it is commonly used in literature[46, 103], while other promoters were selected due to their low variance (implying they are constitutive) and broad range of activity. A literature search was also performed for each promoter to confirm that no information exists to suggest that any are in fact inducible promoters.

In order to extract the exact promoter sequence, the estimate of the exact TSS was taken from the Nicolas dataset, which predicts the start of transcription from a combination of probe upshift data and the location of the predicted sigma factor binding site. The sequence from 54 bases upstream to the TSS was then extracted for use as the promoter region. This relatively short sequence length was chosen such that promoter DNA fragments would be relatively short ( 80bp including BsmBI insertion sites and compatible overlaps) and appropriate for synthesis from complementary primers. The selection of only a short sequence also reduced the chance that the inducer binding domain was included in the synthetic promoter should

| # | Name | TU | Mean | Variance | Sequence |
|---|---|---|---|---|---|
| 1 | penP | – | 11.2 | 0.98 | TTTCGTCTCACGTATCATCATTTCCTTCCGAAAAAACGGT TGCATTTAAATCTTACATATGTAATACTTTCAAAGACTGA GACGAAA |
| 2 | veg | – | 14.9 | 0.45 | TGTCAAAATAATTTTATTGACAACGTCTTATTAACGTTGA TATAATTTAAATTTTATTTGACAAAAATGGGCTCGTGTTG TACAATAAATGTAGTGAGGTGGATGCAATGGCGAAGACGT TGTCCGATATTAAAAGATCGCTTGATGGGAATTTAGGTAA AAGGCTGACGTTAAAAGACTGAGACGAAA |
| 3 | ssrA | U2644 | 16.4 | 0.01 | TTCGTCTCACGTATAGTGCTTGATTCGAAAATCAGGCCTG TGCTATACTGTGTTCACGATCAGATCAAGACTGAGACGAA |
| 4 | yhdI | U753 | 8.8 | 0.19 | TTCGTCTCACGTAGGATCATCTTCTCAATCCCCTTTCCTT TCGTTACAATGATCATAGCTAATAAATAGACTGAGACGAA |
| 5 | yrrK | U2113 | 13.8 | 0.31 | TTCGTCTCACGTAGAAATCCGTTTTATAATTCGCCCATGT AGTGTAGAATGATGGTAAGATTAGGACAGACTGAGACGAA |
| 6 | yneK | U1449 | 10.9 | 0.26 | TTCGTCTCACGTATTTACTGTGTTCTTCTAATCTCTCTTC GTATACAATATGGTTAGAATAAGGGGAGACTGAGACGAA |
| 7 | ycgJ | U235 | 8.2 | 0.28 | TTCGTCTCACGTATATCTTGGACGAATCATAACAGAAATT GCTAACATAATCCATATCATCATTTAGACTGAGACGAA |

Table 4.4 Overview of promoters selected for characterisation

it later transpire that any of the selected promoters are inducible. Were we attempting to characterise the native promoter this would of course be unacceptable, but since we are instead generating a library of promoter which are characterised and easy to work with, it seems sensible to minimise the length of the promoter sequence. Information about each promoter and its sequence is displayed in Table 4.4.

### 4.3.5 Choice of Ribosome Binding Sites

The latest version of the *B. subtilis* 168 genome[5] has 4,457 annotated genes, each of which would be expected to have an RBS immediately upstream. Unlike the case with promoters above, no data on whole-genome translation rate exists for *B. subtilis*, and so this wide array of possible RBSs cannot be condensed into a handful of candidates in the same way as with promoter regions.

We can reduce the number of candidates by only considering those RBSs which are included within 5' UnTranslated Regions (UTRs), as our reporter gene will be the first (and only) within its transcriptional unit. The Nicolas dataset can help us here, as it predicts the location of the start site of 3,242 transcriptional units. For predicted transcriptional start site, the 5' UTR sequence was estimated by extracting the bases between the start site and the start of the next annotated gene on that strand. The length of these predicted UTRs is shown in the left-hand plot of Figure 4.13. This data is likely to be somewhat noisy due to the fact that the TSS estimation has not been confirmed experimentally, and thus could be inaccurate in many cases, and similarly it is likely that some gene annotations are missing and some are incorrect. Despite these problems, there is a clear preference for UTRs in the range of 30 to 50bp in length, although the median length is 64bp, due to the spurious prediction of several excessively long RNAs.

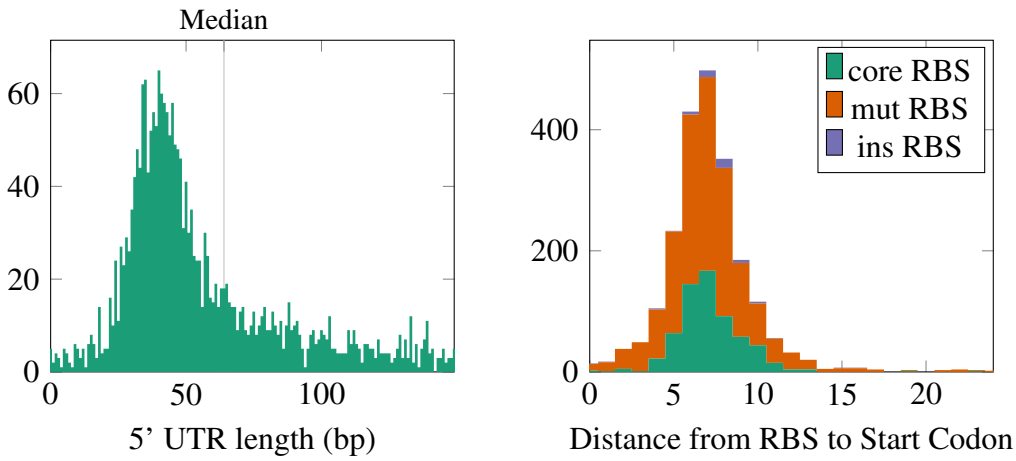



Fig. 4.13 (Left) Length of predicted UTR sequences, predicted by distance between predicted Transcriptional Start Site (TSS) from Nicolas et al. [112] and the next start codon; (Right) Shortest gap between predicted RBS sequence and start codon. 'Core' RBS refers to the sequence AGGAGG, while 'mut' and 'ins' refer to single base mutations and insertions of that sequence respectively.

Multiple sequence alignment of all of the extracted UTRs aligned strongly to the core RBS sequence in *B. subtilis*, *AGGAGG*, as expected. A simple algorithm was then used to predict the location of the RBS within the UTR. Each sequence was first searched for exact matches to the core sequence. Each sequence which did not contain this sequence was then searched for sequences with a single base mutation from the core sequence, and finally the remaining sequences were searched for sequences with a single base insertion from the core sequence. In cases were multiple matches were found, the one closest to the start codon was preferred.

Using this method, RBSs were identified in 2,571 of the 3,242 original sequences, with the majority of those found (1,719) being single base mutations from the core sequence, while 802 exact matches to the core sequence were found. Only 51 insertion matches were found, which was to be expected as binding to these domains requires the ribosome's binding domain to accommodate an extra base.

The right-hand side of Figure 4.13 shows a histogram of these predicted RBSs against the gap between the RBS and the start codon of the downstream gene. This distance is defined by the size of the ribosome as most efficient translation initiation occurs when the p-site is able to contact a start codon when the ribosome is bound to a RBS, and the distance of 6-8bp agrees with previous data[162].

Having understood more about the natural variations present in native UTRs and RBSs present in *B. subtilis*, we are better prepared to select a handful of native UTRs for use in the

| # | Gene | Core Sequence | Length | Gap[a] | Sequence |
|---|---|---|---|---|---|
| 1 | gsiB | AGGAGG | 23 | 8 | AAACGTCTCAAGACACCAATTAAAGGAGGAATTCAAAATGGTGAGACGAAA |
| 2 | yjzC | AGGAGG | 15 | 7 | AAACGTCTCAAGACAAAGGAGGAAAAACAATGGTGAGACGAAA |
| 3 | metA | TGGAGG | 28 | 7 | AAACGTCTCAAGACTAAACGGGGAAATAATGGAGGTGGCACGATGGTGAGACGAAA |
| 4 | tgl | AGGGGG | 29 | 7 | AAACGTCTCAAGACTATCTTATAAAAAACAAGGGGGGCTAAACATGGTGAGACGAAA |
| 5 | comFA | AGGAGG | 29 | 8 | AAACGTCTCAAGACGCATACTGTTTCGAAAGGAGGCGTGCTATATGGTGAGACGAAA |
| 6 | gsiB7 | AGGAGG | 22 | 7 | AAACGTCTCAAGACACCAATTAAAGGAGGAATTCAAATGGTGAGACGAAA |
| 7 | gisB6 | AGGAGG | 21 | 6 | AAACGTCTCAAGACACCAATTAAAGGAGGAATTCAATGGTGAGACGAAA |
| 8 | gsiB5 | AGGAGG | 20 | 5 | AAACGTCTCAAGACACCAATTAAAGGAGGAATTCATGGTGAGACGAAA |

Table 4.5 Overview of 5' UTRs selected for characterisation. (a) separation between core RBS sequence and start codon

library. 264 of these predicted 5'UTRs are 30bp or less in length, which makes them ideal for synthetic projects, as their brevity makes them cheaper and easier to synthesise and also reduces the chance that they contain significant secondary structures, as it is easier for larger pieces of DNA to form secondary structures.

Table 4.5 shows details of the four UTRs which were selected to complement the reference UTR, that of *gsiB*. *yjzC* was selected because if it's extreme compactness at only 15 bp in total, while *metA* and *tgl* were selected as two examples of single base mutations from the core sequence, and *comFA* was selected as its core sequence is 8 bases away from the start codon.

In addition, three variants of the *gsiB* UTR were constructed, *gsiB7-5*, which contain deletions between the core sequence and start codon such that the gap becomes 7, 6, and 5 bp respectively. The relative strength of these four UTRs together will tell us something about the preference of the *B. subtilis* ribosome for gaps of different length – although we should not completely disregard the possibility that these deletions change the secondary structure around the RBS and affect the translation initiation rate in that way.

### 4.3.6 Choice of Loci

As in Eukaryotic cells, the chromosome of bacterial cells is very long when compared with the size of the cell and therefore must be tightly packed in order to fit. Prokaryotic cells lack a nucleus, but DNA is instead organised into a region called the nucleoid and the level of compaction is controlled by various Nucleoid Associated Proteins[164]. Areas of the genome which are more tightly coiled are less accessible to DNA binding proteins such as sigma factors, and so genes in these regions are effectively down regulated. Thus, an expression

system which has been characterised in one locus may behave differently when inserted into another region of the genome due to differences in genome availability.

Two commonly used loci in *B. subtilis* research are the *amyE* locus and the *lacA* locus. The *amyE* locus first established as a convenient locus for homologous recombination due to the ease of screening for AmyE$^-$ mutants as these mutants are unable to metabolise starch[62]. *lacA*, which encodes beta-galactosidase, was later developed as an alternative insertion locus to *amyE* which allowed two regions of recombinant DNA to be inserted sequentially into the same organism[59].

Given how commonly these loci are used, it seems sensible that they should be included in any effort to characterise different genome regions, however we have no particular reason *a priori* to suspect that either of these locations should be particularly repressed. One such region is the *cotVWX* region, which contains three late stage sporulation genes which are involved in the formation of the spore coat, see Section 1.3. Since these genes are only required during the latter stages of sporulation, shortly before the mother cell lyses, it seems likely that these genes will be tightly repressed at other points in the cell cycle, and since these genes are three of several sporulation specific genes clustered together in the genome it seems feasible that a change in nucleoid compaction may be partly responsible for this. For this reason, the *cotVWX* locus was also chosen as a third locus for comparison.

## 4.4 Results and Discussion

### 4.4.1 Ratiometric Experiments

Ratiometric experiments were carried out by growing strains overnight in LB media at 37 °C with shaking and then inoculating 1 ml of pre-warmed fresh media from those overnights. These were then grown again at 37 °C with shaking for a further 3-4 hours until the optical density (OD) had visibly increased. Samples where the OD had increased above around 0.5 were diluted and then all samples were transferred to appropriate plasticware for analysis by flow cytometry which was carried out immediately.

Figure 4.14 shows the dual channel output of the system under a range of different query systems. While the query reporter channel varies significantly, the output of the reference system is stable, only increasing very slightly as the query system is lower. This implies that, under these conditions, the two reporters are not causing saturation of the cells' expression system or the need for molecular oxygen for proper fluorescent reporter maturation. In other conditions, however, more significant changes in reference channel output have been observed, but these changes are corrected by the dual reporter system. No difference in

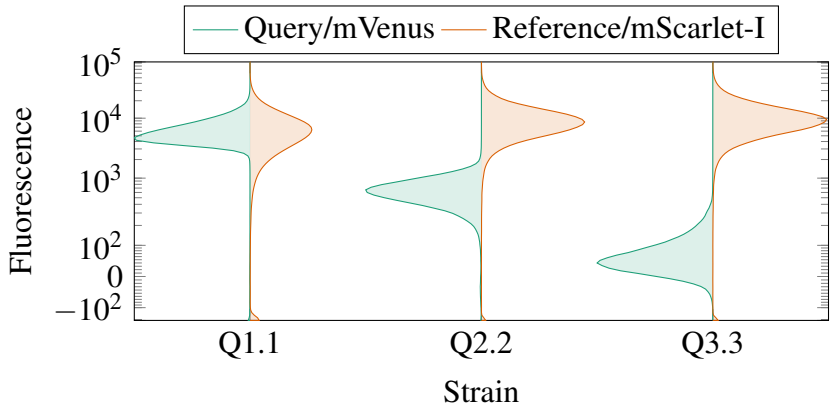



Fig. 4.14 Split violin diagram of autofluorescence corrected Query and Reference channels under different query expression systems. Cultures were grown in LB overnight, then used to inoculate fresh pre-warmed LB and grown for 3 hours. Gain of the two channels was adjusted such that the ratiometric output for the reference strain Q1.1 was approximately unity. The strength of the query reporter varies significantly with the different expression systems, but there is also a slight increase in reference output as the reporter decreases. Strains are named in the form Q<*Promoter ID*>.<*RBS ID*>, where 'Promoter ID' and 'RBS ID' refer to values in tables 4.4 and 4.5 respectively.

growth rate was observed between the query strains and the original strain, implying that the production of fluorescent reporter does not having a large effect on cells' metabolic behaviour.

Since the saturation observed in some conditions is a cell-wide effect, there is no reason to believe that it would favour the production of any one type of fluorescent reporter protein more than another. This was the reason for choosing fluorescent reporters with similar maturation rates in Section 4.2.3. Since the observed levels of both the query and reference system should therefore be affected equally, taking the ratio of the query to the reference will remove the effect of the saturation, resulting in a more accurate characterisation of the specific system being tested.

### 4.4.2 Promoter and RBS Characterisation

Query strains were assembled for each combination of promoter (see Table 4.4) and RBS (see Table 4.5), as described in Section 4.2.4. Initial experiments showed that fluorescence could only be detected for promoters $P_{pen}$, $P_{veg}$, $P_{ssrA}$, and $P_{yrrK}$, which correspond to promoter numbers 1, 2, 3, and 5. However, the activity of promoter 5, $P_{yrrK}$, was so low as to be indistinguishable from autofluorescence. The three promoters which produced no fluorescence were $P_{yhdI}$, $P_{yneK}$, and $P_{ycgJ}$, which happen to be the weaker three promoters

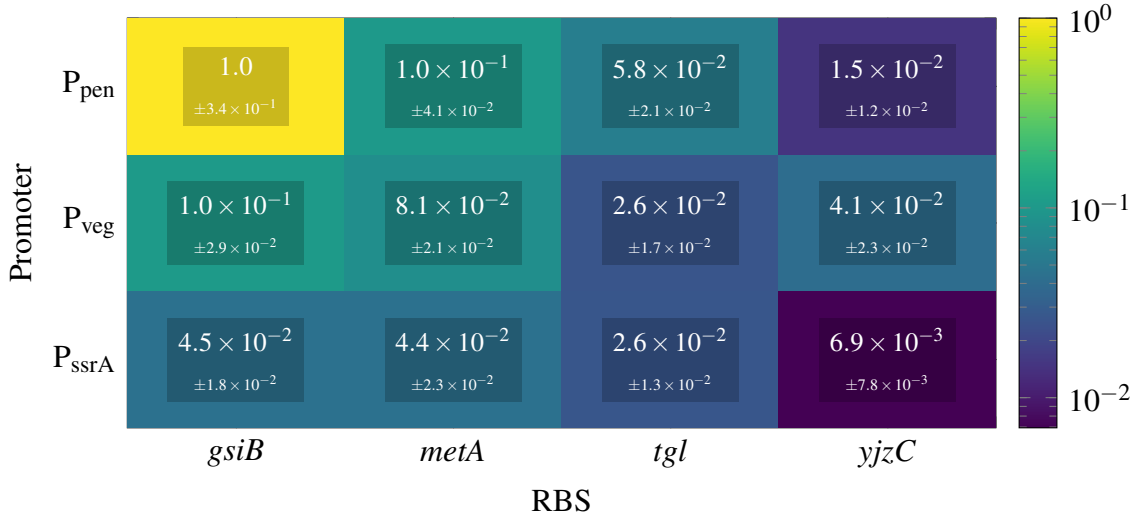



Fig. 4.15 Median ratiometric output from the promoter and RBS characterisation screen. See Tables 4.4 and 4.5 for details of each promoter and RBS respectively.

selected on the basis of the Nicolas dataset. Whether the lack of observable fluorescence is due to the low strength of the native promoters or due to the constrictive way in which the promoters were extracted is unknown.

Fluorescence was observed for all RBSs except numbers 5 and 8 (Table 4.5), which correspond to the *comFA* RBS and the *gisB5* variant, which will be discussed in Section 4.4.4.

A complete screen of each of the three functional promoters and four functional RBSs was made, using the assembly strategy described above in Section 4.2.4. Each strain was sequence verified before being quantified ratiometrically as described in the previous section. The median results of this analysis are shown in Figure 4.15.

Let us first consider the results shown in the left hand half of Figure 4.15, those involving the RBSs associated with *gsiB* and *metA*. These six results show a clear pattern in the strength of the promoters, particularly that $P_{pen} > P_{ssrA} > P_{veg}$. This is somewhat of a surprise, as $P_{ssrA}$ was chosen specifically because of *ssrA*'s exceptionally high expression as measured in the Nicolas dataset[112], but the associated promoter appears to be of only intermediate strength. It is possible that there is some other transcription initiation element upstream of *ssrA* which increases mRNA content in the studied strain, or that there is some other technical reason which caused a spuriously high reading for *ssrA* in the Nicolas dataset such as non-specific probe binding.

Secondly these results show consistently that the RBS associated with *gsiB* is stronger than that associated with *metA*, though we rapidly encounter problems when attempting to quantify the scale of this difference. Starting with the strongest expression system, $P_{pen}$-*gsiB*,

we can see that replacing the promoter with $P_{ssrA}$ or the RBS with *metA* causes the ratiometric output to one tenth of its previous value. One might logically expect replacing both, such that we are left with $P_{ssrA}$-*metA*, would reduce the output to $\frac{1}{100}^{\text{th}}$ of the original, but the observed value is actually around $\frac{1}{12}^{\text{th}}$ of the original. The effect of the RBS is further reduced in the case of $P_{veg}$, which has a strength of around $4.5 \times 10^{-2}$ times that of $P_{pen}$ when the *gsiB* RBS is used. While in the case of $P_{pen}$, changing the RBS from *gsiB* to *metA* reduces the output to $\frac{1}{10}^{\text{th}}$ of its original value, the same change under $P_{veg}$ has almost no effect.

Let us consider a simple model of a cell's expression machinery, where there are $N$ different mRNA molecules, each present at copy number $M_n$ for $1 \leq n < N$. Assuming that each mRNA molecule is equally likely to encounter a ribosome, let us define $R_n$ as the probability that translation is initiated when a ribosome encounters the $n^{\text{th}}$ mRNA, or, put otherwise, the strength of the $n^{\text{th}}$ mRNA. The rate at which translation of the $n^{\text{th}}$ mRNA occurs is therefore proportional to the fraction of that mRNA present in the cell multiplied by the probability that the encounter will result in translation initiation, or

$$\frac{M_n}{\sum_{i=1}^{N} M_i} \cdot R_n.$$

This model predicts that the number of copies of the mRNA – which is proportional to the strength of the promoter – is independent from the strength of the RBS. As we have discussed, data presented in the left hand side of Figure 4.15 do not support this conclusion, and indicate that there is indeed some connection between the translation of an mRNA and its transcription.

The picture becomes more complicated when we consider the results with the two other RBSs, *tgl* and *yjzC*. When used in conjunction with either of the previous RBSs, $P_{veg}$ is approximately half as strong as $P_{ssrA}$, however when the *tgl* RBS is used the two promoters result in the same level of output, while the *yjzC* RBS reduces the relative strength of $P_{veg}$ to about $\frac{1}{6}^{\text{th}}$ that of $P_{ssrA}$. In fact, when both using the *yjzC* RBS, the output of $P_{ssrA}$ is just over twice that of $P_{pen}$, the complete opposite of when the *gsiB* RBS is used.

It is clear that the overall strength of the expression system is not a linear function of the strength of the promoter and the RBS, as the two factors clearly interact in some surprising ways. This is not wholly a surprise, as transcription and translation are known to be linked in prokaryotes, but the fact that the stronger of two promoters can be made the weaker simply by changing the RBS is a surprise. There are of course other effects which could be at work here, especially given the small sample size. Each promoter was selected such that, based on predictions of the location of the transcriptional start site, none of the promoter sequence should appear in the mRNA and thus changing the promoter should not affect the resulting

mRNA. However, transcriptional start site estimation is not necessarily wholly accurate – and the exact start site may of course vary randomly – such that differences in mRNA structure may in fact arise depending on the promoter, which could account for some of the variation present in the dataset. Performing RNA sequencing of these transcripts would confirm or deny whether or not this is in fact the case.

### 4.4.3 Loci Characterisation

The response of each of the three functional promoters inserted into each of the three loci introduced above in Section 4.3.6 is shown in Figure 4.16. The median ratiometric output for promoters inserted into the *cotVWX* locus are between $\frac{1}{4}$ and $\frac{2}{5}$ of the output of the same promoter inserted into the *amyE* locus. Surprisingly, the two promoters which were successfully inserted into the *lacA* locus were also around a half and a third as active as when inserted into the *amyE* locus.

While these factors do not seem enormous – the Nicolas dataset shows that transcriptional activity varies by a factor of roughly $2^9$ – the fact that the choice of locus can affect the overall rate of expression by a factor of at least $2^2$ is noteworthy. For example, consider the task of efficiently expressing a recombinant heterodimer consisting of two different large protein subunits in *B. subtilis*. If the two subunits cannot be easily inserted in a single recombination event – for example if they are too large – then one might consider inserting each subunit at separate loci, with *amyE* and *lacA* being by far the most likely candidates as well established recombination loci in the organism. However, if the same expression system was used for each, the subunit inserted into the *amyE* locus would be produced in higher quantities than the other. This over-production would reduce the efficiency of the system, as metabolic flux would be wasted in producing an excess of one of the components.

### 4.4.4 RBS variants

Three variants of the *gsiB* RBS were created by deleting bases from between the core binding site (`AGGAGG`) and start codon. The original has eight bases between these two, while the others have 7, 6, and 5, respectively. Initial experiments did not detect any fluorescence from the *gsiB*-5 variant, and so the others were tested as per the procedure described above.

Figure 4.13 shows the distribution of this distance found in native RBSs, showing that 6, 7, and 8 are the most commonly observed distances. The most commonly observed distance is 7 bp, while the native *gsiB* RBS has a distance of 8. The results of this experiment, shown in Figure 4.17, demonstrate that reducing this gap to 7 bp slightly increases the strength of the RBS, while reducing it further to 6 bp dramatically reduces the strength of the RBS.

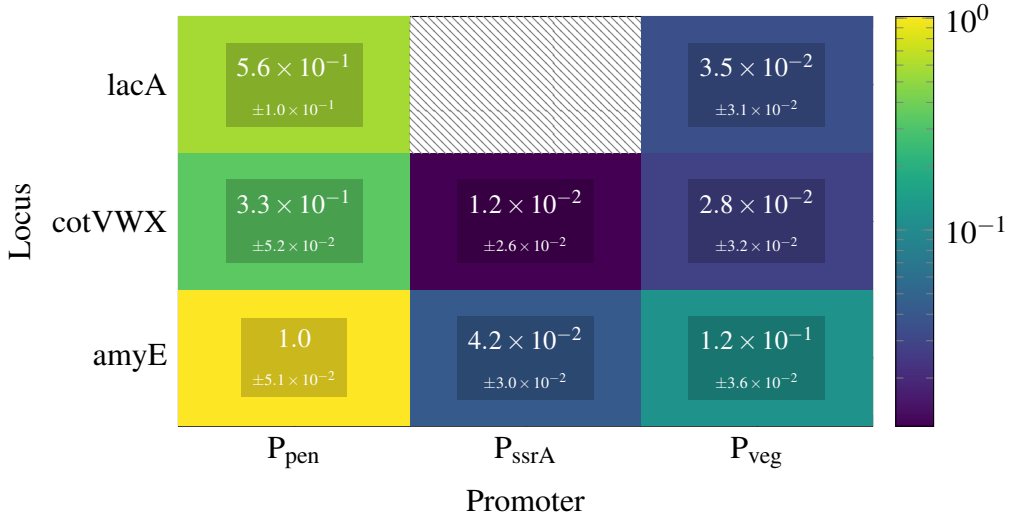



Fig. 4.16 Median response of each functional promoter in each of the three loci. All cultures were grown in plain LB media, and RBS 1 (gisB) was used in the query system throughout. Hatched boxes show strains for which query system integration failed.

Bases were removed from adjacent to the start codon to minimise the possibility of disrupting some non-core binding interaction with the ribosome, however there is still the possibility that the removal of bases has some other effect such as disrupting or introducing some form of secondary structure which could have an effect on ribosome recruitment. However, the addition or removal of single bases between the RBS and the start codon appears to be a simple way to make small modifications to the ability of an RBS to recruit ribosomes.

### 4.4.5 Growth Condition Comparisons

In order to test the robustness of the dual reporter system, characterisations of four different growth conditions were made. 'LB' refers to growth in LB media, which was conducted as before, while 'MG' and 'SV' refer to growth in Minimal Growth and Starvation media as defined in Section 2.5.5 respectively. 'S' refers to growth on LB agar solid media. Colonies were incubated at 37 °C for 24 h, before being scraped using an inoculation loop and gently resuspended in LB media by pipetting up and down. The resuspended cells were then immediately analysed using flow cytometry, within 2 to 3 min of resuspension such that fluorescent output was not affected by the change in conditions.

Figure 4.18 shows how the condition affects the output from each of the three promoters in combination with the reference (*gsiB*) RBS. $P_{pen}$ is used as a positive reference in each condition, and the results normalised such that the output from that promoter is 1.0 in each

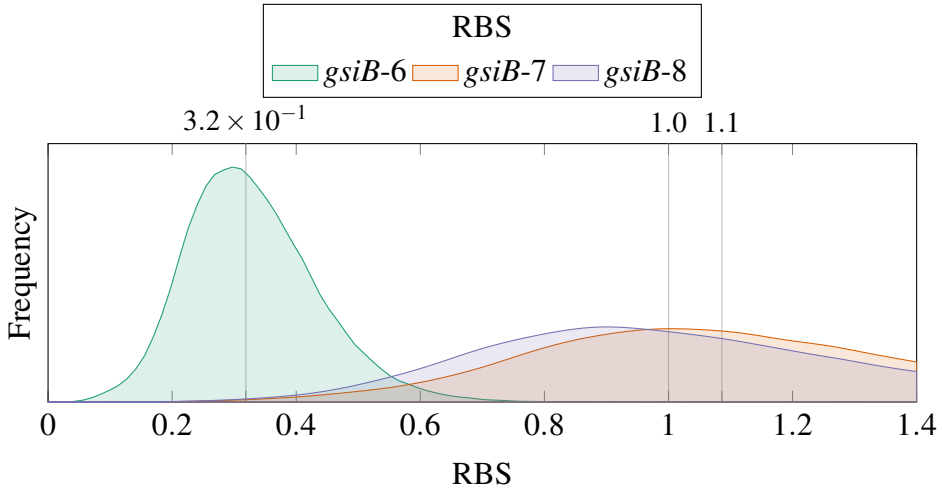



Fig. 4.17 Ratiometric output for the *gsiB* RBS variants, under the $P_{pen}$ reference promoter.

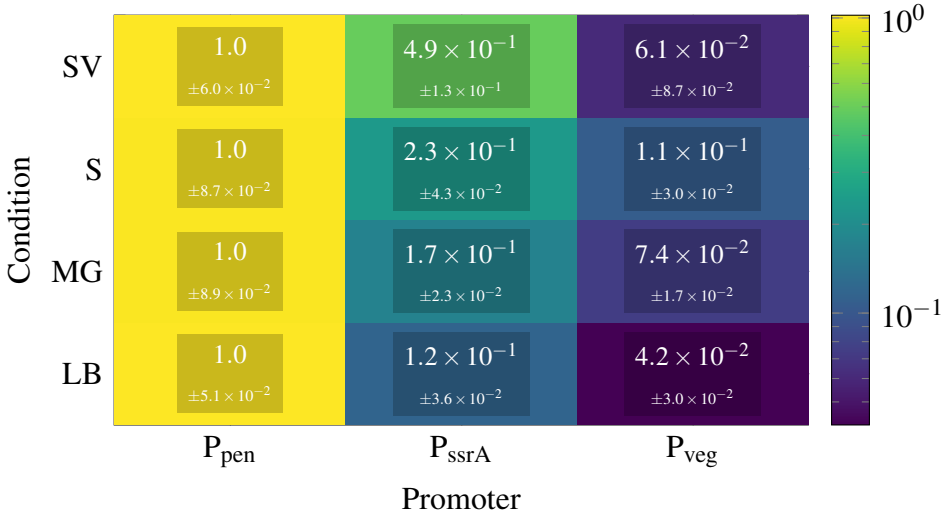



Fig. 4.18 Effect of Growth Condition on each Promoter using the reference (*gsiB*) RBS

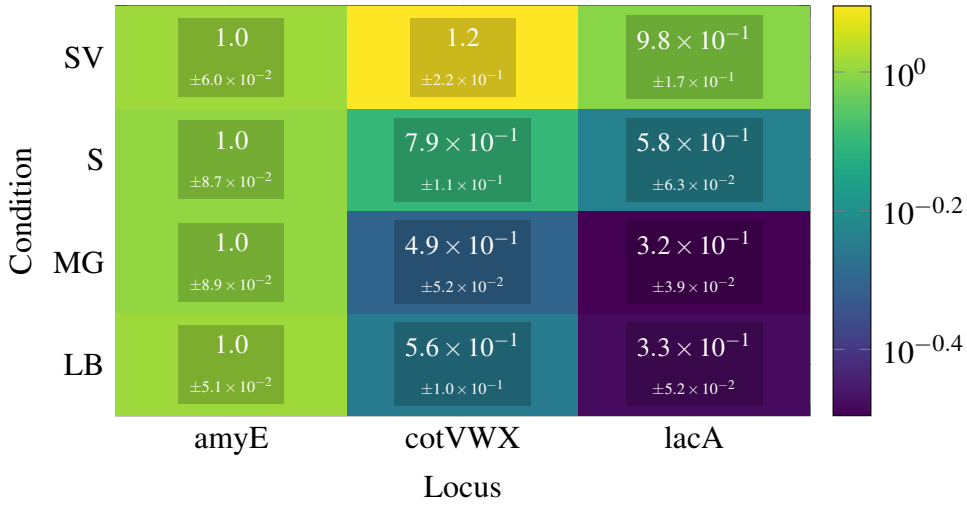



Fig. 4.19 Effect of Growth Condition on each Locus under $P_{pen}$

condition. This is to account for effects of the condition on the reporter system. For example, the ratiometric output for the reference system ($P_{pen}$-*gsiB*) is around 1.1 times that of the same system when grown in Minimal Growth media as opposed with LB media. As possible explanation for this is the difference in the availability of tryptophan in each media, since the strains we are working with are tryptophan auxotrophs. The reference reporter, mScarlet-I, contains three tryptophan residues compared with one in the query reporter, mVenus. It is therefore feasible that tryptophan starvation may affect the reference reporter more severely than the reference, causing the apparent ratiometric reading to increase. By using a reference strain, effects such as this can be minimised in order to focus on effects which are real and specific to the system under test.

The results shown in Figure 4.18 show the effects of growth condition on the ratiometric output from $P_{ssrA}$ and $P_{veg}$. In minimal growth media and on solid media, the relative output of both promoters increases by a similar proportion, while in starvation media the relative strength of $P_{ssrA}$ increases to around four times the value observed in LB, while $P_{veg}$ returns to close to the value observed in LB. The fact that the two promoters appear to behave differently suggests that there may in fact be some regulatory reason for this rather than simply differences in reporter protein expression and maturation.

Figure 4.19 shows the effect of growth condition on the different loci that were investigated previously. The expression system is kept constant in this experiment (the $P_{pen}$-*gsiB* reference is used throughout), and the *AmyE* locus is used as a reference for each condition. Again, this should minimise the effect that changing the growth condition has on fluorescent reporter expression an maturation has on the output, which should be specific to the locus in question.

The behaviour of each locus appears similar when grown in minimal growth media as when grown in LB, however growth in solid media and in particular starvation media appears to increase the activity observed in the *cotVWX* locus to over twice the original value, and three times the original for *lacA*. The case of *cotVWX* is particularly interesting as under LB the system inserted there is observed to be about half as active as *amyE*, while in starvation media the genes in the *cotVWX* locus are more highly expressed than those in the *amyE* locus. A possible explanation for this is that during normal growth the *cotVWX* locus is tightly packed to prevent leaky expression of these sporulation related genes, but that during nutrient limitation such as during growth in starvation media this region is less tightly packed allowing greater access to transcription factors and a higher level of transcription overall.

## 4.5 Conclusions

The goal of this chapter was the establishment of a dual channel reporter system in *B. subtilis* appropriate for use in the characterisation of the *ars* operon's response to environmental arsenic. Secondary goals included an exploratory characterisation of a small set of expression systems using the dual reporter system as well as investigating the effects of locus and growth conditions on the behaviour of those expression systems.

These goals have been achieved, although some weaknesses in the approach have been exposed. The output of the expression systems shown in Figure 4.15 varies by almost three orders of magnitude, although the measured RNA concentrations due to Nicolas et al. [112] vary by five orders of magnitude. While it is possible that this difference is due to the difference in approach between the two studies, it is clear that the dynamic range offered by fluorescent reporters – in particularly the sensitivity at low range – is not as great as the chip technology used in Nicolas et al. [112]. Selecting a wider range of promoters and RBSs for characterisation may help to improve this, particularly by studying promoters and RBSs with lower activities. The data processing in such a study would need to take into account the greater stochasticity expected at low expression levels as proteins are produced in short bursts when mRNA is transcribed.

The use of flow cytometry in combination with dual channel reporters is a departure from how characterisation studies have been traditionally carried out in synthetic biology. The greater level of detail provided by this technology will prove invaluable in characterising the underlying behaviour of stochastic biological systems as compared with more traditional bulk assays such as plate reader assays.

# Chapter 5

# Characterisation of the *ars* Operon of *B. subtilis*

## 5.1 Introduction

The native *ars* operon in *B. subtilis* is induced by arsenite through the interaction between arsenite and the autorepressive protein ArsR, which unbinds from the *ars* promoter in response to arsenite allowing transcription of the operon. This system was introduced in detail in Section 3.1.1, and a thorough mathematical model of the operon in a cellular context was developed throughout the rest of Chapter 3.

In this chapter, the dual channel characterisation technology developed in Chapter 4 is applied to the *ars* operon in order to improve the practical understanding of the operon.

Due to the complexity and number of parameters of the full model, fluorescent reporter based characterisation alone will not be sufficient to fully parametrise it. However, characterising the arsenite response of the system will first help to validate the modelling work done in Chapter 3, and will help us to gain valuable insights into the functioning of the operon and its suitability for use in a biosensor.

## 5.2 Applying the Dual Reporter to the *ars* Operon

### 5.2.1 Deleting the *ase* Operon

An important decision when designing characterisation strains for the *ars* operon is what to do about *B. subtilis*'s second arsenite mediating operon, the *ase* operon. While the overall effect of the *ase* operon is reported to be relatively weak when compared with the more

complex *ars* operon, it also hasn't been rigorously characterised, and so it was decided to delete the *ase* operon in its entirety and base all further characterisation on the Δase strain.

A Δase strain was initially sourced from a collaborator, but a colony PCR of this strain showed that the *ase* locus had not in fact been removed from the organism. Instead, the entire transcriptional unit was removed using the Xer recombinase system discussed in Section 2.3.1. A plasmid was constricted which contained ∼1 kb homologous regions from upstream and downstream of the *ase* operon, with the operon itself replaced with the *cat* antibiotic marker which encodes for chloramphenicol resistance. This marker was in turn flanked by diff sites – 28 base sequences which are recognised by the Xer recombinase system. The plasmid was then linearised and transformed in to competent wildtype *B. subtilis*, which were grown on chloramphenicol media overnight. Positive colonies were then suspended in plain LB media and grown at 37 °C for around 6 hours to allow for resolution of the dif regions and excision of the resistance cassette, before being re-plated on plain media this time and grown overnight. Excision of the cassette was confirmed by a lack of growth on chloramphenicol media, and the deletion of the *ase* operon was then confirmed by colony PCR.

This deletion leaves one dif site in the location where the *ase* operon was previously present. This would be problematic if the *ase* and *amyE* loci were nearby as the remaining dif site would be resolved with the one left behind after insertion of the reference reporter system. While these loci are actually not too far apart at a whole genome level, the quarter of a mega-base of sequence which separates them should prove enough to prevent excision[11].

### 5.2.2 Implementing the Reporter

#### Extracting the Promoter

The mScarlet-I based reference reporter system described in Section 4.2.4 was the inserted into the amyE locus as before. The choice of reference system – $P_{pen}$ with *gsiB* UTR – was unchanged. The insertion was again verified with PCR and Sanger sequencing as before.

While the reference system is unchanged, the query system is different from the previously characterised systems in this work in one key respect in that the promoter region associated with the *ars* operon, $P_{ars}$, is not constitutive. The first important thing to take into account because of this is that it is important that the extracted promoter sequence includes as much of its natural context as possible, and in particular the DNA binding domain of the $ArsR_2$ dimer must be included. To ensure this, the 341bp upstream of the TSS predicted by the Nicolas dataset were included, up to the stop codon of the final gene of the adjacent transcription unit. The full sequence of the promoter which was used in the study is shown in appendix B.

### Controlling for the Extra Copy of the Promoter

The second consideration is of how the extra copy of the promoter influences the native system, which we can determine by referring to our knowledge of the operon. For any given arsenite concentration, there exists a specific balance between the concentration of the repressor $ArsR_2$, the induction level of the promoter, and the proportion of the repressor bound to arsenite which cannot interact with the promoter. This balance then sets the rate at which arsenite is extruded from the cell, finally determining the concentration of arsenite within the cell. Increasing the concentration of arsenite changes this balance, increasing the level of induction and therefore concentration of $ArsR_2$ but reducing the proportion of it which is free to bind to the promoter, such that the rate of extrusion of arsenite increases to keep internal arsenite concentrations stable.

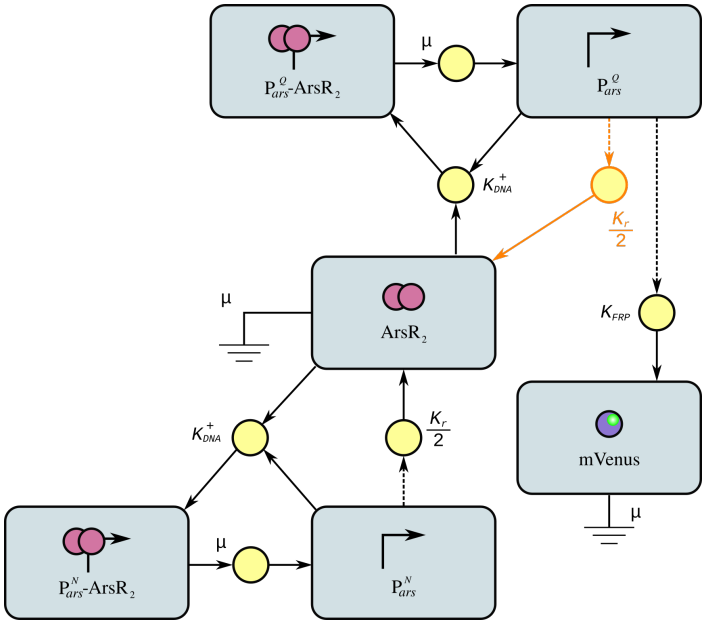
Doubling the number of copies of the promoter doubles the number of $ArsR_2$ binding sites present in the cell, which one expects to affect this balance. If the total number of copies of $ArsR_2$ which are not associated with arsenite and are thus free to interact with the promoter is much higher than the number of copies of the binding site, then the addition of the extra binding domain will have little effect on the level of induction.

The level of perturbation caused by this extra copy can be investigated by constructing a different strain in which an extra copy of the *arsR* gene is also included with the $P_{ars}$ promoter, which we shall refer to as the $P^{+}_{ars}$ promoter. Assuming that the query promoter drives expression at the same rate as the native copy – a core assumption of our characterisation effort – then the addition of the second copy of *arsR* downstream of the query promoter will double the rate of *arsR* production, which as we will show provides a lower bound on the behaviour of the native system.

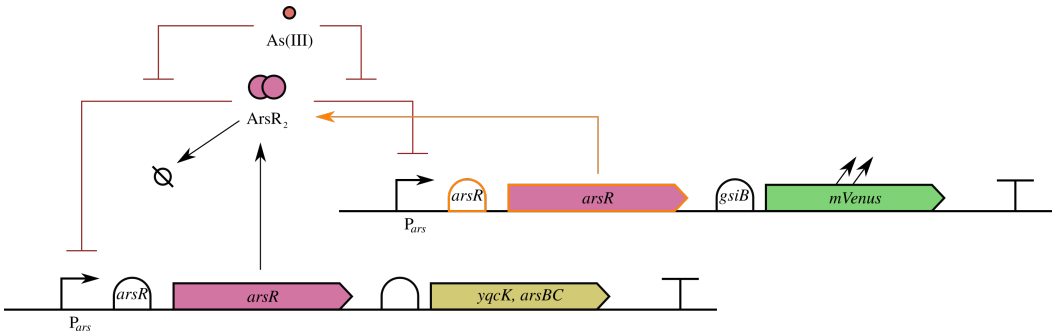
Figure 5.1 shows a model diagram of this system, based on the arsenic free model derived in Section 3.2.4. In addition to the interactions with the native system, $ArsR_2$ also represses the query promoter, reducing expression of the reporter. By including or excluding the reaction shown in orange, the model can predict the behaviour of the system with or without the extra copy of *arsR*.

This model has three important parameters, the rate of production of $ArsR_2$, $\frac{k_r}{2}$, the rate of binding of $ArsR_2$ with DNA, $k^{+}_{DNA}$, and the rate of growth dilution, $\mu$. As in Chapter 3, the rate of growth dilution is assumed to be far greater than the rate at which $ArsR_2$ unbinds from DNA, and so the latter is ignored. Similarly, the rate at which monomeric ArsR forms the homodimer $ArsR_2$ is assumed to be very fast when compared with the rate at which the monomer is diluted, and so that step is ignored in the model.

Understanding the physical interpretation of these parameters is crucial to understanding whether or not the effect of the extra copy is significant. Let $[ArsR_2]_{max}$ be the maximum




(a) Model diagram of the effect of the reporter systen on the native system in the absence of arsenite




(b) SBOL diagram of the two reporter systems along side the native *ars* operon, showing the interaction between the two

Fig. 5.1 Model diagram of the effect of the reporter system on the native system in the absence of arsenite. Figure 3.7 describes the native system; this model also incorporates the query promoter which drives expression of the reporter protein mVenus, as well as the native promoter which drives expression of the repressor, ArsR$_2$. Shown in orange in both diagrams is the effect of including an extra copy of *arsR*, by which the query promoter also contributes to ArsR$_2$ production. Including or excluding this reaction in simulations will help up to understand the effect of the extra copy of the gene. $k_F RP$ and mVenus are shown for completeness; their effects will be normalised out by the dual reporter system.

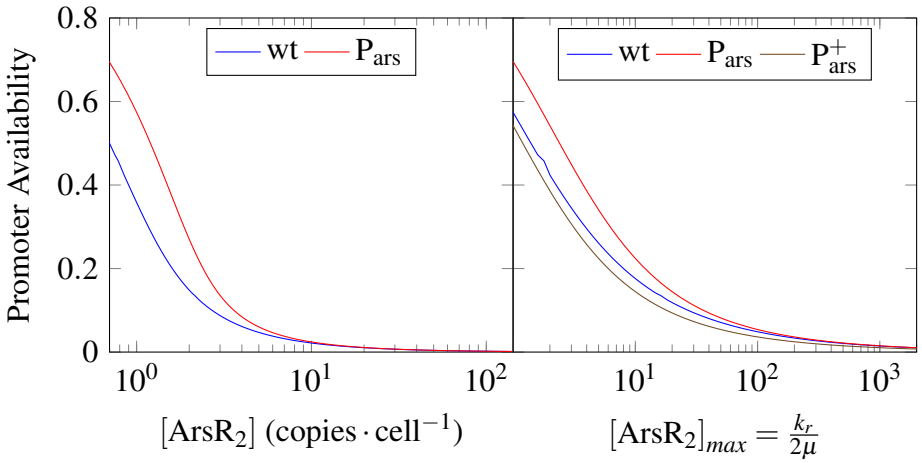



Fig. 5.2 Simulated output of the model from Figure 5.1, with $k^+_{DNA} = 3.46 \times 10^{-4}$ and $\mu = 3.46 \times 10^{-2}$ and varying promoter strength $k_r$. Left: Promoter availability as a function of total $ArsR_2$, showing the output converging for larger amounts of $ArsR_2$. A smaller value of $k^+_{DNA}$ causes the two plots to converge faster. Right: Promoter availability against normalised promoter strength, showing the outputs of both query systems, $P_{ars}$ and $P^+_{ars}$, which form the upper and lower bounds of the wild type (wt) system respectively. The wild type system always converges towards the upper bound if the promoter is strong enough, and smaller values of $k^+_{DNA}$ cause the wild type to converge with the upper limit for lower promoter strength.

possible concentration of the repressor $ArsR_2$, which is equal to the ratio of the maximum rate of production over the rate of growth dilution, $\frac{k_r}{2\mu}$. Now, imagine a system in which the concentration of $ArsR_2$ is maintained at its maximum possible value independently to the unbound fraction of the promoter. This corresponds to an upper bound on the level of repression of the system, in which the fraction of the promoter available to initiate transcription is equal to

$$[P_{ars}]_{min} = \frac{\mu}{\mu + k^+_{DNA} \cdot [ArsR_2]_{max}} \tag{5.1}$$

Since we know from basic observation of the system that tight repression of the system is possible, $[P_{ars}]_{min}$ must be close to zero and so $k^+_{DNA} \cdot [ArsR_2]_{max} >> \mu$.

The question of whether the extra copy of the binding domain significantly perturbs the system boils down to which of $k^+_{DNA}$ or $[ArsR_2]_{max}$ dominate in the above inequality. If the former is small enough and the latter is large enough, then the large pool of $ArsR_2$ will not be significantly perturbed by the extra binding domain, however if $k^+_{DNA}$ is larger and the pool of $ArsR_2$ is smaller, then the effect of the extra binding domain will be significant. This is demonstrated in the left hand side of Figure 5.2.

This simplified model does not include the effect of arsenite on the system, which is clearly important as it will affect the amount of free $ArsR_2$ which is available to bind the promoters. However, since we are not particularly concerned with modelling the interactions with arsenic in detail, we can think of the addition of arsenite as 'siphoning off' a proportion of the $ArsR_2$ which is produced, as arsenite bound $ArsR_2$ is assumed not to interact with DNA in this model. This siphoning effect causes a reduction in the concentration of active $ArsR_2$, which we can think of as a reduction in the rate of $ArsR_2$ production, which corresponds with a weakening of the promoter. Thus, the introduction of arsenite into the system is equivalent to a reduction in the value of $k_r$, which increases the possibility of the upper and lower bounds of the wild type system separating, depending on the value of $k^+_{DNA}$ and the amount of arsenite introduced.

### Influences on the Extracellular Environment

*B. subtilis* takes up arsenite from the environment, most likely through glycerol transporter channels, and extrudes it through the ArsB membrane protein, see Section 3.1.1. The purpose of the extrusion is to maintain an internal arsenite concentration that is below toxic levels, however the relationship between the internal and external arsenite concentrations is not clear. At high external arsenite concentrations – in particular concentrations greater than the toxic threshold – the internal arsenite concentration must be lower than the external arsenite concentration in order for the cells to survive.

It is theoretically possible that at lower external arsenite concentrations the internal arsenite concentration could be higher than the external concentration if the glycerol transport system is an efficient arsenite transporter at concentrations too low to trigger much activity from the arsenite detoxification system. This seems unlikely, based on the model derived in Chapter 3, and seems particularly unlikely in cases where the arsenite concentration is high enough to have any real relevance.

Assuming that the former is true and internal arsenite is always lower than external arsenite, then as a population of cells grows in a fixed volume of media, the concentration of arsenite in that media will increase as the total intra-cellular volume increases, due to the lower concentration within the intracellular volume. The higher external concentration means more arsenite entering into the cell, and thus the level of induction of the operon is expected to grow as the cell count increases.

This is a problem for our characterisation circuit, which assumes that a dynamic equilibrium has been reached at the time of data acquisition. We can avoid violating this equilibrium provided we perform our measurements while the cell count is small, as the total intracellular volume will then be negligible compared with the extracellular volume of the media, and thus

the external concentration of arsenite will not have changed appreciably. There is therefore a limited window in which measurements can take place – long enough after inoculation that an equilibrium has formed and enough cells have grown, but soon enough that the cell volume hasn't become an appreciable fraction of the total volume of the media.

## 5.3 Results and Discussion

### 5.3.1 Basal Strength of the Promoter

Before characterising the response of the system to arsenite, it is informative to characterise the unregulated, or basal, strength of the promoter. To do this, a strain lacking any copies of the regulator, *arsR*, has to be generated.

Two methods for doing this were attempted, both using the Xer recombinase system described in Section 2.3.1, which allows an antibiotic to be cured out after selection. First, the entire sequence of the *ars* operon including 1 kB homology regions either side was extracted by PCR and circularised using a high copy plasmid backbone with ampicillin resistance. This plasmid, designated pHK001 was then thoroughly sequence verified by Sanger sequencing.

Five variants of this *ars* plasmid were created by using PCR and Golden Gate assembly to replace first the entire operon and then each of the four genes individually with the 'dcat' sequence – a chloramphenicol resistance cassette flanked by *dif* sites which are resolved by the Xer recombinase system. Each of these plasmids was again sequence verified before being transformed into a competent *B. subtilis* which already included the reference and query systems.

The aim was to generate five separate deletions, Δ*ars*, Δ*arsR*, Δ*yqcK*, Δ*arsB*, and Δ*arsC*, to be used in further characterisation. Since the Δ*ase* strain was used as a starting point for all *ars* characterisation strains, the Δ*ars* strain is expected to remove any arsenite or arsenate responsiveness from the cells. The deletion of *arsR* removes the regulation of the operon, such that the remaining genes are expressed constitutively[141]. The Δ*arsB* and Δ*arsC* strains remove the ability to export arsenite or reduce arsenate to arsenite, and result in a strain which is sensitive to both arsenite and arsenate or just arsenate, as observed previously[141]. Since the role of *yqcK* is not confirmed, it is hard to predict what the effect of a deletion would be, and previous attempts have been inconclusive.

Unfortunately, after several attempts, only the Δ*arsR* deletion proved successful, with no colonies forming from any of the other transformations. Each of the plasmids was thoroughly sequence verified and found to be correct, but despite this only the Δ*arsR* variant gave rise to colonies. This transformation was used as a positive control in further attempts at generating

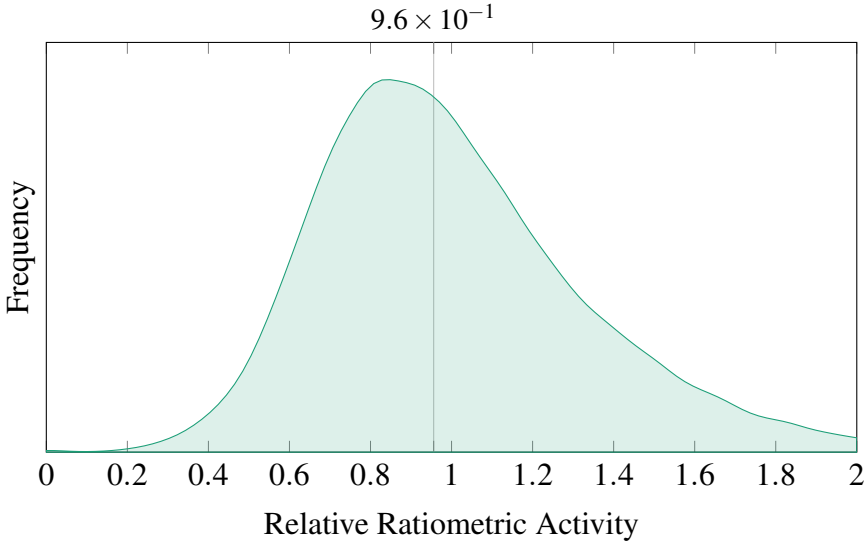



Fig. 5.3 Ratiometric strength of the $P_{ars}$ promoter relative to $P_{pen}$, showing the median strength of 0.96

the other deletions, and consistently gave of the order of 10 colonies, suggesting that the competent cells were not at fault.

Further diagnostic work is possible – for example verifying the linearisation of the plasmid by ScaI-HF, or increasing the mass of DNA used for the transformation, or even redesigning the plasmid with different homology regions. However, since the Δ*arsR* strain demonstrates constitutive expression of the operon, it is sufficient to characterise the basal strength of the $P_{ars}$ promoter, and it was decided to proceed with that characterisation before attempting to more thoroughly troubleshoot the other transformations.

Ratiometric output from the $P_{ars}$ promoter driving the mVenus/mScarlet-I based ratiometric system develiped in chapter 4 in the constitutive Δ*arsR* context and normalised against the $P_{pen}$ reference is shown in Figure 5.3. At 96% of the strength of $P_{pen}$, the promoter is remarkably strong – $P_{pen}$ was chosen as a reference precisely because of its strength.

### 5.3.2 Relative Strength of RBS

The strengths of each of the four RBSs associated with each gene of the operon were also characterised relative to the gsiB control. RBS sequences were extracted taking the entire sequence from one base after the stop codon of the previous gene until the base before the start codon of the gene in question, in the case of *arsR*, the first gene of the operon, the entire 5'UTR was taken using the transcriptional start site predicted by Nicolas et al. [112].

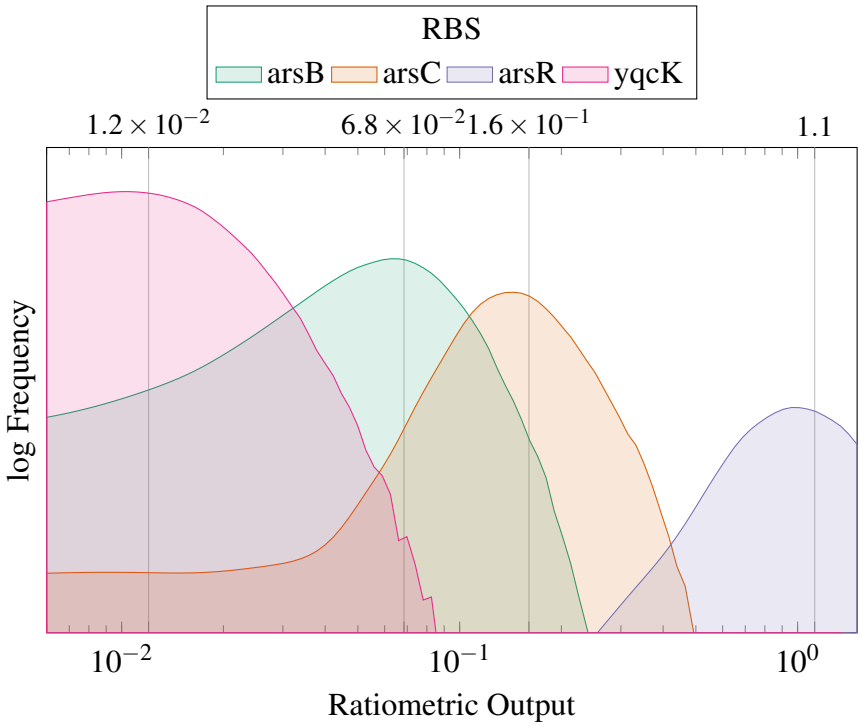



Fig. 5.4 The relative ratiometric strength of the four RBSs of the *ars* transcriptional unit, measured relative to the gsiB RBS and each under $P_{pen}$. The RBS associated with the repressor, *arsR* is the strongest by almost ten-fold, with the other two genes whose function has been determined showing similar activities.

The results of this characterisation are shown in Figure 5.4. The RBS associated with *arsR* is the strongest, and is slightly stronger than gisB. The RBSs associated with the two other genes with known function, *arsB* and *arsC* are of similar strength to each other, roughly $\frac{1}{10}^{\text{th}}$ that of *arsR* when under $P_{pen}$. The fourth RBS, that of *yqcK*, is weaker, about 100 fold less strong than *arsR* under the reference promoter.

As was seen in Section 4.4.2, the scale of these differences in terms of fold change between the RBSs will be reduced when they in a less transcriptionally active context.

### 5.3.3 Normalised Induction Curve

Induction curves were generated by growing strains in plain LB media overnight and using these to inoculate media with a range of sodium arsenite concentrations. Growth was then continued at 37 °C with shaking for around 2.5 h, until growth was just beginning to become visible to the naked eye in the wildtype control strain. The cultures were then transferred into appropriate plasticware and analysed using flow cytometry as described previously in Section 2.4.

Based on the form of the response predicted by the modelling effort from Chapter 3, a logarithmically distributed set of characterisation points seemed most appropriate as the response changes most rapidly at low arsenite concentrations and changes much more slowly as the concentration grows. This simplified the preparation of the arsenite media considerably, as an accurate stock of the highest chosen concentration could first be made, and all others could be made by serial dilution. After some initial experiments, a concentration of $2.048\,\mathrm{mg l^{-1}}$ was chosen as the maximum concentration as this appeared to be above the quasi-linear response range and as a power of 2 makes for simple two-fold dilutions. In all, twelve arsenite concentrations were investigated, $0\,\mu\mathrm{g l^{-1}}$ and $2^n\,\mu\mathrm{g l^{-1}}$ for $n \in \{1-11\}$.

Two different strains were tested, $P_{ars}$, which includes only the promoter in the query position, and $P_{ars}$-arsR, also known as $P^{+}_{ars}$, which also includes an extra copy of the arsR gene in the query locus as discussed in Section 5.2.2.

The results of these experiments are shown in Figures 5.5 and 5.6, shown relative to the basal strength of the $P_{pen}$ promoter, measured as described in Section 5.3.1. The form of the response closely matches that predicted by the mathematical modelling from Chapter 3, as shown in Figure 3.13. In particular, the response increases monotonically from an initial value, while the gradient decreases monotonically towards zero from an initial maximum.

The error bars shown in the figures refer to the 95% confidence interval in the ratiometric output of cells which passed gating and reference thresholding, and represent a large amount of noise in the observed value. Individuals within the population display a wide range of fluorescent output around the median, which is shown marked. The population behaves much more predictably when taken together, as evidenced by the fact that the median values are in such close alignment with one another.

The introduction of the second copy of *arsR* has a considerable effect on the output of the system, reducing the relative level of induction in the absence of arsenite by almost 3 orders of magnitude from around $10^{-1}$ to around $2 \times 10^{-4}$. The introduction of the extra copy of the *arsR* binding domain, which doubles the number of binding domains present, has therefore had a significant effect on the dynamic equilibrium between bound and unbound $P_{pen}$, suggesting that the number of copies of free $ArsR_2$ is typically very low. We can now answer the question posed by Equation 5.1, as to whether $k^{+}_{DNA}$ or $[ArsR_2]_{max}$ dominates – clearly a very small amount of $ArsR_2$ is able to have a very strong repressive effect on the promoter, implying that the association constant between it and DNA, $k^{+}_{DNA}$ must indeed be very large.

At the highest concentration of arsenite investigated, the fold difference between the two versions has decreased to under 10, suggesting that the magnitude of the effect caused by the extra copy decreases as arsenite concentration increases, as predicted in Figure 5.2. The point

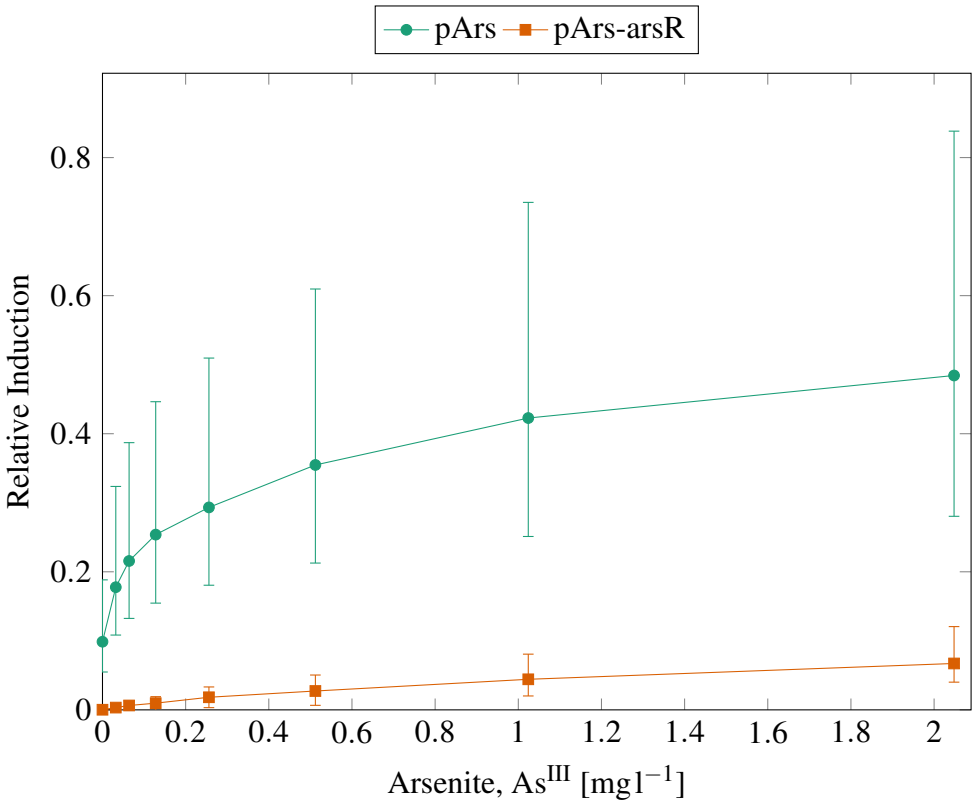



Fig. 5.5 Induction of the arsenic operon by a range of arsenite concentrations, showing both the $P_{ars}$ and $P_{ars}$-*arsR* versions of the system from no induction to near full induction (2 mg l$^{-1}$ As$^{III}$). Each query system was inserted into the ratiometric system developed in chapter 4, using the Δ*ase* strain as a starting point. As predicted, the rate of induction grows quickest at low levels of induction, and there is a considerable difference in the level of the response between the two systems, implying that the addition of the extra copy of *arsR* has a significant effect on the output of the system. An SBOL diagram of the two systems is available in figure 5.1.

at which the two converge is unknown – and possibly occurs at an arsenite concentration too high for normal growth – but the fact that this convergence is slow is consistent with a high value of $k^{+}_{DNA}$.

As discussed before, the two curves shown in Figure 5.5 represent the upper and lower bounds for the wildtype system, and although the high strength of the promoter suggests that the wildtype system will tend towards the upper bound, it is impossible to say this with certainty. However, the ultimate purpose of this research is to understand the behaviour of a biosensor based on the *ars* operon. Figure 5.6 shows the behaviour of the system under lower arsenite concentrations, including those relevant to the Bangladeshi and WHO limits

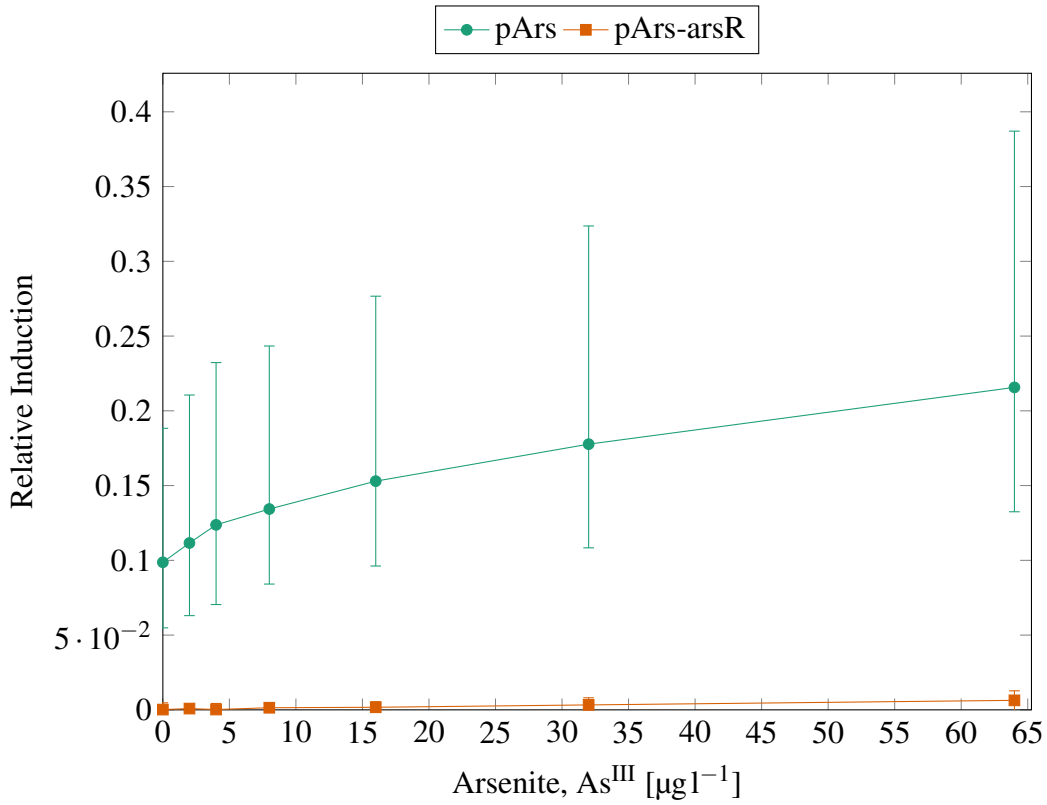



Fig. 5.6 Induction of the arsenic operon by arsenite from zero to 64 µg l$^{-1}$, all other information equal to figure 5.5.

on arsenite concentration in water sources, which are 50 mg l$^{-1}$ and 10 mg l$^{-1}$ respectively, as discussed in Section 1.1.3.

Encouragingly, both the $P_{ars}$ and $P_{ars}$-arsR variants show measurable differences between arsenite free media and media containing 10 mg l$^{-1}$ and 50 mg l$^{-1}$, which signifies that the *ars* operon is able to distinguish between these conditions, however neither of these systems are suitable to drive a biosensor circuit in the wild. In the case of $P_{ars}$-arsR, the problem is clear – the magnitude of the increase caused by even 50 mg l$^{-1}$ is simply not enough to be read reliably. This is also the case for the $P_{ars}$ system, where 10 mg l$^{-1}$ causes an increase in expression of between 4 and 5% of the de-repressed strength, which rises to almost 10% for 50 mg l$^{-1}$. This can be observed easily using precise equipment such as a flow cytometer, but is harder to determine in the field for example when observing only the colour of a colony of cells.

The question of how to build an effective arsenite sensor which is sensitive enough for use in the field is thus how to amplify this change such that the difference in output between the 0 and 10 mg l$^{-1}$ is easily observable.

## 5.4 Conclusions and Further Work

The goal of this characterisation work was to improve our understanding of the native operon, as well as to validate the mathematical modelling that was done previously in Chapter 3. The predictions made by the mathematical modelling as to the form of the response have been borne out by experiment as the form shown in Figures 5.5 and 5.6 closely corresponds to that predicted by modelling.

We have also learnt something about the parameters of the model. The *ars* operon promoter, $P_{ars}$, is a very strongly active promoter which is very tightly repressed by even a small number of $ArsR_2$ repressor proteins. This tight repression is somewhat relieved by $As^{III}$, but does not reach more than around 50% of the maximum possible level of induction – when $ArsR_2$ no longer represses $P_{ars}$ – for arsenite concentrations below $2\,mgl^{-1}$. This contrasts with the observations made in Sato and Kobayashi [141] who originally reported on the *ars* operon in *B. subtilis*, and believed that maximal induction occurs with between 1 and $10\,mmoll^{-1}$ (between 75 and $750\,mgl^{-1}$). Here we have shown that the level of induction actually continues to grow with arsenite concentration, albeit very slowly at higher concentrations, this is likely due to the higher precision offered by ratiometric flow cytometry as compared with the $\beta$-galactosidase method employed by Sato and Kobayashi [141].

### 5.4.1 Measuring Arsenite Extrusion

While this characterisation work has been informative in validating the previous modelling effort and in learning something of the scale of the response and the internal workings of the system, we have not learnt a great deal about the central balance affecting the *ars* operon – that between the internal and external arsenite concentrations. As was seen during Chapter 3, this balance is key to setting the scale and sensitivity of the response, since although we wish to measure the external arsenic concentration, it is the internal concentration which is visible to the operon.

The internal arsenite concentration is set by a balance of forces, on one side the entry of arsenite into the cell, either through the glycerol uptake mechanism as in *E. coli* (see Section 3.1.1) or otherwise, and on the other the joint forces of growth dilution and extrusion through the ArsB membrane protein.

Considerable time and effort was invested into generating single gene deletions from the *ars* operon which were so far unsuccessful, as discussed in Section 5.3.1. Of particular interest in this context would be an *arsB* deletion, which would remove extrusion from the balance of forces acting on internal arsenite. If the rate of extrusion is large compared with the rate of growth dilution, then this deletion will shift the balance and increase the internal

arsenite concentration, in turn increasing the level of induction observed. However, if the rate of extrusion is small (as would possibly be the case if extrusion is mainly relevant in stationary phase), then the deletion of *arsB* would have little effect on the relative induction level until stationary phase is reached.

Careful modelling of this effect may allow us to say something as to the relative scales of dilution and extrusion as a means of arsenite mitigation, and this would also tell us something about the rate at which arsenite enters the cell. Returning to the parameters introduced in Chapter 3 (see Figure 3.12), the steady state internal arsenite concentration for the Δ*arsB* mutant is given by

$$[As_i] = \frac{1}{\mu} \cdot \frac{U_{max} \cdot [As_e]}{U_m + [As_e]}$$

where $\mu$ is the growth rate, $U_{max}$ and $U_m$ are the Michaelis-Menten parameters for arsenite uptake, and $[As_i]$ and $[As_e]$ are the internal and external arsenite concentrations. For growth dilution to be a significant cause of a reduced internal arsenite concentration, as would be evidenced by only a modest increase in induction following an *arsB* deletion, the rate at which arsenite enters the cell must be small compared with the rate of growth dilution.

A further experiment would be to reintroduce ArsB under a constitutive promoter – for example using a selection from the promoter and RBS combinations characterised in Chapter 4. The internal arsenite concentration is then given by the solution to

$$\frac{U_{max} \cdot [As_e]}{U_m + [As_e]} = \mu \cdot [As_i] + \frac{E_{cat} \cdot [ArsB] \cdot [As_i]}{E_m + [As_i]}$$

where $[ArsB]$ is the concentration of ArsB and $E_{cat}$ and $E_m$ are the catalytic Michaelis-menten parameters for ArsB mediated extrusion. By gathering data from a range of $[As_e]$ and $[ArsB]$, estimates of the value of the parameters $U_{max}$, $U_m$, $E_{cat}$, $E_m$ could be made.

### 5.4.2 Testing Induction by Arsenate

In the original paper on the *ars* operon in *B. subtilis*, Sato and Kobayashi [141] found that around 10-fold more arsenate was required to produce a similar level of induction as with arsenite. This is not a great surprise as arsenate is removed from the cell by first being reduced to arsenite by *arsC* and expelled from the cell as normal (see Section 3.1.1). The lower level of induction is likely caused by the fact that it takes time for the level of arsenite to build up once arsenate has entered the cell – eventually one would expect the induction to reach the same level once all the in the sample arsenate has been converted to arsenite.

The detection of arsenate is an important problem which must be solved in order for the biosensor to be reliably used in the field as both forms or arsenic are harmful. The most

reliable field tests currently available reduce the $As^{V}$ to $As^{III}$ chemically before continuing with the test, however this is not particularly attractive as the added complexity of operation is one of the key things the biosensor design seeks to avoid.

Two other options are available – somehow tune the sensor such that arsenate sensitivity is the same as arsenite sensitivity, either by over-expressing the reductase or by increasing the rate at which arsenate enters the cell by characterising and up-regulating whichever membrane proteins are responsible for its entry, currently believed to be related to the phosphate transport system.

The other solution is to develop two sensors, one responsive to both $As^{III}$ and $As^{V}$ and the other only responsive to $As^{III}$ due to deletion of the *arsC* reductase. Having quantified the concentration of arsenite using the second version, this output can be used to calculate the amount of output from the first sensor which is due to $As^{III}$, and use the remainder to calculate the $As^{V}$ concentration. While this system could be used to specifically distinguish between the two oxidation states, the non-linear nature of the reporter would make it complex to characterise and use, and must necessarily be used to quantify the exact arsenic concentration, rather than directly answering the question "is this sample above or below the permissible threshold for arsenic contamination?"

It seems likely that in the short term at least, a more reliable first generation biosensor would be better off by converting all the available arsenate into arsenite before testing.

### 5.4.3 Investigating Other Potential Inducers and Response Modulators

Arsenite and arsenate are not the only two factors capable of causing de-repression of $P_{ars}$, leading to the possibility of false positive readings. For example, antimonite is also known to interact with $ArsR_2$, causing some degree of de-repression[141]. This leads to the possibility of antimonite being incorrectly recorded as arsnite. Antimonite also affects the rate at which ArsB-mediated extrusion of $As^{III}$ occurs[84, 100], which causes a change in the balance between internal and external arsenite. Therefore, the presence of a small amount of antimonite in a sample has the potential to significantly affect the output of an arsenic biosensor. This effect needs to be studied in detail before a reliable sensor can be constructed, and more data on the prevalence of antimonite in water samples from affected tube wells needs to be collected.

Glycerol is another factor which may modulate the system's response to arsenite. Since arsenite is believed to enter into the cell through the glycerol uptake pathway, it is possible that growth on glycerol media may affect the rate at which arsenite enters the cell, and again

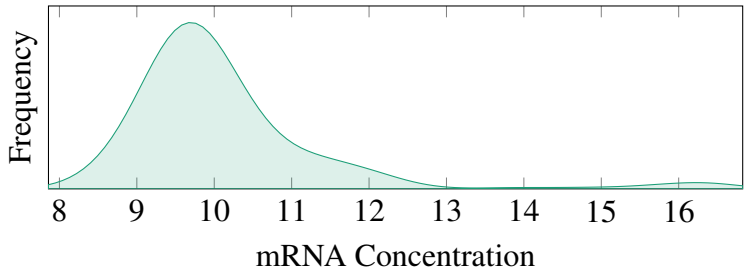



Fig. 5.7 Distribution of mean mRNA concentration observed for the four *ars* genes

change the balance between internal and external arsenite. However, initial experiments did not indicate a significant difference between arsenite sensitivity of cells grown in LB media and minimal growth media with glucose replaced with glycerol.

We can again turn to the mRNA concentration data collected by Nicolas et al. [112] to find other potential inducers of the *ars* operon. Figure 5.7 shows the distribution of recorded mRNA concentrations for *arsR*, *yqcK*, *arsB*, and *arsC* for each of the 104 conditions investigated in the study. In the majority of conditions, the operon is very tightly repressed and shows concentrations at the very low end of reported concentrations from the study. However, the distribution appears to have an appreciable shoulder on the right hand side, and a few conditions appear very highly active, implying that they do indeed cause induction of the operon.

Table 5.1 shows the most significant p-values for t-tests conducted for each condition tested by Nicolas et al. [112]. The top three conditions involve incubation with diamide, a reagent which causes the oxidation of thiols[80, 81]. The p-value of each of the 9 diamide related conditions taken together is $2.5 \times 10^{-52}$, indicating that diamide does indeed cause the de-repression of the $P_{ars}$ promoter. This could be due to diamide induced oxidation of the cysteine thiol residues of $ArsR_2$ which normally bind with $As^{III}$ and trigger the release of DNA[175]. Although this effect appears to be very strong, causing what appears to be complete de-repression of the operon, the chances of encountering diamide in a natural water sample are very small and none of the other oxidative reagents tested caused de-repression of the operon. This reaction is therefore unlikely to be of concern in the context of a biosensor, but could prove useful in further research as a way of causing full de-repression of the operon.

Other conditions which appear to demonstrate induction of the *ars* operon include phosphate limitation, sporulation, and ethanol stress. These conditions are again unlikely to be brought about unexpectedly during testing of a water sample, however further study of the mechanism or mechanisms by which they appear to induce the operon would perhaps shed light on other potential sources of false positives.

| Condition | Repeats | p-value | Description |
|---|---|---|---|
| dia5 | 3 | $5.2 \times 10^{-19}$ | 0.5 mM Diamine for 5 minutes |
| Diami | 3 | $7.3 \times 10^{-17}$ | 0.6 mM Diamide for 10 minutes |
| dia15 | 3 | $3.4 \times 10^{-10}$ | 0.5 mM Diamide for 15 minutes |
| LPhT | 3 | $4.9 \times 10^{-3}$ | 3 h after outset of the phosphate-limitation induced stationary phase |
| S4 | 3 | $1.4 \times 10^{-2}$ | 4 h after resuspension in sporulation media |
| S5 | 3 | $1.4 \times 10^{-2}$ | 5 h after resuspension in sporulation media |
| S6 | 3 | $1.8 \times 10^{-2}$ | 6 h after resuspension in sporulation media |
| S3 | 3 | $2.1 \times 10^{-2}$ | 3 h after resuspension in sporulation media |
| Etha | 3 | $3.8 \times 10^{-2}$ | 10 min after 4% (v/v) Ethanol stress |

Table 5.1 P-values below 0.05 for induction of *ars* operon in each condition

### 5.4.4 Developing an Arsenic Biosensor

This work has demonstrated that there is a small but detectable difference in the output of the *ars* detoxification system between zero arsenite and the WHO and Bangladeshi standards of 10 and 50 $\mathrm{mg\,l^{-1}}$ respectively. While the small size of the difference shown by the natural system is unlikely to be large enough to be useful in the field on its own, it demonstrates the potential of the technology in the detection of arsenic.

The initial characterisation of the *ars* operon discussed in this chapter is consistent with the modelling predictions made in Chapter 3, which allows for some confidence in the model.

The question of how best to increase the dynamic range of the naturally occurring system in order to accurately detect whether arsenic concentration is above or below a given tolerable threshold remains to be answered. In addition, there are several other technical, regulatory, and social issues which must be overcome in order for a viable arsenic biosensor to be successful. While progress has been made on the shipping and containment of the *B. subtilis* spores, further evidence is required to understand the effects of other potential contaminates in the water samples, and regulatory approval is required for the use of a GMO in the field. Furthermore, education initiatives are required on the ground in areas such as Bangladesh and Nepall, and local communities will need to be supported in taking control of the regular testing of their water sources.

# Chapter 6

# Codon Optimisation for *B. subtilis* and Other Organisms

## 6.1 Introduction

There are 64 possible permutations of the nucleotide triplets which make up the genetic code, but these 64 codons encode for only 20 natural amino acids and three stop codons which terminate translation. The mapping between the sequence of codons making up a coding sequence and the sequence of amino acids present in the encoded protein is therefore necessarily degenerate, meaning information is lost in the process of translation as many possible sequences of codons encode the same amino acid. For most encoded amino acids there are between two and six possible codons, each of which will cause the insertion of that particular residue into the polypeptide chain.

As has long been known, the choice of which codon to use for each amino acid is not uniform, with certain codons occurring significantly more frequently than others. Which particular codons are popular or unpopular within a genome varies with different organisms, but are generally similar within genes from a particular genome. This observation led rapidly to the genome hypothesis of codon bias[56], that codon preference is specified at the genome level rather than the level of individual genes.

The 'mutationist' explanation of codon bias is highly consistent with this viewpoint, as it posits that codon bias is driven by the fact that the probability of the introduction of a mutation during genome duplication is not uniform and depends on the sequence being duplicated. In particular, certain codons appear to be more mutable than others, and so codon usage tends to drift towards the less mutable sequences. Multiples studies[27, 79] have shown that the GC content of the entire genome has the biggest effect on changes in

codon bias between different genomes. This confirms that whole genome processes do affect the choice of codons within coding sequences, and do not act solely on intergenic space.

This hypothesis does not help explain the observations that codon bias is correlated with both the tRNA content within the organism[71] and the translation rate of the resulting messenger RNA[53]. Both of these observations suggest that selective pressure is also involved in codon choice, with genes which are highly transcribed having been better optimised for the tRNA context experienced in their host organism. In addition, the choice of codon affects the accuracy of the included amino acid[3, 105], with the pattern of codon choice varying when the resultant amino acid is essential for protein function.

Inserting recombinant genes from one organism into another is a common activity in synthetic biology. In this project, the fluorescent reporter proteins used to develop the reporter systems in Chapter 4 were taken from a variety of contexts and inserted into *B. subtilis* after optimisation.

It is therefore important to understand the specific codon bias present in *B. subtilis* in order to reliably insert recombinant or synthetic genes into the organism and have them be translated efficiently. While there is much commercial interest in codon optimisation strategies which can optimise the translational efficiency and thus yield of an arbitrary gene product, these techniques are often proprietary and rely on large numbers of comparative experiments in which the possible sequence space of the amino acid is explored.

While such methods do well in finding an optimal codon strategy, the large amount of synthesis required makes them expensive and time consuming. In many cases, particularly within academia, it is not vital that the recombinant gene is translated optimally, but merely that it is expressed reasonably well and consistently. It is therefore useful to study codon optimisation with a view to building simple *in silico* methods for optimising codon selection to generate gene sequences which are likely to be well expressed.

Such a method would not have access to the wealth of statistical information that is generated by a high-throughput empirical screening method such as those used commercially. Without such information, it is currently impossible to accurately account for the broader effects of codon choice such as secondary structure, interactions with other cellular RNAs, or the complex interactions at translation initiation and termination. Despite this limitations, it is still useful for synthetic biology as a whole to develop a better understanding of how codon choice influences gene expression in native genes in order to apply that knowledge when designing synthetic systems such as the fluorescent reporter proteins used in Chapter 4.

| 1st Base | 2nd Base T | | | | 2nd Base C | | | | 2nd Base A | | | | 2nd Base G | | | | 3rd Base |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Δ[c] | Codon | %[b] | AA[a] | Δ | Codon | % | AA | Δ | Codon | % | AA | Δ | Codon | % | AA | |
| T | | TTT | 68.50 | F | | TCT | 20.50 | S | | TAT | 65.41 | Y | | TGT | 45.33 | C | T |
| | | TTC | 31.50 | F | | TCC | 12.70 | S | | TAC | 34.59 | Y | | TGC | 54.67 | C | C |
| | | TTA | 19.83 | L | | TCA | 23.62 | S | | TAA | 63.19 | * | | TGA | 22.73 | * | A |
| | | TTG | 15.90 | L | | TCG | 10.03 | S | | TAG | 14.08 | * | | TGG | 100.00 | W | G |
| C | | CTT | 23.93 | L | | CCT | 28.64 | P | | CAT | 67.35 | H | | CGT | 18.19 | R | T |
| | | CTC | 11.23 | L | | CCC | 8.89 | P | | CAC | 32.65 | H | | CGC | 20.68 | R | C |
| | | CTA | 5.09 | L | | CCA | 19.06 | P | | CAA | 51.28 | Q | | CGA | 9.86 | R | A |
| | | CTG | 24.01 | L | | CCG | 43.41 | P | | CAG | 48.72 | Q | | CGG | 15.61 | R | G |
| A | | ATT | 50.47 | I | | ACT | 16.10 | T | | AAT | 56.52 | N | | AGT | 10.59 | S | T |
| | | ATC | 36.72 | I | | ACC | 15.89 | T | | AAC | 43.48 | N | | AGC | 22.55 | S | C |
| | | ATA | 12.81 | I | | ACA | 41.20 | T | | AAA | 70.17 | K | | AGA | 26.28 | R | A |
| | | ATG | 100.00 | M | | ACG | 26.81 | T | | AAG | 29.83 | K | | AGG | 9.38 | R | G |
| G | | GTT | 28.41 | V | | GCT | 24.73 | A | | GAT | 64.00 | D | | GGT | 18.38 | G | T |
| | | GTC | 25.63 | V | | GCC | 20.72 | A | | GAC | 36.00 | D | | GGC | 33.94 | G | C |
| | | GTA | 19.75 | V | | GCA | 28.25 | A | | GAA | 67.99 | E | | GGA | 31.46 | G | A |
| | | GTG | 26.22 | V | | GCG | 26.30 | A | | GAG | 32.01 | E | | GGG | 16.22 | G | G |

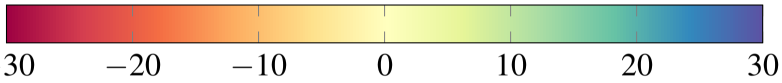



Table 6.1 Codon preference table for *B. subtilis* 168. (*a*) The Amino Acid (AA) encoded by the codon, (*b*) percentage of amino acid residues encoded by the given codon, (*c*) deviation from expected percentage given uniform codon preferences.

## 6.2 Variations in Codon Preference

The codon preference of an entire genome can be easily calculated by extracting the DNA sequence of each annotated gene and recording the number of occurrences of each codon. This is then normalised by dividing by the total number of occurrences of the amino acid for which each codon encodes to give the fraction of instances in which each codon is chosen over the other homologous codons to encode the amino acid. Assuming uniform codon preference, the value for the $i^{\text{th}}$ codon would be expected to be $\frac{1}{M(i)}$ where $M(i)$ is the box number (i.e. the number of codons which encode the same amino acid) of the codon.

The codon preference table for *B. subtilis* is presented in Table 6.1, which shows the percentage of times each codon is selected and indicates via the coloured boxes whether each codon is more or less frequently observed than expected given the amino acid content. From the table, one can easily see that some amino acids are very close to the expected distribution, for example glutamine (Q) is encoded by `CAA` and `CAG` with very close to 50% frequency for each. However, other amino acids demonstrate a strong skew in favour of particular codons, for example both lysine (K) and glutamic acid (E) are significantly more likely – by a ratio of almost 7:3 – to include an `A` in the third base wobble position as opposed to a `G`. This is interesting because if codon preference were driven solely by whole-genome effects such as GC content, one would expect the apparent pressure to prefer `AAA` over `AAG` in the case of lysine or `GAA` over `GAG` in the case of glutamic acid to be mirrored by a similar pressure in favour of `CAA` over `CAG` in the case of glutamine, but this is not observed.

In fact, codon preference is not the same for every gene in a genome, and codon preference varies appreciably between different genes. This was first observed in *B. subtilis* by Kanaya et al. [76] and reviewed in Moszer et al. [105], who represented the codon preference of each gene as a vector $\boldsymbol{x}$ in 61-dimensional space, encoding the normalised usage probability for each codon. Since 61-dimensional space is not particularly useful, dimensional reduction was carried out using Principal Component Analysis (PCA). PCA reduces the number of dimensions by finding the 61-dimensional vector $\boldsymbol{b}$ along which the maximum amount of variance in the data can be explained, equivalent to finding the 'line of best fit' for two dimensional data. This explained variance is subtracted from the sample and the process is repeated until the desired number of dimensions have been found, resulting in a list of PCA components, $\{\boldsymbol{b}_0, \boldsymbol{b}_1, \dots\}$, which provide our new basis. Taking the dot product of these vectors with the original data gives the PCA scores, $\{z_0, z_1, \dots\}$, for each gene. Since the dimensions of the input dataset are not wholly independent – increasing the probability of choosing one codon necessarily decreases the probability of choosing another – we can be certain that our input data is not truly 61-dimensional, but can actually be represented accurately in far fewer dimensions.

PCA is very sensitive to normalisation, and input data must be well normalised for PCA to be effective. The normalisation scheme proposed by Kanaya et al. [76] is

$$x_i = \frac{f_i}{\frac{1}{M(i)} \sum_{\forall j \in \boldsymbol{m}(i)} f_j} \tag{6.1}$$

where $f_i$ is the number of occurrences of the $i^{\text{th}}$ codon and $\boldsymbol{m}(i)$ is the set of all codons which encode the same amino acid as the $i^{\text{th}}$ codon, which is of size $M(i)$. This normalisation removes effects of gene length, amino acid content, and codon box number (where codons encoding for amino acids that are encoded by several codons are more rare than ones which are encoded by fewer codons), such that the expected value for each dimension of the resulting vector, $\boldsymbol{x}$, is 1 under uniform codon preference.

This normalisation is re-used here, with one important improvement. 42% of all known *B. subtilis* genes contain no incidences of at least one amino acid, and all 4,255 of them contain 1 or fewer instances of at least one amino acid. When a gene contains only one instance of an amino acid, the frequentist approach outlined by Kanaya et al. [76] will assign a probability of 1 to the codon used on that occasion, and zero to all other codons which could encode that amino acid. Since effects of amino acid content are removed through normalisation, this gene will then appear to have a very extreme codon bias in the case of that amino acid, despite that estimate being based on only one single observation. This is not just a problem for amino acids which are only encoded by one or two codons as 32% of genes contain six or fewer instances of one of the three amino acids which are encoded by six codons, arginine, leucine and serine. This is clearly not enough samples to give a confident estimate of the true codon preference of the gene, and is likely to lead to extreme estimates of codon bias based on little evidence.

This problem can be avoided by taking a Bayesian approach to estimating the codon usage probabilities for each gene. In Bayesian statistics, a prior probability is assigned for each outcome, based on our experience or knowledge of the system we are modelling. When observations are made, the prior probability is updated to reflect the new information, to give the posterior probability distribution which represents all our knowledge to date. When the number of observations is small, the posterior closely resembles the prior distribution as the observations don't have enough statistical power to change our belief about the underlying system. However, with more and more observations the posterior begins to resemble the frequentist solution as our belief about the system changes.

To model codon bias using a Bayesian approach, two decisions must be made

1. what should the prior distribution be?

2. how many observations are required before we change our initial belief about the underlying codon distribution?

For simplicity, rather than considering the entire distribution at once, let us consider the case of a single amino acid, for example lysine (K), which is encoded by the codons $\{AAA, AAG\}$. Without making any observations of the particular gene in question, our prior expectation of the distribution of lysine codons in that gene is equal to the distribution found throughout the whole genome, approximately $\{0.7, 0.3\}$ as shown in Table 6.1. This is a categorical Dirichlet distribution, which can very conveniently be used as a Bayesian prior using *pseudocounts*. First set the number of pseudocounts, for example $N = 5$, and initialise the counts for each codon with this value multiplied by the prior probability, in this case $\{3.5, 1.5\}$. Now add to this the observed number of each codon in the gene and divide by the total to give the posterior distribution, or apply the normalisation given in Equation 6.1 to proceed with PCA analysis. Supposing the gene contains only a single lysine and that it happens to be encoded by the rarer *AAG* codon, our Bayesian model predicts an underlying distribution of $\{0.58, 0.42\}$, which is quite close to the prior and much more reasonable than the extreme $\{0, 1\}$ estimate which would be made by the frequentist approach. However suppose five lysines are present, and every one of them is encoded by the rarer codon. Our model now predicts a distribution of $\{0.35, 0.65\}$, which is quite different to the prior, representing the fact that we now have more evidence that the true distribution of lysine codons for this gene is significantly different from the prior than in the previous example.

This question remains as to what value should be set for $N$, the number of pseudocounts. If $N$ is too small, then we return to the previous problem where a small number of observations completely change the estimated distribution, while if $N$ is too large our estimate of per-gene codon bias will never stray far from the codon bias of the whole genome, no matter how many observations we make. Ultimately, the choice for $N$ is an arbitrary trade off between these extremes, but we can use the idea of *explained variance* to help guide the choice. The explained variance of a PCA component vector is the percentage of the total variance which occurs along that vector. Since PCA sequentially finds the vectors which explain the maximum amount of variance – hence the term principal component – the variance explained by successive components must be less than or equal to the previous vector.

Figure 6.1 shows the percentage of explained variance for the first five principal components for *B. subtilis* codon usage data for a range of pseudocounts, $N$, normalised using Equation 6.1. As the value of $N$ increases, the amount of random variation in the dataset caused by sampling noise reduces, and so we expect the proportion of variation due to real changes in codon bias to increase. Thus, the percentage of variance explained by principal components which represent true underlying variation in the data should increase, while the

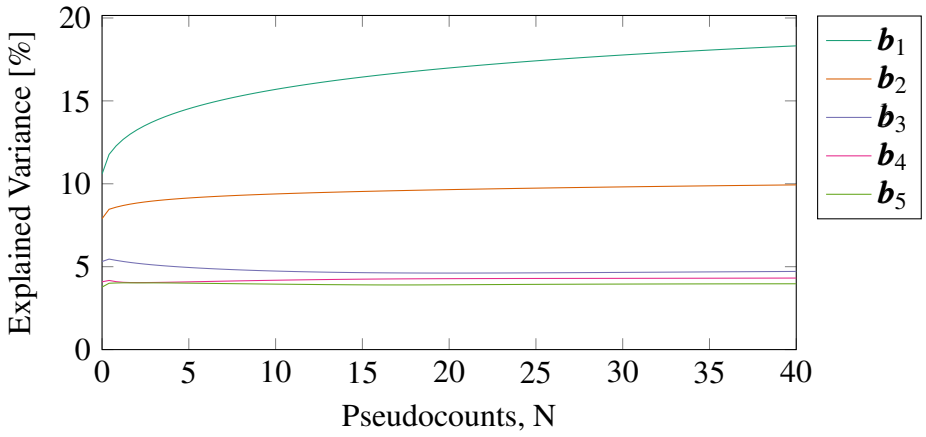



Fig. 6.1 Percentage of explained variance of the first five PCA component vectors for a range of values of $N$

percentage variance explained by principal components which only represent noise in the data should decrease towards some minimal value. This is reflected in Figure 6.1, where the first two principal components progressively explain a greater fraction of the total variance as $N$ increases, while the variance explained by the third component decreases slightly. The proportion explained by the fourth and fifth components does not appear to vary significantly with $N$.

This suggests that only the first two principal components are relevant to the underlying variation in per-gene codon usage, a theory which is borne out by looking at the distribution of the z-scores (dot products) for each principal component, shown in figure 6.2. The distribution of the z-scores for the first two principal components clearly contain some form of bimodal structure, while the z-scores of principal components of higher order than 2 are distributed normally implying that they are the result of noise alone.

A scatter plot of the z-scores of the first two principal components is shown in Figure 6.3, which closely resembles the 'rabbit head' structure found in Figure 1 of Moszer et al. [105]. The data has been clustered according to a Gaussian Mixture Model (GMM) with 3 groups and optimised using the standard Expectation Maximisation (EM). After some experimentation, it was found that a value of $N = 7$ provided a good trade off between increasing the proportion of explained variance (Figure 6.1) and stable clustering properties, and so this value was used throughout the rest of this section.

Moszer et al. [105] posited that these three categories related to three different classes of genes present in the *B. subtilis* genome. The majority of genes belong, predictably, to a 'general' category whose codon preference is closely aligned to that of the genome as a whole. This category, according to Moszer et al. [105], mostly contains genes of 'intermediary

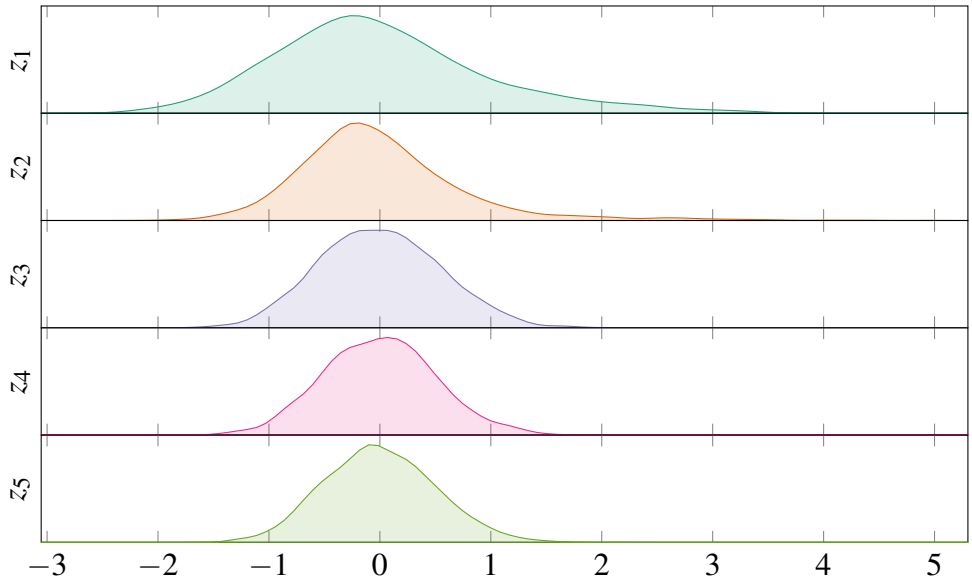



Fig. 6.2 Distribution of PCA scores for the first five components and 7 pseudocounts (i.e. $N = 7$). Both $z_1$ and $z_2$ have larger variance than the following distributions, as well as having a positive skew, that is, a larger tail on the right hand side. These observations support the hypothesis that only $\mathbf{b}_1$ and $\mathbf{b}_2$ capture true shifts in the underlying codon distrubution while $\mathbf{b}_3$ and above represent noise.

metabolism', as well as genes involved in core carbon assimilation. The second category contains genes which are highly active under exponential growth, while the third category contains genes which are less actively expressed and appear to cluster together in the genome, leading to the hypothesis that these genes are of foreign origin, labelled as *phage* in figure 6.3.

The first five principal components are shown in Table 6.2. While no component is completely characterised by a single shift in codon usage, we can see that the largest values in $\boldsymbol{b}_1$ correspond to a shift away from $CTG$ and towards $TTA$ when encoding leucine (L), and thus genes with a large $Z_1$ will be more likely to use $TTA$ than the background genetic content. The biggest shift shown in $\boldsymbol{b}_2$ is away from $CGG$, $AGG$ and particularly $AGA$ and towards $CGC$ and $CGT$ when encoding arginine. Genes with a large $Z_2$ – which are likely to be classified in the 'active' category of genes – thus favour $CGC$ and $CGT$ to encode arginine when compared to other genes.

Interestingly, both of these shifts represent not just changes in codon usage, but also changes in which tRNA recognises the preferred codon. One possible explaination is that this difference in tRNA preference between the different classes of genes serves to somewhat isolate genes of each class from each other. For example, tRNA depletion caused by high expression of a gene in the 'general' class will have less of an effect on a gene from the 'active' class, as the tRNAs that genes of these classes tend to use are different. A possible

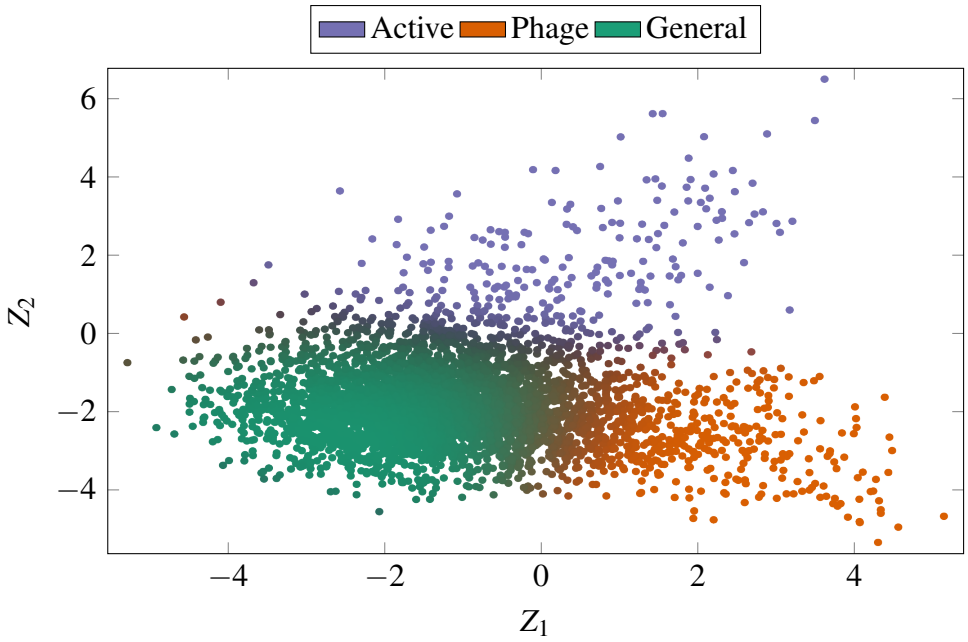



Fig. 6.3 GMM clustering of principal component scores of each gene's codon usage. The model was fitted using expectation maximisation (EM), and the number of components was set to 3 based on literature. Slightly over 3,000 genes are classified as belonging to the "general" category, while $\sim$ 860 and $\sim$ 360 are categorised as "phage" and "active" respectively.

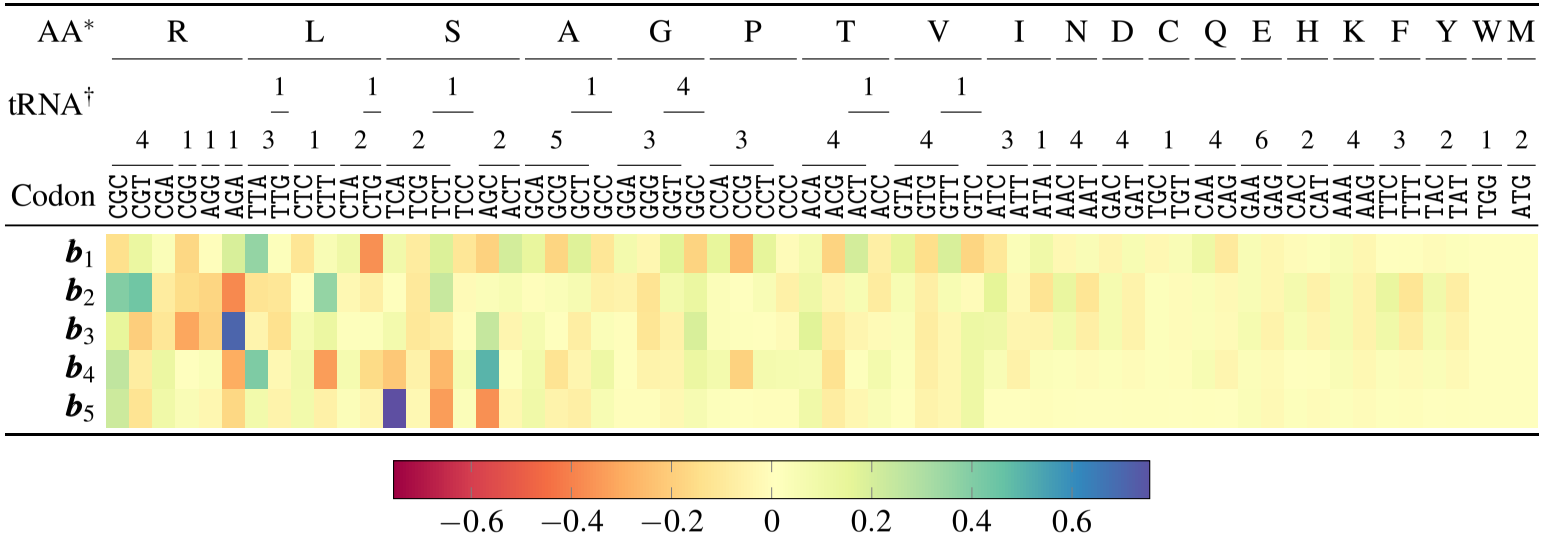



Table 6.2 The first five principal components of *B. subtilis* codon usage data. (*) Amino Acid (†) Genome copy number of each tRNA, underlining indicates recognised codons, data from Kanaya et al. [76]

counterhypothesis is that the higher dependence of genes in the 'active' class on particular tRNAs mean that modulating the avilability of those tRNAs particularly affects expression of genes within that class. While the former does seem more probable since the tRNAs preferred by 'active' group genes are not used exclusively by those genes, the data as currently available do not confirm any particular explaination, and it is possible that the truth is in fact some combination of them.

The hypothesis that genes within the 'active' group are highly expressed was made by Moszer et al. [105] based on the relatively little knowledge of gene function or of expression data available at the time. Since then, however, Nicolas et al. [112] has provided a wealth of information on the transcriptional landscape across the whole of the *B. subtilis* genome under a variety of conditions by using a tiling array technique to measure RNA concentrations across the genome, as introduced briefly in Section 4.3.4.

Figure 6.4 combines the new transcriptional data with gene codon usage information, by showing transcriptional activity in a variety of conditions overlaid onto the $Z_1$ and $Z_2$ scores presented in Figure 6.3. While some genes within the 'active' group are not highly transcribed, many of the genes from that class are very transcriptionally active, to levels which are quite rare among genes in the general group, which supports the hypothesis that codon bias found in members of the group bears some relation to their being highly expressed.

It is noteworthy, though perhaps not surprising, that this correspondence between codon bias and transcription activity exists in the first place, since codon bias is typically thought to affect the translational efficiency of genes rather than their transcription. The relationship between translational efficiency and transcription activity is expected because it would be metabolically inefficient to spend energy generating large numbers of transcripts of a gene which has poor translational efficiency. While this may happen occasionally in individual cases, in general the translational efficiency of highly transcribed genes is expected to increase over time.

## 6.3 First Order Codon Preference

Codon choice is not only influenced by large-scale factors such as whole genome codon bias or even the codon bias of individual genes, but also on the context in which a codon is used[9]. Codon preference is known to be different towards the start of genes than elsewhere, where typically codons with lower GC content are preferred[105], possibly to reduce secondary structure around the start codon leading to easier translation initiation.

In addition, the distribution of codons depends on the codons surrounding it. Codon bias is typically presented as the probability of choosing a particular codon given the amino

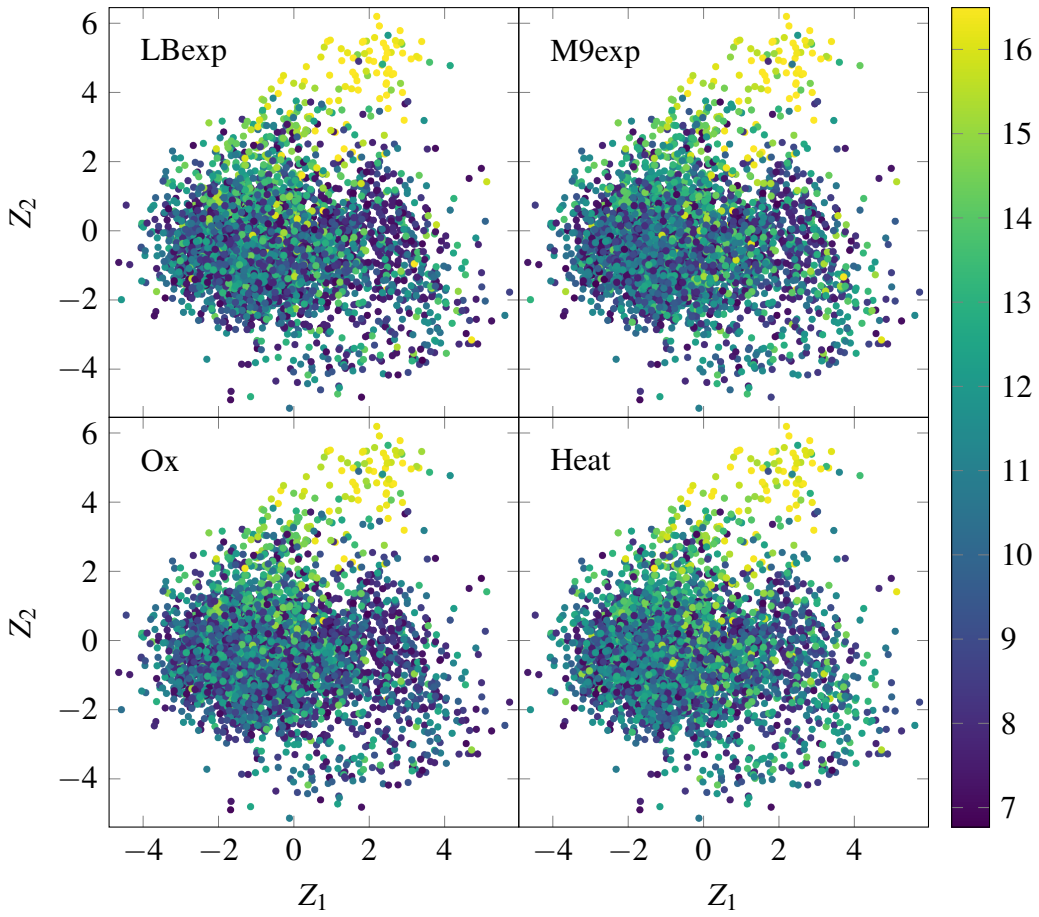



Fig. 6.4 Pincipal component scores overlaid with transcriptional data from Nicolas et al. [112] under four different test conditions. "LBexp" and "M9exp" refer to exponential growth in LB and M9 minimal media, respectively, while "Ox" refers to cells exposed to oxidative stresses for 10 minutes (induced independently with diamide, paraquat, and $H_2O_2$) and "Heat" refers to cells after undergoing rapid temperature up-shift to 48 °C. In each of the plots, the genes predicted to be members of the 'active' group in figure 6.3 generally recorded higher levels of activity than those belonging to either of the other two groups. The fact that there are no significant differences in this broad trend across a range of conditions is consistent with the hypothesis that genes in this group are highly expressed under wide range of conditions.

acid that it is known to encode, or mathematically as $p(CDN_i|AA_i)$, where $CDN_i$ and $AA_i$ are the codon and amino acid in the $i^{\text{th}}$ position within a gene, respectively. We can also calculate $p(CDN_i|CDN_{i-1})$, the probability of choosing a particular codon given the previous codon encountered. Table 6.3 shows the percentage difference between these values for each possible previous codon, or

$$\frac{p(CDN_i|CDN_{i-1}) - p(CDN_i)}{p(CDN_i)}$$

which shows the amount by which the appearance of a particular codon influences the next codon. This pattern, whereby each position depends on the previous position is similar to a first order Markov chain, where the probability of entering to each state depends only on the previous state, hence the choice of the name first order codon bias to refer to this model. Note that this data is no longer normalised by amino-acid content, as doing so can lead to misinterpretation of the data, which will be discussed later.

The previous codons in Table 6.3 are ordered in inverse, first by last, then middle, then first base. This ordering reveals the largest pattern present in the table, which is that the final base of the previous codon has the most impact over the subsequent codon.

For example, all codons ending with `T` significantly reduce the chance that any of the `AGN` codons will be encountered next. Most codons ending in `T` also repress `GGN` and `CGN` codons, and a subset of them (those ending in `GT`) also repress `TGN` codons, such that codons ending with a `T` repress most codons of the form `NGN`. This pattern is seen in reverse for codons that end with a `C`, which tend to increase the probability of observing an `NGN` immediately afterwards.

Similarly, codons beginning with a `C` are less commonly observed after codons ending with an `A`, although this effect is strongest when the previous codon is of the form `CNA`. A comparable though slightly weaker phenomenon is observed after codons ending in `G`, where codons beginning in `T` or `G` are less commonly observed.

No single consistent biological explanation of these effects is forthcoming, though several have been suggested. Before discussing each in turn, however, let us return to the question of amino-acid content normalisation. Typically when constructing a codon table such as in Table 6.1, the codon frequency is calculated by dividing the number of occurrences of the codon by the number of occurrences of the amino acid it encodes. This makes sense in the case of a codon table because normally when consulting such a table we are not interested in the frequency of the amino acid itself, just the relative frequencies of the codons which encode it.

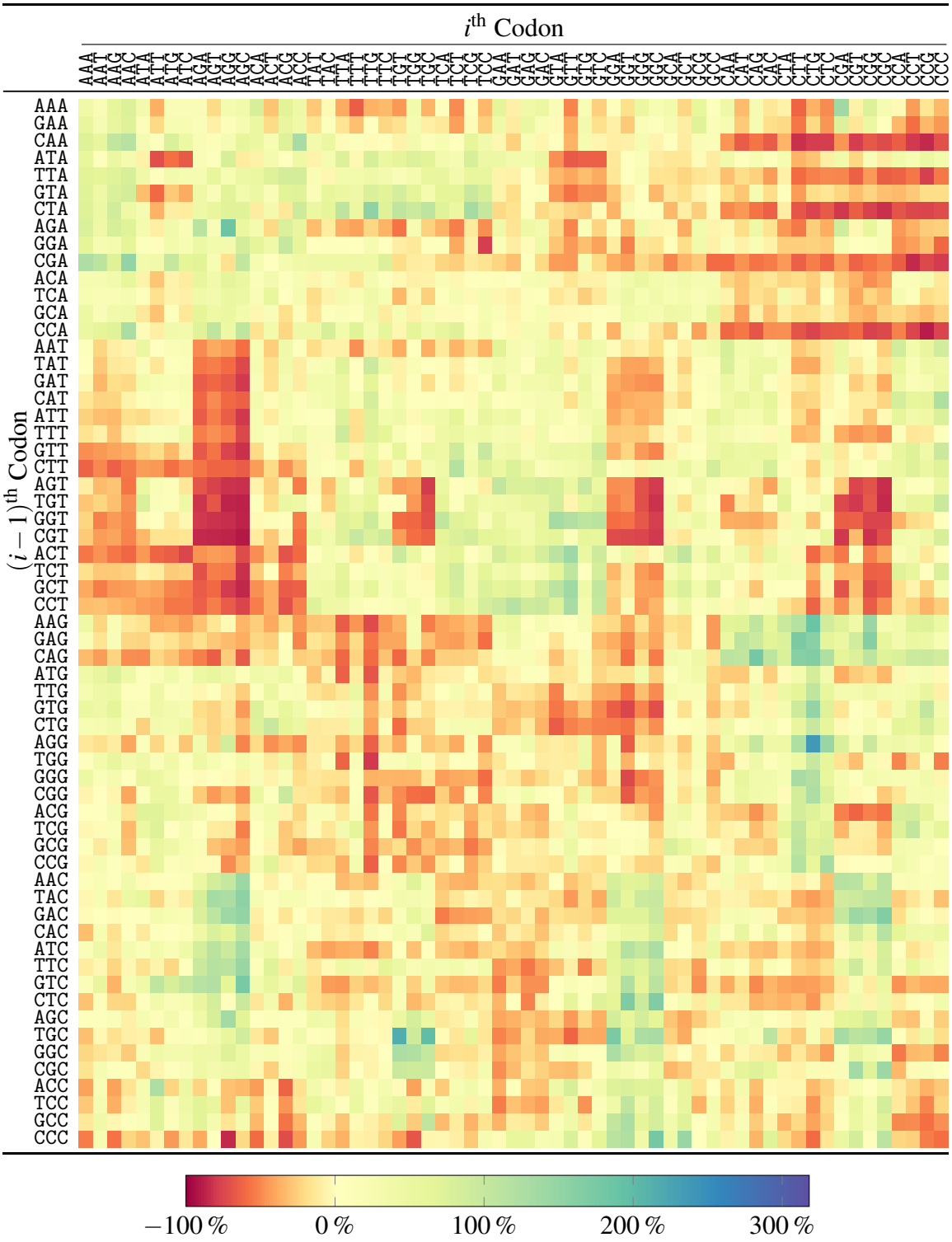



Table 6.3 Percentage change in codon preference due to previous codon. Separate counts of codon usage were made for each possible previous codon for the entire *B. subtilis* genome, and normalised to give the probability of observing a codon, given the previous codon observed. Percentage change was calculated by subtracting then dividing by the background probability of observing each codon. (∗ & †) See Table 6.2

Let us consider the effect of applying that normalisation to the data presented in Table 6.3. As an example, let us consider the observation that codons beginning with C are consistently less likely to be observed when the preceding codon is of the form CNA. Of the 16 codons beginning with a C, four encode leucine (L), four encode proline (P), four encode arginine (R), two encode histidine (H), and the remaining two encode glutamine (G). Of these five amino acids, glutamine, histidine, and proline are exclusively encoded by codons beginning with C, such that a preceding CNA codon reduces the overall probability of encountering any codon which encodes these amino acids. If this data were normalised by amino acid, the change in probability would be hidden for these three amino acids, as the *relative* change in codon bias remains unchanged. The only effect of a preceding CNA which would not be normalised away would be in the case of leucine and arginine, which both have alternate codons which do not begin with C. Under amino-acid normalisation these codons appear to be greatly more likely following a CNA codon because their relative frequency has increased, even though the actual frequency of these codons does not increase drastically after a CNA codon.

One hypothesis to explain why codon preference varies depending on the surrounding content is that there is a preference for codons which maintain the local GC content within a certain range. Since G-C Watson-Crick base pairing is stronger than A-T pairing, the strands in regions of double stranded DNA with a higher GC content are more stably bonded together than regions which are AT rich. Since many cellular processes involve separating the DNA strands – such as DNA replication or RNA transcription – these processes will be affected by how tightly the strands are bound, and there may be some advantage to keeping local GC content within certain parameters. If this effect were significant, one would expect codons which are particularly GC rich to be less common after codons which are also GC rich, and vice versa. Consistent with this is the observation that codons beginning with A are rarer when the preceding codon ends with a T, increasing the chance that the selected codon begins with a G or C. However, codons beginning with a T are not similarly affected by codons ending with a T, and codons ending with an A do not cause the same effect, which is not consistent with the hypothesis that local GC content is the major underlying cause of these effects.

Similarly to the GC content hypothesis, local codon bias could be affected by the secondary structure of the resulting mRNA message. Secondary structure is known to have a significant effect on the translational efficiency of genes, particularly in the region of the Ribosome Binding Site (RBS), although the effect is less pronounced elsewhere as the ribosome seems able to 'iron out' secondary structure once it has bound[31, 83]. Furthermore, the effects shown in Table 6.3 are quite short-range, as the final base of one codon affects

the first base and sometimes the second base of the next, and secondary structure forming interactions typically occur over longer ranges due to the high free energy associated with the small hairpin required for nearby bases to interact.

Interaction between successive tRNAs has also been suggested as a possible cause of some of the codon context effects[9, 105]. Since successive tRNAs come into very close proximity while bound with the ribosome, making is feasible for some form of interaction to occur between them. Such an interaction could aid with translation – for example if it increased recruitment of the next tRNA – or hinder it if for example the interaction were to cause some kind of jamming which required energy to unblock. If this were the case, we would expect the interaction to be dependent on the identities of the two tRNAs involved in codon recognition and amino acid transport for successive codons. This prediction is not consistent with our previous observation that it is in fact the final base of the previous codon which has most weight in determining the effect on the codon bias. Many tRNAs recognise several different codons, typically because they are able to accept several different bases in the third position (hence the term wobble base) and so no clear pattern is seen which could be explained directly by tRNA-tRNA interactions.

The behaviour of rare codons may be important for some of the interactive effects seen in Table 6.3. In *E. coli*, `AGG` is a rare codon which encodes around 3% of arginine residues and is known to be very slowly translated[12]. In fact, it has also been shown that the presence of two or more `AGG` codons immediately after each other increases the probability of frame shift errors during the translation process[26], which result in a non-functional protein. In *E. coli*, nearly 80% of arginine is encoded by only two of the six possible codons, `CGT` and `CGC`, while in *B. subtilis* the same two codons only encode just under 40% of arginine. This greater codon diversity causes the relative frequency of `AGG` to increase from 3% in *E. coli* to 9% in *B. subtilis* – a small increase compared to `AGA` which increases from 4% to 26%. Table 6.3 confirms that, unlike in *E. coli*, `AGG` is in fact slightly more likely to appear when the previous codon was also `AGG` than elsewhere in the *B. subtilis* genome, suggesting that multiple `AGG` codons do not have the same adverse effect in the latter bacteria. While it is perfectly possible that codon abundance has an effect on the frequency of other codon pairs, it does not seem to be a satisfying explanation of the broader patterns seen in Table 6.3.

Other explanations are of course possible – for example, we have not discussed the possibility that some of the effects are driven by some form of grammar at the amino acid level – but there appears to be no single model based on our observations to explain all of the effects noted in Table 6.3. It is possible that these effects are due to a combination of factors rather than one single explanation, or that these effects are due to some higher level process which is as yet hidden from us.

## 6.4 A Simple Algorithm for first Order Codon Optimisation

Despite not being fully able to explain the mechanisms underlying this first order codon bias, we can still use the measured codon distributions in order to generate coding sequences for a given amino acid which closely match the distributions found in native genes. In this work, the following process was used to generate coding sequences given a desired amino acid sequence:

1. Calculate $p(CDN_i|CDN_{i-1}, AA_i)$, the probability of each codon given the previous codon and the amino acid encoded
2. Count the number of incidences of each amino acid
3. For each amino acid, decide how many of each codon should be used in order to given the desired first order codon bias
4. Start with `ATG` and assign codons with probability proportional to $p(CDN_i|CDN_{i-1}, AA_i)$ multiplied by the number of codons remaining to be used
5. Always use the most common stop codon, `TAA`
6. Generate several variants, and score them according to $\sum_{i=2}^{N} \log p(CDN_i|CDN_{i-1}, AA_i)$, the log likelyhood of that ordering

This method does not find the maximum likelyhood ordering of the codon, but instead generates several variants each of which will closely fit the native distribution. The first advantage of this is computational – testing all possible codon orderings for a long gene would be computationally infeasible – but also practical, as most services offering gene synthesis have specific limitation on what sequences can be made, such as repeats or areas of high or low GC content. Generating several permissible variants increases the chances that at least one of them will be synthesisable while it seems unlikely that the slight reduction in fitness from the optimal ordering will have a large impact on the expressibility of the gene.

## 6.5 Conclusions

The goal of this chapter was to investigate codon bias in *B. subtilis* in order to optimise Fluorescent Reporter Proteins (FRPs) for expression in the organism in the dual reporter

system introduced in Chapter 4. In this it was successful, the algorithm described in Section 6.4 above was used to generate optimised coding sequences for published amino acids for a number of FRPs, including eForRed, mCherry, and mScarlet-I. All of these appeared to express well in the organism, with fluorescent output proportional to published brightness levels.

Of course, we cannot conclude that this proposed method is ideal for codon optimisation, and we have not directly demonstrated that this method of codon optimisation is better or worse than any other. To demonstrate this, one would have to generate several variants of genes under various codon selection strategies – of particular interest would be variants with exactly the same codons, but ordered either according to the idea of first order codon bias as introduced above or completely at odds with it. Whether such variants showed any significant difference in translational efficiency would be particularly interesting.

However, such experiments are not trivial to complete. Codon choice or codon ordering cannot be studied in isolation, as both affect several other factors such as GC content and mRNA secondary structure, which can have a significant impact on translation initiation. Secondary structure prediction can help with this – for example by adding a constraint that only sequences with no predicted secondary structure around the RBS are permissible – but secondary structure prediction is not wholly reliable and so such a study should focus on many different proteins and coding sequences to avoid discovering effects specific to the coding sequence of a particular gene.

Furthermore, such a study should not focus solely on the optimisation of FRP proteins, which, while being simple to assay are broadly related and thus likely do not represent a wide selection of sequence space. An optimisation technique which works well for FRPs may not work well for another gene.

What is clear from the data presented in this work is that codon optimisation is not simply a case of selecting the most common codons found in an organism. In *B. subtilis*, codon preference varies depending on the expression context of the gene, with genes which are consistently expressed at high levels having a different codon preference to other more general genes. Native sequences also contain some form of grammar which dictates which codons are more or less likely to follow each other, in the same way as in English where the letter ‘Q’ is highly likely to be followed by a ‘U’, particular codons significantly change the distribution of codons which are likely to succeed them.

Ultimately, it may be that we must move away from the idea of codon ‘optimisation’ altogether, and think instead of codon translation. While it is a remarkable fact that the genetic code which maps codons to amino acids is almost ubiquitous, it is clear that different organisms do speak different dialects of the same language. As we have noted above, `AGG`

is slowly translated in *E. coli*. In some cases, this slight delay in translation may allow the binding of some cofactor, or allow time for the growing poly-peptide chain to fold into some configuration which would be lost if translation continued rapidly. In *B. subtilis*, it appears that `AGG` does not have the same meaning. In the latter the sequence `AGGAGG`, which is rarely seen in *E. coli* due to causing translation errors, actually occurs more frequently than would be expected by random chance. Similarly, the probability of mis-translation – where an incorrect amino acid is inserted into the sequence – varies with different codons[12, 26], such that regions where correct translation is critical for protein function are likely to use more reliable codons than regions where occasional miss-translation is more acceptable.

The subject of codon optimisation clearly has several important applications within synthetic biology, both from a research and an industrial perspective, and there is undoubtedly more to learn about this important topic.

# Chapter 7

# Conclusions

There is a clear need for better detection and management of arsenic contamination, particularly in the developing world. Estimates of the number of people affected by arsenic contaminated drinking water reached 100 million in 2010[43], but efforts to improve testing procedures and the supply of clean water have stagnated over the last decade[43, 68]. Improving access to reliable testing technology has been found to be the most effective solution to tackling the crisis in rural regions where practical limitations prevent the supply of clean water directly[2], however commercially available field testing kits remain outside the reach of local communities without the efforts of government, non-governmental organisations, or the international community.

Furthermore, contemporary kits remain complex and potentially hazardous to use, even with proper training. Early kits were known to emit significant quantities of highly toxic arsine gas[70], presenting a serious health hazard to the user and anyone in the vicinity. While later kits have become better at containing the deadly gas, their function relies on the same chemistry and so the possibility of release due to misuse remains. In addition, the chemicals required for the tests are also hazardous and require correct disposal. The shallow tube wells at the heart of this mass poisoning, the largest to effect an entire population in history[151], were installed in order to reduce deaths due to cholera and other surface-borne diseases. In this they were successful[22], but the lack of research and monitoring of the programme has lead to a situation where arsenic contamination is responsible for up to 15% of deaths in some areas[44]. Careless disposal of millions of traditional arsenic sensors has the potential to also become a serious public health issue[131].

Previous research has indicated that biology may be able to provide a radically different approach to arsenic detection[149, 153], however a quantitative understanding of the mechanisms by which biological systems defend themselves from the deleterious effects of arsenic was lacking. In addition to the practical issues of distributing engineered biological material

to end-users in remote locations in a device which is simple and easy to use and as cheap as possible is a challenge that has yet to be solved.

In **Chapter 4**, a method for ratiometric characterisation of biological parts was established in the spore-forming gram-negative bacterium *B. subtilis*. While the arsenic mitigation system of this bacterium is slightly different to the better studied version that is found in *E. coli*, the ability to form highly resistant spores which are able to remain dormant for many years makes them an ideal vector for distribution. However, as *B. subtilis* is less commonly used in synthetic biology than *E. coli*, much of the progress in characterisation techniques which have been made in *E. coli* have yet to be applied to it. Further work investigating gene expression in *B. subtilis* was carried out in **Chapter 6**, in which the nature of codon optimisation found in the bacteria was investigated. The application of ratiometric characterisation to the bacteria will hopefully serve as a useful platform for further synthetic biology in the organism.

Early work on a biological biosensor established the possibility that bacteria are capable of sensing arsenic at concentrations relevant for sensing applications, but failed to approach the question from an analytical or engineering standpoint. **Chapter 3** attempts to rectify this by developing a mathematical model of the mechanism through which *B. subtilis* protect themselves from arsenic. While much work has previously been done to identify the mechanism by which *B. subtilis* are able to detect arsenic, the precise details of the mechanism are not fully understood. However, parameter variations have shown that the native regulatory architecture is such that the form of the response to arsenic is invariant to these changes. Key topics for further research are the mechanisms by and rate at which difference forms of arsenic enter the cell, as well as studying the rate at which arsenite efflux occurs do the the arsnite transporter, ArsB.

Having developed a mathematical understanding of the response of *B. subtilis* to arsenic in Chapter 3, the rigorous characterisation methodology developed in Chapter 4 was applied to improving the practical knowledge of the arsenic response in **Chapter 5**. The form of the observed response matched that which was previously predicted very closely, confirming the validity of the model which was developed. The system was found to respond detectably at arsenite concentrations as low as $2\,\mu g\,l^{-1}$, and modelling suggests that there is no fundamental lower detection threshold other than that set by the sensitivity of the instruments used and the limitations of the stochastic nature of gene expression.

The response to arsenic which was predicted and observed previously is not ideal for a sensor, as differentiating between a reading which is just above or just below a given threshold is not straightforward to do by eye. This very problem is also an issue with current arsenic sensors, where the user must distinguish between similar shades of yellow to determine whether the sample is above or below the permissible threshold. One possible solution to

this issue could be the use of a synthetic toggle switch similar to the one first published by Gardner et al. [47]. Under the right conditions, such a switch has two stable points and can be toggled between these points based on the concentration of an inducer.

While this work has demonstrated that an arsenic biosensor is within our technical grasp, considerable further work is required in order to make this project a reality. A better understanding of the relationship between how the arsenic in the environment affects the concentration within the cell is necessary in order to build a well functioning whole-cell biosensor. In particular, the mechanism through which arsenite and arsenate enter the cell is not fully understood, and little to no quantitative data is available on this. By better understanding these mechanisms, we may be able to re-engineer the systems responsible for uptake in order to harmonise the effects of arsenite and arsenate such that the sensor can be used for both. The new knowledge would also give greater control over the internal arsenic concentration in relation to the external arsenic concentration, drastically increasing the range of concentrations which can be detected.

There is also an increasing interest in cell-free systems as an alternative to whole-cell systems, as these systems avoid the regulation associated with the use of GMOs. Cell-free systems have many benefits over whole-cell systems, such as the speed of design and testing and by using freeze-drying technology can also be shipped in a manner not unlike bacterial spores. However, cell-free systems necessarily lack a cell membrane and are thus more exposed to their environment, making them more susceptible to contamination from other chemicals present in the samples. Thorough controls would be required to ensure false negatives were not generated by deleterious contaminants, as well as a wide range of experiments to test for other potential inducers of the system.

Ultimately, the success or failure of any such device, irrespective of the technology behind it, depends upon successful engagement with local communities in affected areas. While education on mitigating the risks of diseases such cholera is widespread, understanding of the risks of arsenic contamination and the steps which can be taken to manage that risk have yet to catch up.

# References


[1] Ahmann, D., Roberts, A. L., Krumholz, L. R., and Morel, F. M. M. (1994). Microbe grows by reducing arsenic. *Nature*, 371(6500):750–750.

[2] Ahmed, M. F., Ahuja, S., Alauddin, M., Hug, S. J., Lloyd, J. R., Pfaff, A., Pichler, T., Saltikov, C., Stute, M., and van Geen, A. (2006). Ensuring Safe Drinking Water in Bangladesh. *Science*, 314(5806).

[3] Akashi, H. and Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6):3695–700.

[4] Alieva, N. O., Konzen, K. A., Field, S. F., Meleshkevitch, E. A., Hunt, M. E., Beltran-Ramirez, V., Miller, D. J., Wiedenmann, J., Salih, A., and Matz, M. V. (2008). Diversity and evolution of coral fluorescent proteins. *PloS one*, 3(7):e2680.

[5] Barbe, V., Cruveiller, S., Kunst, F., Lenoble, P., Meurice, G., Sekowska, A., Vallenet, D., Wang, T., Moszer, I., Medigue, C., and Danchin, A. (2009). From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later. *Microbiology*, 155(6):1758–1775.

[6] Beijer, L., Nilsson, R.-P., Holmberg, C., and Rutberg, L. (1993). The glpP and glpF genes of the glycerol regulon in Bacillus subtilis. *Journal of General Microbiology*, 139(2):349–359.

[7] Bennett, M. S., Guan, Z., Laurberg, M., and Su, X. D. (2001). Bacillus subtilis arsenate reductase is structurally and functionally similar to low molecular weight protein tyrosine phosphatases. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13577–82.

[8] Bennett, R. L. and Malamy, M. H. (1970). Arsenate resistant mutants of and phosphate transport. *Biochemical and Biophysical Research Communications*, 40(2):496–503.

[9] Berg, O. and Silva, P. J. N. (1997). Codon bias in Escherichia coli: the influence of codon context on mutation and selection. *Nucleic Acids Research*, 25(7):1397–1404.

[10] Bindels, D. S., Haarbosch, L., van Weeren, L., Postma, M., Wiese, K. E., Mastop, M., Aumonier, S., Gotthard, G., Royant, A., Hink, M. A., and Gadella, T. W. J. (2016). mScarlet: a bright monomeric red fluorescent protein for cellular imaging. *Nature Methods*, 14(1):53–56.

[11] Bloor, A. E. and Cranenburgh, R. M. (2006). An efficient method of selectable marker gene excision by Xer recombination for gene replacement in bacterial chromosomes. *Applied and environmental microbiology*, 72(4):2520–5.

[12] Bonekamp, F. and Jensen, K. F. (1988). The AGG codon is translated slowly in *E. coli* even at very low expression levels. *Nucleic Acids Research*, 16(7):3013–3024.

[13] Bralatei, E., Lacan, S., Krupp, E. M., and Feldmann, J. (2015). Detection of Inorganic Arsenic in Rice Using a Field Test Kit: A Screening Method. *Analytical Chemistry*, 87(22):11271–11276.

[14] Briley Jr, K., Dorsey-Oresto, A., Prepiak, P., Dias, M. J., Mann, J. M., and Dubnau, D. (2011a). The secretion ATPase ComGA is required for the binding and transport of transforming DNA. *Molecular Microbiology*, 81(3):818–830.

[15] Briley Jr, K., Prepiak, P., Dias, M. J., Hahn, J., and Dubnau, D. (2011b). Maf acts downstream of ComGA to arrest cell division in competent cells of B. subtilis. *Molecular Microbiology*, 81(1):23–39.

[16] Burkholder, P. R. and Giles, N. H. (1947). Induced Biochemical Mutations in Bacillus subtilis. *American Journal of Botany*, 34(6):345.

[17] Busenlehner, L. S., Apuy, J. L., and Giedroc, D. P. (2002a). Characterization of a metalloregulatory bismuth(III) site in Staphylococcus aureus pI258 CadC repressor. *JBIC Journal of Biological Inorganic Chemistry*, 7(4-5):551–559.

[18] Busenlehner, L. S., Cosper, N. J., Scott, R. A., Rosen, B. P., Wong, M. D., and Giedroc, D. P. (2001). Spectroscopic Properties of the Metalloregulatory Cd(II) and Pb(II) Sites of S. aureus pI258 CadC.

[19] Busenlehner, L. S., Pennella, M. A., and Giedroc, D. P. (2003). The SmtB/ArsR family of metalloregulatory transcriptional repressors: structural insights into prokaryotic metal resistance. *FEMS Microbiology Reviews*, 27(2-3):131–143.

[20] Busenlehner, L. S., Weng, T.-C., Penner-Hahn, J. E., and Giedroc, D. P. (2002b). Elucidation of Primary ($\alpha$3N) and Vestigial ($\alpha$5) Heavy Metal-binding Sites in Staphylococcus aureus pI258 CadC: Evolutionary Implications for Metal Ion Selectivity of ArsR/SmtB Metal Sensor Proteins. *Journal of Molecular Biology*, 319(3):685–701.

[21] Butcher, B. G., Deane, S. M., and Rawlings, D. E. (2000). The chromosomal arsenic resistance genes of Thiobacillus ferrooxidans have an unusual arrangement and confer increased arsenic and antimony resistance to Escherichia coli. *Applied and environmental microbiology*, 66(5):1826–33.

[22] Carrel, M., Escamilla, V., Messina, J., Giebultowicz, S., Winston, J., Yunus, M., Streatfield, P. K., and Emch, M. (2011). Diarrheal disease risk in rural Bangladesh decreases as tubewell density increases: a zero-inflated and geographically weighted analysis. *International journal of health geographics*, 10:41.

[23] Cavet, J. S., Meng, W., Pennella, M. A., Appelhoff, R. J., Giedroc, D. P., and Robinson, N. J. (2002). A nickel-cobalt-sensing ArsR-SmtB family repressor. Contributions of cytosol and effector binding sites to metal selectivity. *The Journal of biological chemistry*, 277(41):38441–8.

[24] Cervantes, C. (1994). Resistance to arsenic compounds in microorganisms. *FEMS Microbiology Reviews*, 15(4):355–367.

[25] Charvin, G., Cross, F. R., and Siggia, E. D. (2008). A Microfluidic Device for Temporally Controlled Gene Expression and Long-Term Fluorescent Imaging in Unperturbed Dividing Yeast Cells. *PLoS ONE*, 3(1):e1468.

[26] Chen, G.-F. T. and Inouye, M. (1990). Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the Escherichia coli genes. *Nucleic Acids Research*, 18(6):1465–1473.

[27] Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L., and McAdams, H. H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(10):3480–5.

[28] Chen, Y. and Rosen, B. P. (1997). Metalloregulatory properties of the ArsD repressor. *The Journal of biological chemistry*, 272(22):14257–62.

[29] Christen, K. (2001). Chickens, manure, and arsenic. *Environ. Sci. Technol.*, 35(9):184A–185A.

[30] Churchill, M. E. and Travers, A. A. (1991). Protein motifs that recognize structural features of DNA. *Trends in Biochemical Sciences*, 16:92–97.

[31] Davis, J. H., Rubin, A. J., and Sauer, R. T. (2011). Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic acids research*, 39(3):1131–41.

[32] de Jong, H., Ranquet, C., Ropers, D., Pinel, C., and Geiselmann, J. (2010). Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. *BMC systems biology*, 4(1):55.

[33] Dey, S., Dou, D., and Rosen, B. P. (1994). ATP-dependent arsenite transport in everted membrane vesicles of Escherichia coli. *The Journal of Biological Chemistry*, 269(41):25442–25446.

[34] Dey, S. and Rosen, B. P. (1995). Dual mode of energy coupling by the oxyanion-translocating ArsB protein. *Journal of bacteriology*, 177(2):385–9.

[35] Earl, A. M., Losick, R., and Kolter, R. (2008). Ecology and genomics of Bacillus subtilis. *Trends in Microbiology*, 16(6):269–275.

[36] Ehrenberg, C. G. (1835). Physikalische Abhandlungen der Koeniglichen Akademie der Wissenschaften zu Berlin aus den Jahren 1833–1835. pages 145–336.

[37] Eicken, C., Pennella, M. A., Chen, X., Koshlap, K. M., VanZile, M. L., Sacchettini, J. C., and Giedroc, D. P. (2003). A Metal–Ligand-mediated Intersubunit Allosteric Switch in Related SmtB/ArsR Zinc Sensor Proteins. *Journal of Molecular Biology*, 333(4):683–695.

[38] Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science (New York, N.Y.)*, 297(5584):1183–6.

[39] Engler, C. and Marillonnet, S. (2013). Combinatorial DNA assembly using Golden Gate cloning. *Methods in molecular biology (Clifton, N.J.)*, 1073:141–56.

[40] Erickson, B. E. (2003). Field kits fail to provide accurate measure of arsenic in groundwater. *Environmental Science & Technology*, pages 35–38.

[41] European Commission (2003). Commission Directive 2003/2/EC of 6 January 2003 relating to restrictions on the marketing and use of arsenic.

[42] FDA (2015). CVM Updates - FDA Announces Pending Withdrawal of Approval of Nitarsone.

[43] Fendorf, S., Michael, H. A., and van Geen, A. (2010). Spatial and temporal variations of groundwater arsenic in South and Southeast Asia. *Science (New York, N.Y.)*, 328(5982):1123–7.

[44] Flanagan, S., Johnston, R., and Zheng, Y. (2012). Arsenic in tube well water in Bangladesh: health and economic impacts and implications for arsenic mitigation. *Bulletin of the World Health Organization*, 90(11):839–846.

[45] French, C. E., de Mora, K., Joshi, N., Elfick, A., Haseloff, J., and Ajioka, J. (2011). Synthetic Biology and the Art of Biosensor Design.

[46] Fukushima, T., Ishikawa, S., Yamamoto, H., Ogasawara, N., and Sekiguchi, J. (2003). Transcriptional, functional and cytochemical analyses of the veg gene in Bacillus subtilis. *Journal of biochemistry*, 133(4):475–83.

[47] Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000). Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403(6767):339–342.

[48] George, C. M., Zheng, Y., Graziano, J. H., Rasul, S. B., Hossain, Z., Mey, J. L., and van Geen, A. (2012). Evaluation of an Arsenic Test Kit for Rapid Well Screening in Bangladesh. *Environmental Science & Technology*, 46(20):11213–11219.

[49] Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods*, 6(5):343–5.

[50] Gladysheva, T. B., Oden, K. L., and Rosen, B. P. (1994). Properties of the Arsenate Reductase of Plasmid R773. *Biochemistry*, 33(23):7288–7293.

[51] Goedhart, J., van Weeren, L., Hink, M. A., Vischer, N. O. E., Jalink, K., and Gadella, T. W. J. (2010). Bright cyan fluorescent protein variants identified by fluorescence lifetime screening. *Nature Methods*, 7(2):137–139.

[52] Goedhart, J., von Stetten, D., Noirclerc-Savoye, M., Lelimousin, M., Joosen, L., Hink, M. A., van Weeren, L., Gadella, T. W. J., Royant, A., and Royant, A. (2012). Structure-guided evolution of cyan fluorescent proteins towards a quantum yield of 93%. *Nature communications*, 3:751.

[53] Gouy, M. and Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, 10(22):7055–7074.

[54] Government of Bangladesh (2004). National Policy for Arsenic Mitigation.

[55] Gralla, J. D. (1990). Promoter recognition and mRNA initiation by Escherichia coli E$\sigma$70. pages 37–54.

[56] Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pavé, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8(1):197–197.

[57] Haldenwang, W. G. (1995). The sigma factors of Bacillus subtilis. *Microbiological reviews*, 59(1):1–30.

[58] Harold, F. M. and Baarda, J. R. (1966). Interaction of arsenate with phosphate-transport systems in wild- type and mutant Streptococcus faecalis. *Journal of bacteriology*, 91(6):2257–62.

[59] Härtl, B., Wehrl, W., Wiegert, T., Homuth, G., and Schumann, W. (2001). Development of a new integration site within the Bacillus subtilis chromosome and construction of compatible expression cassettes. *Journal of bacteriology*, 183(8):2696–9.

[60] Harvie, D. R., Andreini, C., Cavallaro, G., Meng, W., Connolly, B. A., Yoshida, K.-i., Fujita, Y., Harwood, C. R., Radford, D. S., Tottey, S., Cavet, J. S., and Robinson, N. J. (2006). Predicting metals sensed by ArsR-SmtB repressors: allosteric interference by a non-effector metal. *Molecular Microbiology*, 59(4):1341–1356.

[61] Herbel, M. J., Switzer Blum, J., Hoeft, S. E., Cohen, S. M., Arnold, L. L., Lisak, J., Stolz, J. F., Oremland, R. S., N.S., B., and D., W. (2002). Dissimilatory arsenate reductase activity and arsenate-respiring bacteria in bovine rumen fluid, hamster feces, and the termite hindgut. *FEMS Microbiology Ecology*, 41(1):59–67.

[62] Hidenori, S. and Henner, D. J. (1986). Construction of a single-copy integration vector and its use in analysis of regulation of the trp operon of Bacillus subtilis. *Gene*, 43(1-2):85–94.

[63] Hileman, B. (2007). Arsenic In Chicken Production. *Chemical & Engineering News*, 85(15):34–35.

[64] Hodgman, C. E. and Jewett, M. C. (2012). Cell-free synthetic biology: Thinking outside the cell. *Metabolic Engineering*, 14(3):261–269.

[65] Holmgren, A. and Aslund, F. (1995). Glutaredoxin. pages 283–292.

[66] Hossain, M. (2006). Arsenic contamination in Bangladesh—An overview. *Agriculture, Ecosystems & Environment*, 113(1-4):1–16.

[67] Hotter, G. S., Wilson, T., and Collins, D. M. (2001). Identification of a cadmium-induced gene in Mycobacterium bovis and Mycobacterium tuberculosis. *FEMS Microbiology Letters*, 200(2):151–155.

[68] HRW (2016). The Failing Response to Arsenic in the Drinking Water of Bangladesh's Rural Poor. Technical report, Human Rights Watch.

[69] Hua, L., Li, M., Sun, X., Wang, J., Li, Z., Xu, Y., Hu, S., and Chen, H. (2012). A dual color fluorescent reporter system for the real time detection of promoter activity. *Biotechnology letters*, 34(5):823–30.

[70] Hussam, A., Alauddin, M., Khan, A. H., And, S. B. R., and Munir, A. K. M. (1999). Evaluation of Arsine Generation in Arsenic Field Kit.

[71] Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, 2(1):13–34.

[72] Jackson, C. R. and Dugas, S. L. (2003). Phylogenetic analysis of bacterial and archaeal arsC gene sequences suggests an ancient, common origin for arsenate reductase. *BMC evolutionary biology*, 3:18.

[73] Jakariya, M., Vahter, M., Rahman, M., Wahed, M. A., Hore, S. K., Bhattacharya, P., Jacks, G., and Persson, L. Å. (2007). Screening of arsenic in tubewell water with field test kits: Evaluation of the method from public health perspective. *Science of The Total Environment*, 379(2-3):167–175.

[74] Johnston, R. B. and Sarker, M. H. (2007). Arsenic mitigation in Bangladesh: National screening data and case studies in three upazilas. *Journal of Environmental Science and Health, Part A*, 42(12):1889–1896.

[75] Kabir, A. and Howard, G. (2007). Sustainability of arsenic mitigation in Bangladesh: Results of a functionality survey. *International Journal of Environmental Health Research*, 17(3):207–218.

[76] Kanaya, S., Yamada, Y., Kudo, Y., and Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238(1):143–155.

[77] Kar, S. R., Adams, A. C., Lebowitz, J., Taylor, K. B., and Hall, L. M. (1997). The Cyanobacterial Repressor SmtB Is Predominantly a Dimer and Binds Two Zn2+ Ions per Subunit†.

[78] Kinniburgh, D. G., Smedley, P. L., and (Editors) (2001). Arsenic contamination of groundwater in Bangladesh. Technical report, British Geological Survey: Keyworth.

[79] Knight, R. D., Freeland, S. J., and Landweber, L. F. (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology*, 2(4):research0010.1.

[80] Kosower, N. S. and Kosower, E. M. (1995). Diamide: An oxidant probe for thiols. *Methods in Enzymology*, 251(C):123–133.

[81] Kosower, N. S., Kosower, E. M., Wertheim, B., and Correa, W. S. (1969). Diamide, a new reagent for the intracellular oxidation of glutathione to the disulfide. *Biochemical and Biophysical Research Communications*, 37(4):593–596.

[82] Kredel, S., Oswald, F., Nienhaus, K., Deuschle, K., Röcker, C., Wolff, M., Heilker, R., Nienhaus, G. U., and Wiedenmann, J. (2009). mRuby, a Bright Monomeric Red Fluorescent Protein for Labeling of Subcellular Structures. *PLoS ONE*, 4(2):e4391.

[83] Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. *Science (New York, N.Y.)*, 324(5924):255–8.

[84] Kuroda, M., Dey, S., Sanders, O. I., and Rosen, B. P. (1997). Alternate energy coupling of ArsB, the membrane subunit of the Ars anion-translocating ATPase. *The Journal of biological chemistry*, 272(1):326–31.

[85] Laverman, A. M., Blum, J. S., Schaefer, J. K., Phillips, E., Lovley, D. R., and Oremland, R. S. (1995). Growth of Strain SES-3 with Arsenate and Other Diverse Electron Acceptors. *Applied and environmental microbiology*, 61(10):3556–61.

[86] Levin, P. A. and Grossman, A. D. (1998). Cell cycle and sporulation in Bacillus subtilis. *Current Opinion in Microbiology*, 1(6):630–635.

[87] Li, Q., Li, C., Xie, L., Zhang, C., Feng, Y., and Xie, J. (2017). Characterization of a putative ArsR transcriptional regulator encoded by Rv2642 from Mycobacterium tuberculosis. *Journal of Biomolecular Structure and Dynamics*, 35(9):2031–2039.

[88] Li, S., Chen, Y., and Rosen, B. P. (2001). Role of vicinal cysteine pairs in metalloid sensing by the ArsD As(III)-responsive repressor. *Molecular Microbiology*, 41(3):687–696.

[89] Lin, Y.-F., Walmsley, A. R., and Rosen, B. P. (2006). An arsenic metallochaperone for an arsenic detoxification pump. *Proceedings of the National Academy of Sciences of the United States of America*, 103(42):15617–22.

[90] Lin, Y.-F., Yang, J., and Rosen, B. P. (2007). ArsD residues Cys12, Cys13, and Cys18 form an As(III)-binding site required for arsenic metallochaperone activity. *The Journal of biological chemistry*, 282(23):16783–91.

[91] Liu, Z., Shen, J., Carbrey, J. M., Mukhopadhyay, R., Agre, P., and Rosen, B. P. (2002). Arsenite transport by mammalian aquaglyceroporins AQP7 and AQP9. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):6053–8.

[92] Locke, J. C. W., Young, J. W., Fontes, M., Hernández Jiménez, M. J., and Elowitz, M. B. (2011). Stochastic pulse regulation in bacterial stress response. *Science (New York, N.Y.)*, 334(6054):366–9.

[93] Lu, J. and Holmgren, A. (2014). The thioredoxin antioxidant system. *Free Radical Biology and Medicine*, 66:75–87.

[94] Maass, S., Sievers, S., Zühlke, D., Kuzinski, J., Sappa, P. K., Muntel, J., Hessling, B., Bernhardt, J., Sietmann, R., Völker, U., Hecker, M., and Becher, D. (2011). Efficient, Global-Scale Quantification of Absolute Protein Amounts by Integration of Targeted Mass Spectrometry and Two-Dimensional Gel-Based Proteomics. *Analytical Chemistry*, 83(7):2677–2684.

[95] Marsh, J. (1836). Account of a method of separating small quantities of arsenic from substances with which it may be mixed. *Edinburgh New Philosophical Journal*, 21:229–236.

[96] Martin, P., DeMel, S., Shi, J., Gladysheva, T., Gatti, D. L., Rosen, B. P., and Edwards, B. F. (2001). Insights into the Structure, Solvation, and Mechanism of ArsC Arsenate Reductase, a Novel Arsenic Detoxification Enzyme. *Structure*, 9(11):1071–1081.

[97] Mateos, L. M., Ordóñez, E., Letek, M., and Gil, J. A. (2006). Corynebacterium glutamicumas a model bacterium for the bioremediation of arsenic. *International Microbiology*, 9(3):207–215.

[98] McCarty, K. M., Senn, D. B., Kile, M. L., Quamruzzaman, Q., Rahman, M., Mahiuddin, G., and Christiani, D. C. (2004). Antimony: an unlikely confounder in the relationship between well water arsenic and health outcomes in Bangladesh. *Environmental health perspectives*, 112(8):809–11.

[99] Melamed, D. (2004). Monitoring arsenic in the environment: a review of science and technologies with the potential for field measurements. *Analytica Chimica Acta*, 532(1):1–13.

[100] Meng, Y.-L., Liu, Z., and Rosen, B. P. (2004). As(III) and Sb(III) uptake by GlpF and efflux by ArsB in Escherichia coli. *The Journal of biological chemistry*, 279(18):18334–41.

[101] Milton, A. H., Hore, S. K., Hossain, M. Z., and Rahman, M. (2012). Bangladesh arsenic mitigation programs: lessons from the past. *Emerging health threats journal*, 5.

[102] Moore, C. M., Gaballa, A., Hui, M., Ye, R. W., and Helmann, J. D. (2005). Genetic and physiological responses of Bacillus subtilis to metal ion stress. *Molecular microbiology*, 57(1):27–40.

[103] Moran, C. P., Lang, N., LeGrice, S. F. J., Lee, G., Stephens, M., Sonenshein, A. L., Pero, J., and Losick, R. (1982). Nucleotide sequences that signal the initiation of transcription and translation inBacillus subtilis. *MGG Molecular & General Genetics*, 186(3):339–346.

[104] Moré, J. J., Garbow, B. S., and Hillstrom, K. E. (1980). User Guide for MINPACK-1.

[105] Moszer, I., Rocha, E. P., and Danchin, A. (1999). Codon usage and lateral gene transfer in Bacillus subtilis. *Current Opinion in Microbiology*, 2(5):524–528.

[106] Mukhopadhyay, R., Rosen, B. P., Phung, L. T., Silver, S., A., D., M., D. G., J.M., W., R., W., L., W., and I., Z. (2002). Microbial arsenic: from geocycles to genes and enzymes. *FEMS Microbiology Reviews*, 26(3):311–325.

[107] Nagai, T., Ibata, K., Park, E. S., Kubota, M., Mikoshiba, K., and Miyawaki, A. (2002). A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nature Biotechnology*, 20(1):87–90.

[108] Nakano, M. M. and Zuber, P. (1998). Anaerobic Growth of a "Strict Aerobe". *Annual Review of Microbiology*, 52(1):165–190.

[109] National Research Council (U.S.). Subcommittee on Arsenic in Drinking Water. (1999). *Arsenic in drinking water*. National Academy Press.

[110] NEB (2015). PCR Using Q5® High-Fidelity DNA Polymerase (M0491) | NEB.

[111] Nicolas, P., Leduc, A., Robin, S., Rasmussen, S., Jarmer, H., and Bessières, P. (2009). Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. *Bioinformatics (Oxford, England)*, 25(18):2341–7.

[112] Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., Becher, D., Bisicchia, P., Botella, E., Delumeau, O., Doherty, G., Denham, E. L., Fogg, M. J., Fromion, V., Goelzer, A., Hansen, A., Härtig, E., Harwood, C. R., Homuth, G., Jarmer, H., Jules, M., Klipp, E., Le Chat, L., Lecointe, F., Lewis, P., Liebermeister, W., March, A., Mars, R. A. T., Nannapaneni, P., Noone, D., Pohl, S., Rinn, B., Rügheimer, F., Sappa, P. K., Samson, F., Schaffer, M., Schwikowski, B., Steil, L., Stülke, J., Wiegert, T., Devine, K. M., Wilkinson, A. J., van Dijl, J. M., Hecker, M., Völker, U., Bessières, P., and Noirot, P. (2012). Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science (New York, N.Y.)*, 335(6072):1103–6.

[113] Nies, D. H. and Silver, S. (1995). Ion efflux systems involved in bacterial metal resistances. *Journal of Industrial Microbiology*, 14(2):186–199.

[114] Oden, K. L., Gladysheva, T. B., and Rosen, B. P. (1994). Arsenate reduction mediated by the plasmid-encoded ArsC protein is coupled to glutathione. *Molecular Microbiology*, 12(2):301–306.

[115] Ohtsuka, T., Yamaguchi, N., Makino, T., Sakurai, K., Kimura, K., Kudo, K., Homma, E., Dong, D. T., and Amachi, S. (2013). Arsenic Dissolution from Japanese Paddy Soil by a Dissimilatory Arsenate-Reducing Bacterium Geobacter sp. OR-1. *Environmental Science & Technology*, page 130528105708003.

[116] OpenWetWare (2015a). Bacillus Subtilis Transformation.

[117] OpenWetWare (2015b). Sanger sequencing services.

[118] OpenWetWare (2015c). TOP10 chemically competent cells.

[119] Oremland, R., Newman, D., Wail, B., and Stolz, J. (2002a). *Environmental chemistry of arsenic*. Marcel Dekker.

[120] Oremland, R. S., Hoeft, S. E., Santini, J. M., Bano, N., Hollibaugh, R. A., and Hollibaugh, J. T. (2002b). Anaerobic oxidation of arsenite in Mono Lake water and by a facultative, arsenite-oxidizing chemoautotroph, strain MLHE-1. *Applied and environmental microbiology*, 68(10):4795–802.

[121] Oremland, R. S. and Stolz, J. F. (2003). The Ecology of Arsenic. *Science*, 300(5621).

[122] Oremland, R. S. and Stolz, J. F. (2005). Arsenic, microbes and contaminated aquifers. *Trends in Microbiology*, 13(2):45–49.

[123] Pande, S. P., Deshpande, L. S., and Kaul, S. N. (2001). Laboratory and Field Assessment of Arsenic Testing Field Kits in Bangladesh and West Bengal, India. *Environmental Monitoring and Assessment*, 68(1):1–18.

[124] Pardee, K., Green, A., Ferrante, T., Cameron, D., DaleyKeyser, A., Yin, P., Collins, J., Al-Khabouri, S., Fall, C., Noireaux, V., Murray, R., and Lucks, J. (2014). Paper-Based Synthetic Gene Networks. *Cell*, 159(4):940–954.

[125] Pathey, P. (2009). Monitoring the Situation of Children and Women: Multiple Indicator Cluster Survey 2009. Technical report, Bangladesh Bureau of Statistics and United Nations Children's Fund (UNICEF).

[126] Pennella, M. A., Shokes, J. E., Cosper, N. J., Scott, R. A., and Giedroc, D. P. (2003). Structural elements of metal selectivity in metal sensor proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):3713–8.

[127] Praxair (2016). Arsine Arsine Safety Data Sheet P-4565.

[128] Qi, Y., Kobayashi, Y., and Hulett, F. M. (1997). The pst operon of Bacillus subtilis has a phosphate-regulated promoter and is involved in phosphate transport but not in regulation of the pho regulon. *Journal of bacteriology*, 179(8):2534–9.

[129] Qiagen (2015a). QIAprep Spin Miniprep Kit - QIAGEN.

[130] Qiagen (2015b). QIAquick Gel Extraction Kit - QIAGEN.

[131] Rahman, M. M., Mukherjee, D., Sengupta, M. K., Chowdhury, U. K., Lodh, D., Chanda, C. R., Roy, S., Selim, M., Quamruzzaman, Q., Milton, A. H., Shahidullah, S. M., Rahman, M. T., and Chakraborti, D. (2002). Effectiveness and Reliability of Arsenic Field Testing Kits: Are the Million Dollar Screening Projects Effective or Not?

[132] Rao, C. V., Wolf, D. M., and Arkin, A. P. (2002). Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912):231–7.

[133] Ravenscroft, P., Kabir, A., Ibn Hakim, S. A., Ibrahim, A. K. M., Ghosh, S. K., Rahman, M. S., Akhter, F., and Sattar, M. A. (2014). Effectiveness of public rural waterpoints in Bangladesh with special reference to arsenic mitigation. *Journal of Water Sanitation and Hygiene for Development*, 4(4).

[134] Rosen, B. P. (1999). Families of arsenic transporters. *Trends in Microbiology*, 7(5):207–212.

[135] Rosen, B. P. (2002). Biochemistry of arsenic detoxification. *FEBS Letters*, 529(1):86–92.

[136] Rosenberg, H., Gerdes, R. G., and Chegwidden, K. (1977). Two systems for the uptake of phosphate in Escherichia coli. *Journal of bacteriology*, 131(2):505–11.

[137] Saha, K. (1995). Chronic Arsenical Dermatoses from tube-well water in West Bengal during 1983-87. *Indian Journal of Dermatology*, 40(1).

[138] Saikat, S. Q., Selim, A. M., Kessi, J., Wehrli, E., and Hanselmann, K. W. (2001). Transformation of Arsenic Compounds by Bacteria from Groundwater Sediments of Bangladesh. Technical report.

[139] Sankararamakrishnan, N., Chauhan, D., Nickson, R., Tripathi, R., and Iyengar, L. (2008). Evaluation of two commercial field test kits used for screening of groundwater for arsenic in Northern India. *Science of The Total Environment*, 401(1-3):162–167.

[140] Santini, J. M., Sly, L. I., Schnagl, R. D., and Macy, J. M. (2000). A new chemolithoautotrophic arsenite-oxidizing bacterium isolated from a gold mine: phylogenetic, physiological, and preliminary biochemical studies. *Applied and environmental microbiology*, 66(1):92–7.

[141] Sato, T. and Kobayashi, Y. (1998). The ars operon in the skin element of Bacillus subtilis confers resistance to arsenate and arsenite. *Journal of bacteriology*, 180(7):1655–61.

[142] Scharf, C., Riethdorf, S., Ernst, H., Engelmann, S., Völker, U., and Hecker, M. (1998). Thioredoxin is an essential protein induced by multiple stresses in Bacillus subtilis. *Journal of bacteriology*, 180(7):1869–77.

[143] Schifano, J. M., Edifor, R., Sharp, J. D., Ouyang, M., Konkimalla, A., Husson, R. N., and Woychik, N. A. (2013). Mycobacterial toxin MazF-mt6 inhibits translation through cleavage of 23S rRNA at the ribosomal A site. *Proceedings of the National Academy of Sciences of the United States of America*, 110(21):8501–6.

[144] Setlow, P. (2006). Spores of Bacillus subtilis: their resistance to and killing by radiation, heat and chemicals. *Journal of Applied Microbiology*, 101(3):514–525.

[145] Shaner, N. C., Campbell, R. E., Steinbach, P. A., Giepmans, B. N. G., Palmer, A. E., and Tsien, R. Y. (2004). Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nature Biotechnology*, 22(12):1567–1572.

[146] Shcherbo, D., Murphy, C. S., Ermakova, G. V., Solovieva, E. A., Chepurnykh, T. V., Shcheglov, A. S., Verkhusha, V. V., Pletnev, V. Z., Hazelwood, K. L., Roche, P. M., Lukyanov, S., Zaraisky, A. G., Davidson, M. W., and Chudakov, D. M. (2009). Far-red fluorescent tags for protein imaging in living tissues. *The Biochemical journal*, 418(3):567–74.

[147] Shi, J., Vlamis-Gardikas, A., Aslund, F., Holmgren, A., and Rosen, B. P. (1999). Reactivity of glutaredoxins 1, 2, and 3 from Escherichia coli shows that glutaredoxin 2 is the primary hydrogen donor to ArsC-catalyzed arsenate reduction. *The Journal of biological chemistry*, 274(51):36039–42.

[148] Shi, W., Dong, J., Scott, R. A., Ksenzenko, M. Y., and Rosen, B. P. (1996). The Role of Arsenic-Thiol Interactions in Metalloregulation of the ars Operon. *Journal of Biological Chemistry*, 271(16):9291–9297.

[149] Siegfried, K., Endes, C., Bhuiyan, A. F. M. K., Kuppardt, A., Mattusch, J., van der Meer, J. R., Chatzinotas, A., and Harms, H. (2012). Field Testing of Arsenic in Groundwater Samples of Bangladesh Using a Test Kit Based on Lyophilized Bioreporter Bacteria. *Environmental Science & Technology*, 46(6):3281–3287.

[150] Singh, S., Mulchandani, A., and Chen, W. (2008). Highly selective and rapid arsenic removal by metabolically engineered Escherichia coli cells expressing Fucus vesiculosus metallothionein. *Applied and environmental microbiology*, 74(9):2924–7.

[151] Smith, A. H., Lingas, E. O., and Rahman, M. (2000). Contamination of drinking-water by arsenic in Bangladesh: a public health emergency. *Bulletin of the World Health Organization*, 78(9):1093–103.

[152] Smith, M. T., Wilding, K. M., Hunt, J. M., Bennett, A. M., and Bundy, B. C. (2014). The emerging age of cell-free synthetic biology. *FEBS Letters*, 588(17):2755–2761.

[153] Stocker, J., Balluch, D., Gsell, M., Harms, H., Feliciano, J., Daunert, S., Malik, K. A., and van der Meer, J. R. (2003). Development of a Set of Simple Bacterial Biosensors for Quantitative and Rapid Measurements of Arsenite and Arsenate in Potable Water.

[154] Stolz, J. F., Ellis, D. J., Switzer, J., Ahmann, D., Lovley, D. R., and Oremland, R. S. (1999). Sulfurospirillum barnesii. *International journal of systematic bacteriology*, 49(1 999):1177–1180.

[155] Thelwell, C., Robinson, N. J., and Turner-Cavet, J. S. (1998). An SmtB-like repressor from Synechocystis PCC 6803 regulates a zinc exporter. *Proceedings of the National Academy of Sciences*, 95(18):10728–10733.

[156] Tsai, K. J., Yoon, K. P., and Lynn, A. R. (1992). ATP-dependent cadmium transport by the cadA cadmium resistance determinant in everted membrane vesicles of Bacillus subtilis. *Journal of bacteriology*, 174(1):116–21.

[157] Turner, A. (1949). Bacterial Oxidation of Arsenite. *Nature*, 164:76–77.

[158] van Geen, A., Cheng, Z., Seddique, A. A., Hoque, M. A., Gelman, A., Graziano, J. H., Ahsan, H., Parvez, F., and Ahmed, K. M. (2004). Reliability of a Commercial Kit To Test Groundwater for Arsenic in Bangladesh.

[159] vanden Hoven, R. N. and Santini, J. M. (2004). Arsenite oxidation by the heterotroph Hydrogenophaga sp. str. NT-14: the arsenite oxidase and its physiological electron acceptor. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1656(2-3):148–155.

[160] VanZile, M. L., Chen, X., and Giedroc, D. P. (2002a). Allosteric Negative Regulation of smt O/P Binding of the Zinc Sensor, SmtB, by Metal Ions: A Coupled Equilibrium Analysis†.

[161] VanZile, M. L., Chen, X., and P., G. D. (2002b). Structural Characterization of Distinct $\alpha$3N and $\alpha$5 Metal Sites in the Cyanobacterial Zinc Sensor SmtB.

[162] Vellanoweth, R. L. and Rabinowitz, J. C. (1992). The influence of ribosome-binding-site elements on translational efficiency in Bacillus subtilis and Escherichia coli in vivo. *Molecular Microbiology*, 6(9):1105–1114.

[163] Vlamis-Gardikas, A., Aslund, F., Spyrou, G., Bergman, T., and Holmgren, A. (1997). Cloning, overexpression, and characterization of glutaredoxin 2, an atypical glutaredoxin from Escherichia coli. *The Journal of biological chemistry*, 272(17):11236–43.

[164] Wang, W., Li, G.-W., Chen, C., Xie, X. S., and Zhuang, X. (2011). Chromosome organization by a nucleoid-associated protein in live bacteria. *Science (New York, N.Y.)*, 333(6048):1445–9.

[165] Welch, A. H., Westjohn, D., Helsel, D. R., and Wanty, R. B. (2000). Arsenic in Ground Water of the United States: Occurrence and Geochemistry. *Ground Water*, 38(4):589–604.

[166] Willis, S. S., Haque, S. E., and Johannesson, K. H. (2011). Arsenic and Antimony in Groundwater Flow Systems: A Comparative Study. *Aquatic Geochemistry*, 17(6):775–807.

[167] Willsky, G. R. and Malamy, M. H. (1980a). Characterization of two genetically separable inorganic phosphate transport systems in Escherichia coli. *Journal of bacteriology*, 144(1):356–65.

[168] Willsky, G. R. and Malamy, M. H. (1980b). Effect of arsenate on inorganic phosphate transport in Escherichia coli. *Journal of bacteriology*, 144(1):366–74.

[169] Wu, J. and Rosen, B. P. (1993a). Metalloregulated expression of the ars operon. *The Journal of biological chemistry*, 268(1):52–8.

[170] Wu, J. and Rosen, B. P. (1993b). The arsD gene encodes a second trans-acting regulatory protein of the plasmid-encoded arsenical resistance operon. *Molecular Microbiology*, 8(3):615–623.

[171] Wu, J., Tisa, L. S., and Rosen, B. P. (1992). Membrane topology of the ArsB protein, the membrane subunit of an anion-translocating ATPase. *Journal of Biological Chemistry*, 267(18):12570–12576.

[172] Wysocki, R., Chery, C. C., Wawrzycka, D., Van Hulle, M., Cornelis, R., Thevelein, J. M., and Tamas, M. J. (2001). The glycerol channel Fps1p mediates the uptake of arsenite and antimonite in Saccharomyces cerevisiae. *Molecular Microbiology*, 40(6):1391–1401.

[173] Xu, C. and Rosen, B. P. (1997). Dimerization is essential for DNA binding and repression by the ArsR metalloregulatory protein of Escherichia coli. *The Journal of biological chemistry*, 272(25):15734–8.

[174] Xu, C., Shi, W., and Rosen, B. P. (1996). The Chromosomal arsR Gene of Escherichia coli Encodes a trans-acting Metalloregulatory Protein. *Journal of Biological Chemistry*, 271(5):2427–2432.

[175] Xu, C., Zhou, T., Kuroda, M., and Rosen, B. P. (1998). Metalloid resistance mechanisms in prokaryotes. *Journal of biochemistry*, 123(1):16–23.

[176] Yamamura, S. and Amachi, S. (2014). Microbiology of inorganic arsenic: From metabolism to bioremediation. *Journal of Bioscience and Bioengineering*, 118(1):1–9.

[177] Yordanov, B., Dalchau, N., Grant, P. K., Pedersen, M., Emmott, S., Haseloff, J., and Phillips, A. (2014). A computational method for automated characterization of genetic components. *ACS synthetic biology*, 3(8):578–88.

[178] Young, J. W., Locke, J. C. W., Altinok, A., Rosenfeld, N., Bacarian, T., Swain, P. S., Mjolsness, E., and Elowitz, M. B. (2012). Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nature protocols*, 7(1):80–8.

[179] Zeigler, D. R., Prágai, Z., Rodriguez, S., Chevreux, B., Muffler, A., Albert, T., Bai, R., Wyss, M., and Perkins, J. B. (2008). The origins of 168, W23, and other Bacillus subtilis legacy strains. *Journal of bacteriology*, 190(21):6983–95.

[180] Zhang, Y.-N., Sun, G.-X., Williams, P. N., Huang, Q., and Zhu, Y.-G. (2011). Assessment of the solubility and bioaccessibility of arsenic in realgar wine using a simulated gastrointestinal system. *Science of The Total Environment*, 409(12):2357–2360.

# Appendix A

# Influence of Subsequent Codon on Codon Bias

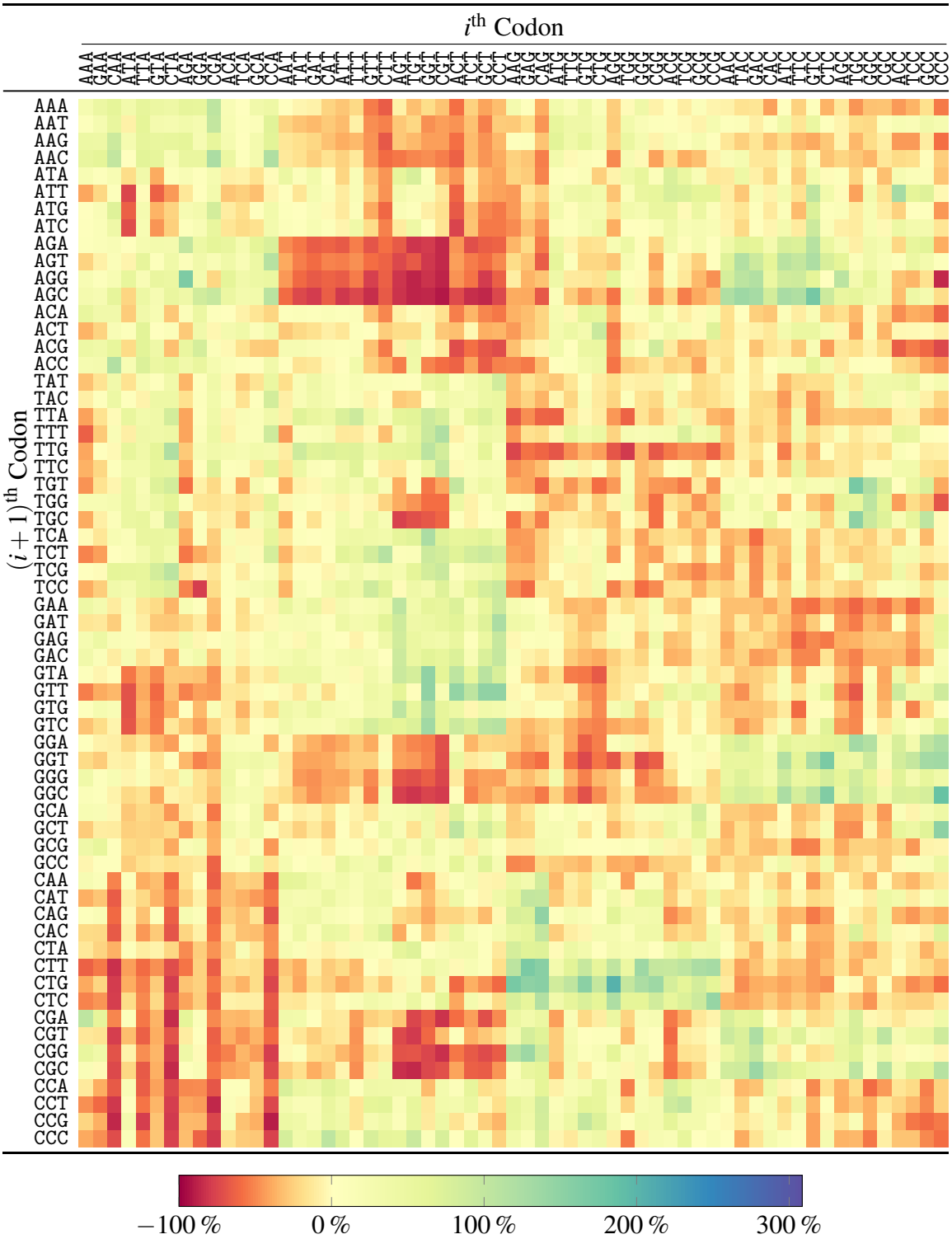



Table A.1 Percentage change in codon preference due to next codon. See Table 6.3 and surrounding text in Section 6.3. (∗ & †) See Table 6.2

# Appendix B

# Sequence of the $P_{ars}$ promoter

The full seqneuce of the *ars* operon promoter which was extracted from the genome by PCR was as follows.

TTTCGTCTCTCGTAATTGTTTAAAACCTTATGTGGAAGGA
TGGATTGAATGGCTTATATCTTCACATAAGGTTTTAATTG
ATGATAAATCTGTGTGCTCTGACTAAATACGTATTATCTC
GAATTATTATCGATTCTACATATTCACAAAAATTATTTAA
TGTTTTTTTCTTGAAATTAGAAATTAGCGTTTATGATCTC
CGTTGCTGTAGTAGCAAAGTAGCGATGGAGGAAATAATAA
AAATGGAAAACCTCTATCTGGTTACACAAGCTTTTAGCGG
GTTTTACAATTAATCAAAATAAATTGATTTATTTGCTTGC
ATTAATTTAAAAATCATGAGTATAATAAATACATCAGACA
GAGACGAAA

# Appendix C

# The *pysim* model file format

The pysim model file defines the mathematical model to be simulated. Examples can be found in the folder `pysim/tests/data`, which are also used to unit test the implementation.

For example, the file `pysim/tests/data/solvetest3.model` contains the following

```
#model without arsenic
#p = free promoter
#r = arsR
#s = arsR2
#t = bound promoter

species p=1.0,q,r,s
constraint q + p - 1.0
param kp=4.4721359549995796, mu=0.22360679774997896
param kd_f=640.31242374328485, kd_r=0.0015617376188860608
param kr_f=8246.2112512353215, kr_r=0.00012126781251816648

p --[kp]> p + r
r --[mu]>
2r <[kd_r]--[kd_f]> s
s --[mu]>
s + p <[kr_r]--[kr_f]> q
q --[mu]> p
```

Lines beginning with '#' are comments and are ignored. The first line defines the species present, with optional initial conditions in arbitrary units. If no initial condition is given, it is assumed to be zero. Species can be specified all on one line separated by commas or

on separate lines. The second line defines a constraint, in this case that the sum of $p$ and $q$ must be 1 at all times. Multiple constraints can be specified, but must be on different lines. The next three lines define the parameters to the model in a manner similar to the species, however all parameters must be given values.

The remaining lines define the reactions present in the network. The values in the square brackets show the rates of each reaction, and can potentially include simple mathematical statements. A Michaelis Menten reaction is shown by providing two comma separated parameters in the brackets, the first being the limiting rate, $V_{max}$, and the second the Menten constant, $K_M$.

# Appendix D

# Plasmid Listings

Sequences for the two template plasmids developed during the project are shown below.

## D.1 pHK025v2

```
LOCUS       pHK025v2                5545 bp ds-DNA     circular     15-FEB-2017
DEFINITION  .
FEATURES             Location/Qualifiers
     deleted_base    complement(1060..1060)
                     /label="Deleted Base"
                     /ApEinfo_revcolor=#f58a5e
                     /ApEinfo_fwdcolor=#f58a5e
     RBS             complement(2710..2715)
                     /label="cat RBS"
                     /ApEinfo_revcolor=#f8d3a9
                     /ApEinfo_fwdcolor=#f8d3a9
     terminator      complement(2007..2038)
                     /label="cat term?"
                     /ApEinfo_revcolor=#ff9ccd
                     /ApEinfo_fwdcolor=#ff9ccd
     terminator      2901..2935
                     /label="t0 terminator of phage lambda"
                     /ApEinfo_revcolor=#ff9ccd
                     /ApEinfo_fwdcolor=#ff9ccd
     CDS             complement(2051..2701)
                     /label="cat"
                     /ApEinfo_revcolor=#84b0dc
                     /ApEinfo_fwdcolor=#84b0dc
     deleted_base    complement(1064..1064)
                     /label="Deleted Base"
                     /ApEinfo_revcolor=#f58a5e
                     /ApEinfo_fwdcolor=#f58a5e
     CDS             complement(3854..4374)
                     /label="amyE (front)"
                     /ApEinfo_revcolor=#84b0dc
                     /ApEinfo_fwdcolor=#84b0dc
```

```
     misc_difference 3361..3361
                     /label="Mut: C–>T (removing PstI site)"
                     /ApEinfo_revcolor=#faac61
                     /ApEinfo_fwdcolor=#faac61
     modified_base   complement(5243..5243)
                     /label="modified_base"
                     /ApEinfo_revcolor=#ffef86
                     /ApEinfo_fwdcolor=#ffef86
     misc_feature    3732..3811
                     /label="BBa_B0010"
                     /ApEinfo_revcolor=#b4abac
                     /ApEinfo_fwdcolor=#b4abac
     misc_feature    complement(1980..2866)
                     /label="cat cassette"
                     /ApEinfo_revcolor=#84b0dc
                     /ApEinfo_fwdcolor=#84b0dc
     misc_feature    4524..5384
                     /label="bla"
                     /ApEinfo_revcolor=#b7e6d7
                     /ApEinfo_fwdcolor=#b7e6d7
     BSub dif        complement(2867..2894)
                     /label="dif"
                     /ApEinfo_revcolor=#f58a5e
                     /ApEinfo_fwdcolor=#f58a5e
     BSub dif        complement(1952..1979)
                     /label="dif"
                     /ApEinfo_revcolor=#c7b0e3
                     /ApEinfo_fwdcolor=#c7b0e3
     rep_origin      1..600
                     /label="Col E1 ori"
                     /ApEinfo_revcolor=#75c6a9
                     /ApEinfo_fwdcolor=#75c6a9
     deleted_base    complement(1014..1014)
                     /label="Deleted Base"
                     /ApEinfo_revcolor=#f58a5e
                     /ApEinfo_fwdcolor=#f58a5e
     stem_loop       3743..3786
                     /label="stem_loop"
                     /ApEinfo_revcolor=#b1ff67
                     /ApEinfo_fwdcolor=#b1ff67
     misc_feature    3010..3720
                     /label="mCherry"
                     /ApEinfo_revcolor=#d59687
                     /ApEinfo_fwdcolor=#d59687
     misc_feature    4395..4523
                     /label="bla_promoter&UTR"
                     /ApEinfo_revcolor=#c6c9d1
                     /ApEinfo_fwdcolor=#c6c9d1
     CDS             complement(620..1648)
                     /label="amyE (back)"
                     /ApEinfo_revcolor=#84b0dc
                     /ApEinfo_fwdcolor=#84b0dc
ORIGIN
        1 AGGATCTTCT TGAGATCCTT TTTTTCTGCG CGTAATCTGC TGCTTGCAAA CAAAAAAACC
       61 ACCGCTACCA GCGGTGGTTT GTTTGCCGGA TCAAGAGCTA TCAACTCTTT TTCCGAAGGT
```

```
 121 AACTGGCTTC AGCAGAGCGC AGATACCAAA TACTGTCCTT CTAGTGTAGC CGTAGTTAGG
 181 CCACCACTTC AAGAACTCTG TAGCACCGCC TACATACCTC GCTCTGCTAA TCCTGTTACC
 241 AGTGGCTGCT GCCAGTGGCG ATAAGTCGTG TCTTACCGGG TTGGACTCAA GACGATAGTT
 301 ACCGGATAAG GCGCAGCGGT CGGGCTGAAC GGGGGGTTCG TGCACACAGC CCAGCTTGGA
 361 GCGAACGACC TACACCGAAC TGAGATACCT ACAGCGTGAG CTATGAGAAA GCGCCACGCT
 421 TCCCGAAGGG AGAAAGGCGG ACAGGTATCC GGTAAGCGGC AGGGTCGGAA CAGGAGAGCG
 481 CACGAGGGAG CTTCCAGGGG GAAACGCCTG GTATCTTTAT AGTCCTGTCG GGTTTCGCCA
 541 CCTCTGACTT GAGCGTCGAT TTTTGTGATG CTCGTCAGGG GGGCGGAGCC TATGGAAAAA
 601 CGCCAGCAAC GCGGCCCGAT CAATGGGGAA GAGAACCGCT TAAGCCCGAG TCATTATATA
 661 AACCATTTAG CACGTAATCA AAGCCAGGCT GATTCTGACC GGGCACTTGG GCGCTGCCAT
 721 TATTAAAAAT CACTTTTGCG TTGGTTGTAT CCGTGTCCGC AGGCAGCGTC AGCGTGTAAA
 781 TTCCGTCTGC ATTTTTAGTC ATTGGTTTTC CAGGCCAAGA TCCGGTCAAT TCAATTACTC
 841 GGCTCCCATC ATGTTTATAG ATATAAGCAT TTACCTGGCT CCAATGATTC GGATTTTGAT
 901 AGCCGATGGT TTTGGCCGAC GCTGGATCTC TTTTAACAAA ACTGTATTTC TCGGTCCTCG
 961 TTACACCATC ACTGTTCGTT CCTTTTAACA TGATGGTGTA TGTTTTGCCA AATTGGATCT
1021 CCTTTTCCGA TTGTGAATTG ATCTCCATCC TTAAACGCCT GTCGTCTGGT CCATTATTGA
1081 TTTGATAAAC GGCTTTTGTT GTATTCGCAT CTGCACGCAA GGTAATCGTC AGTTGATCAT
1141 TGAAAGAATG TGTTACACCT GTTTTGTAAT TCTCAAGGAA AACATGAGGC GCTTTTGCAA
1201 TATCATCAGG ATAAAGCACA GCTACAGACC TGGCATTGAT CGTGCCTGTC AGTTTACCAT
1261 CGTTCACTTG AAATGAACCC GCTCCAGCTT TATTGTCATA CCTGCCATCA GGCAATTTTG
1321 TTGCCGTATT GATAGAGACA GAGGATGAAC CTGCATTTGC CAGCACAACG CCATGTGAGC
1381 CGCGCTGATT CATAAATATC TGGTTGTTTC CATTCGGGTT CGAGAGTTCC TCAGGCTGTC
1441 CAGCCATCAC ATTGTGAAAT CTATTGACCG CAGTGATAGC CTGATCTTCA AATAAAGCAC
1501 TCCCGCGATC GCCTATTTGG CTTTTCCCCG GGAACCTCAC ACCATTTCCG CCTCCCTCAG
1561 GTCTGGAAAA GAAAAGAGGC GTACTGCCTG AACGAGAAGC TATCACCGCC CAGCCTAAAC
1621 GGATATCATC ATCGCTCATC CATGTCGACG CTCTCCCTTA TGCGACTCCT GCATTAGGAA
1681 GCAGCCCAGT AGTAGGTTGA GGCCGTTGAG CACCGCCGCC GCAAGGAATG GTGCATGCAA
1741 GGAGATGGCG CCCAACAGTC CCCCGGCCAC GGGGCCTGCC ACCATACCCA CGCCGAAACA
1801 AGCGCTCATG AGCCCGAAGT GGCGAGCCCG ATCTTCCCCA TCGGTGATGT CGGCGATATA
1861 GGCGCCAGCA ACCGCACCTG TGGCGCCGGT GATGCCGGCC ACGATGCGTC CGGCGTAGAG
1921 GATCTGGAGC TGTAATATAA AAACCTTCGT AAGTTTACAT AATATATATT CTAGGAAGTC
1981 TTCAACTAAC GGGGCAGGTT AGTGACATTA GAAAACCGAC TGTAAAAAGT ACAGTCGGCA
2041 TTATCTCATA TTATAAAAGC CAGTCATTAG GCCTATCTGA CAATTCCTGA ATAGAGTTCA
2101 TAAACAATCC TGCATGATAA CCATCACAAA CAGAATGATG TACCTGTAAA GATAGCGGTA
2161 AATATATTGA ATTACCTTTA TTAATGAATT TTCCTGCTGT AATAATGGGT AGAAGGTAAT
2221 TACTATTATT ATTGATATTT AAGTTAAACC CAGTAAATGA AGTCCATGGA ATAATAGAAA
2281 GAGAAAAAGC ATTTTCAGGT ATAGGTGTTT TGGGAAACAA TTTCCCCGAA CCATTATATT
2341 TCTCTACATC AGAAAGGTAT AAATCATAAA ACTCTTTGAA GTCATTCTTT ACAGGAGTCC
2401 AAATACCAGA GAATGTTTTA GATACACCAT CAAAAATTGT ATAAAGTGGC TCTAACTTAT
2461 CCCAATAACC TAACTCTCCG TCGCTATTGT AACCAGTTCT AAAAGCTGTA TTTGAGTTTA
2521 TCACCCTTGT CACTAAGAAA ATAAATGCAG GGTAAAATTT ATATCCTTCT TGTTTTATGT
2581 TTCGGTATAA AACACTAATA TCAATTTCTG TGGTTATACT AAAAGTCGTT TGTTGGTTCA
2641 AATAATGATT AAATATCTCT TTTCTCTTCC AATTGTCTAA ATCAATTTTA TTAAAGTTCA
2701 TTTGATATGC CTCCTAAATT TTTATCTAAA GTGAATTTAG GAGGCTTACT TGTCTGCTTT
2761 CTTCATTAGA ATCAATCCTT TTTTAAAAGT CAATATTACT GTAACATAAA TATATATTTT
2821 AAAAATATCC CACTTTATCC AATTTTCGTT TGTTGAACTA ATGGGCAGTT TACATAATAT
2881 ATATTCTAGG AAGTACCTTT TTCAGAACGC TCGGTTGCCG CCGGGCGTTT TTTATTTGAA
2941 TAACTAATGT GAGTGTGAGA GTGTAAATCT GCGTATGAGA CGGCTCTGTG ACTATTGTGT
3001 ATCGTCTCTA TGGtgagcaa gggcgaggag gataacatgg ccatcatcaa ggagttcatg
3061 cgcttcaagg tgcacatgga gggctccgtg aacggccacg agttcgagat cgagggcgag
3121 ggcgagggcc gccctacga gggcacccag accgccaagc tgaaggtgac caagggtggc
3181 cccctgccct tcgcctggga catcctgtcc cctcagttca tgtacggctc caaggcctac
3241 gtgaagcacc ccgccgacat ccccgactac ttgaagctgt ccttccccga gggcttcaag
3301 tgggagcgcg tgatgaactt cgaggacggc ggcgtggtga ccgtgaccca ggactcctcc
3361 ttgcaggacg gcgagttcat ctacaaggtg aagctgcgcg gcaccaactt ccctccgac
```

```
     3421 ggccccgtaa tgcagaagaa gaccatgggc tgggaggcct cctccgagcg gatgtacccc
     3481 gaggacggcg ccctgaaggg cgagatcaag cagaggctga agctgaagga cggcggccac
     3541 tacgacgctg aggtcaagac cacctacaag gccaagaagc ccgtgcagct gcccggcgcc
     3601 tacaacgtca acatcaagtt ggacatcacc tcccacaacg aggactacac catcgtggaa
     3661 cagtacgaac gcgccgaggg ccgccactcc accggcggca tggacgagct gtacaagtaa
     3721 taatactaga gccaggcatc aaataaaacg aaaggctcag tcgaaagact gggcctttcg
     3781 ttttatctgt tgtttgtcgg tgaacgctct ctactagagt cacactggct caccttcgCG
     3841 TAGGTCTCAA CGGCGATCAG ACCAGTTTTT AATTTGTGTG TTTCCATGTG TCCAGTTTGG
     3901 AATACTCTTA ACCTCATTGG AAATCGCGGC ATAATCACTG GTGGTATGAT TGATGACCGC
     3961 GTCAACAATG ACCTTTATGC CATATTCTTC AGCGGCTGCA CACATTTCTT TAAATTCTTG
     4021 TTCAGTACCT AAGTAACGGT TGCCAATTTG ATACGATGTC GGCTGATACA GCCAGTACCA
     4081 GTTCGACATG CTTTTATCTC CTTGATTCCC TTCCTTTACT TGGTTAATCG GAGATGTCTG
     4141 AATGGCTGTA TATCCTGCAT CATGAATATC CTTCATATTG TGTTTTAACG TATTGAACGA
     4201 CCAATTCCAT GCATGAAGAA TGGTTCCGCT TTTGATCGAC GGTGCTGTAA GCTCATTCGA
     4261 TTTGTTCGCC GTTTCAGCAC TCGCAGCCGC CGGTCCTGCC AGAACCAAAT GAAACAGCAA
     4321 TAAAAATCCA GCGAATAACG GCAGTAAAGA GGTTTTGAAT CGTTTTGCAA ACATTCTTGA
     4381 CACTCCTTAG TTCAGGTGGC ACTTTTCGGG GAAATGTGCG CGGAACCCCT ATTTGTTTAT
     4441 TTTTCTAAAT ACATTCAAAT ATGTATCCGC TCATGAGACA ATAACCCTGA TAAATGCTTC
     4501 AATAATATTG AAAAAGGAAG AGTATGAGTA TTCAACATTT CCGTGTCGCC CTTATTCCCT
     4561 TTTTTGCGGC ATTTTGCCTT CCTGTTTTTG CTCACCCAGA AACGCTGGTG AAAGTAAAAG
     4621 ATGCTGAAGA TCAGTTGGGT GCACGAGTGG GTTACATCGA ACTGGATCTC AACAGCGGTA
     4681 AGATCCTTGA GAGTTTTCGC CCCGAAGAAC GTTTTCCAAT GATGAGCACT TTTAAAGTTC
     4741 TGCTATGTGG CGCGGTATTA TCCCGTGTTG ACGCCGGGCA AGAGCAACTC GGTCGCCGCA
     4801 TACACTATTC TCAGAATGAC TTGGTTGAGT ACTCACCAGT CACAGAAAAG CATCTTACGG
     4861 ATGGCATGAC AGTAAGAGAA TTATGCAGTG CTGCCATAAC CATGAGTGAT AACACTGCGG
     4921 CCAACTTACT TCTGACAACG ATCGGAGGAC CGAAGGAGCT AACCGCTTTT TTGCACAACA
     4981 TGGGGGATCA TGTAACTCGC CTTGATCGTT GGGAACCGGA GCTGAATGAA GCCATACCAA
     5041 ACGACGAGCG TGACACCACG ATGCCTGCAG CAATGGCAAC AACGTTGCGC AAACTATTAA
     5101 CTGGCGAACT ACTTACTCTA GCTTCCCGGC AACAATTAAT AGACTGGATG GAGGCGGATA
     5161 AAGTTGCAGG ACCACTTCTG CGCTCGGCCC TTCCGGCTGG CTGGTTTATT GCTGATAAAT
     5221 CTGGAGCCGG TGAGCGTGGG TCCCGCGGTA TCATTGCAGC ACTGGGGCCA GATGGTAAGC
     5281 CCTCCCGTAT CGTAGTTATC TACACGACGG GGAGTCAGGC AACTATGGAT GAACGAAATA
     5341 GACAGATCGC TGAGATAGGT GCCTCACTGA TTAAGCATTG GTAACTGTCA GACCAAGTTT
     5401 ACTCATATAT ACTTTAGATT GATTTAAAAC TTCATTTTTA ATTTAAAAGG ATCTAGGTGA
     5461 AGATCCTTTT TGATAATCTC ATGACCAAAA TCCCTTAACG TGAGTTTTCG TTCCACTGAG
     5521 CGTCAGACCC CGTAGAAAAG ATCAA
//
```

# D.2 pHK026

```
LOCUS       pHK026                  5492 bp ds-DNA     circular     10-APR-2017
DEFINITION  .
FEATURES             Location/Qualifiers
     terminator      2846..2880
                     /label="t0 terminator of phage lambda"
                     /ApEinfo_revcolor=#ff9ccd
                     /ApEinfo_fwdcolor=#ff9ccd
     CDS             4968..5488
                     /label="amyE (front)"
                     /ApEinfo_revcolor=#84b0dc
                     /ApEinfo_fwdcolor=#84b0dc
     rep_origin      complement(48..647)
                     /label="Col E1 ori"
                     /ApEinfo_revcolor=#75c6a9
```

```
                     /ApEinfo_fwdcolor=#75c6a9
     modified_base   3629..3629
                     /label="modified_base"
                     /ApEinfo_revcolor=#d59687
                     /ApEinfo_fwdcolor=#d59687
     CDS             4129..4911
                     /label="spc"
                     /ApEinfo_revcolor=#84b0dc
                     /ApEinfo_fwdcolor=#84b0dc
     source          3958..4945
                     /label="spec cassette"
                     /ApEinfo_revcolor=#c6c9d1
                     /ApEinfo_fwdcolor=#c6c9d1
     RBS             4115..4121
                     /label="spc RBS?"
                     /ApEinfo_revcolor=#f8d3a9
                     /ApEinfo_fwdcolor=#f8d3a9
     misc_feature    complement(949..954)
                     /label="misc_feature"
                     /ApEinfo_revcolor=#ffef86
                     /ApEinfo_fwdcolor=#ffef86
     CDS             complement(809..1669)
                     /label="bla"
                     /ApEinfo_revcolor=#84b0dc
                     /ApEinfo_fwdcolor=#84b0dc
     misc_feature    1905..2814
                     /label="amyE (upstream)"
                     /ApEinfo_revcolor=#b4abac
                     /ApEinfo_fwdcolor=#b4abac
     misc_feature    2840..2958
                     /label="landing pad"
                     /ApEinfo_revcolor=#b1ff67
                     /ApEinfo_fwdcolor=#b1ff67
     misc_feature    2955..2978
                     /label="comGA"
                     /ApEinfo_revcolor=#85dae9
                     /ApEinfo_fwdcolor=#85dae9
     misc_feature    2979..3698
                     /label="mVenus"
                     /ApEinfo_revcolor=#9eafd2
                     /ApEinfo_fwdcolor=#9eafd2
     terminator      3710..3789
                     /label="B0010 terminator"
                     /ApEinfo_revcolor=#ff9ccd
                     /ApEinfo_fwdcolor=#ff9ccd
ORIGIN
        1 CATTTTACCA ATCCTGAAAC GGCTGGAATC GGGCCGCGTT GCTGGCGTTT TTCCATAGGC
       61 TCCGCCCCCC TGACGAGCAT CACAAAAATC GACGCTCAAG TCAGAGGTGG CGAAACCCGA
      121 CAGGACTATA AAGATACCAG GCGTTTCCCC CTGGAAGCTC CCTCGTGCGC TCTCCTGTTC
      181 CGACCCTGCC GCTTACCGGA TACCTGTCCG CCTTTCTCCC TTCGGGAAGC GTGGCGCTTT
      241 CTCATAGCTC ACGCTGTAGG TATCTCAGTT CGGTGTAGGT CGTTCGCTCC AAGCTGGGCT
      301 GTGTGCACGA ACCCCCCGTT CAGCCCGACC GCTGCGCCTT ATCCGGTAAC TATCGTCTTG
      361 AGTCCAACCC GGTAAGACAC GACTTATCGC CACTGGCAGC AGCCACTGGT AACAGGATTA
      421 GCAGAGCGAG GTATGTAGGC GGTGCTACAG AGTTCTTGAA GTGGTGGCCT AACTACGGCT
      481 ACACTAGAAG GACAGTATTT GGTATCTGCG CTCTGCTGAA GCCAGTTACC TTCGGAAAAA
```

```
 541 GAGTTGATAG CTCTTGATCC GGCAAACAAA CCACCGCTGG TAGCGGTGGT TTTTTTGTTT
 601 GCAAGCAGCA GATTACGCGC AGAAAAAAAG GATCTCAAGA AGATCCTTTG ATCTTTTCTA
 661 CGGGGTCTGA CGCTCAGTGG AACGAAAACT CACGTTAAGG GATTTTGGTC ATGAGATTAT
 721 CAAAAAGGAT CTTCACCTAG ATCCTTTTAA ATTAAAAATG AAGTTTTAAA TCAATCTAAA
 781 GTATATATGA GTAAACTTGG TCTGACAGTT ACCAATGCTT AATCAGTGAG GCACCTATCT
 841 CAGCGATCTG TCTATTTCGT TCATCCATAG TTGCCTGACT CCCCGTCGTG TAGATAACTA
 901 CGATACGGGA GGGCTTACCA TCTGGCCCCA GTGCTGCAAT GATACCGCGA GACCCACGCT
 961 CACCGGCTCC AGATTTATCA GCAATAAACC AGCCAGCCGG AAGGGCCGAG CGCAGAAGTG
1021 GTCCTGCAAC TTTATCCGCC TCCATCCAGT CTATTAATTG TTGCCGGGAA GCTAGAGTAA
1081 GTAGTTCGCC AGTTAATAGT TTGCGCAACG TTGTTGCCAT TGCTGCAGGC ATCGTGGTGT
1141 CACGCTCGTC GTTTGGTATG GCTTCATTCA GCTCCGGTTC CCAACGATCA AGGCGAGTTA
1201 CATGATCCCC CATGTTGTGC AAAAAAGCGG TTAGCTCCTT CGGTCCTCCG ATCGTTGTCA
1261 GAAGTAAGTT GGCCGCAGTG TTATCACTCA TGGTTATGGC AGCACTGCAT AATTCTCTTA
1321 CTGTCATGCC ATCCGTAAGA TGCTTTTCTG TGACTGGTGA GTACTCAACC AAGTCATTCT
1381 GAGAATAGTG TATGCGGCGA CCGAGTTGCT CTTGCCCGGC GTCAACACGG GATAATACCG
1441 CGCCACATAG CAGAACTTTA AAAGTGCTCA TCATTGGAAA ACGTTCTTCG GGGCGAAAAC
1501 TCTCAAGGAT CTTACCGCTG TTGAGATCCA GTTCGATGTA ACCCACTCGT GCACCCAACT
1561 GATCTTCAGC ATCTTTTACT TTCACCAGCG TTTCTGGGTG AGCAAAAACA GGAAGGCAAA
1621 ATGCCGCAAA AAAGGGAATA AGGGCGACAC GGAAATGTTG AATACTCATA CTCTTCCTTT
1681 TTCAATATTA TTGAAGCATT TATCAGGGTT ATTGTCTCAT GAGCGGATAC ATATTTGAAT
1741 GTATTTAGAA AAATAAACAA ATAGGGGTTC CGCGCACATT TCCCCGAAAA GTGCCACCTG
1801 ACGTCTAAGA AACCATTATT ATCATGACAT TAACCTATAA AAATAGGCGT ATCACtgaaT
1861 GTTGCTAGCC TTTATGGCGG TCATCATTGT TTATCCTCTC CCCGACTGTG TGAACCCGAC
1921 ATCCGGCGTT CTCATGGCGG TGCTTGCCGC CAGCGGTATT CCGTATGTCA AGTGGCTGCG
1981 GTTTATGGTG CCGCTTGCTC TGATTTGGTT CTTGATCGGG CTTGTCTTTA TCGTGATCGG
2041 AGTCATGATC AATTGGGGGC CGTTTTAACG ATTGCTGCCC GCCGGCTTGT ACGGCGGGCT
2101 TTTGAGTTAT TCATTGCAGA AGCGCAGGCT GTTATTGTAA CATGTAAGCC ATAAGCCATT
2161 CGTAAAAGTG CGGGAGGAAG GTCATGAATA ATCTGCGTAA TAGACTTTCA GGCGTGAATG
2221 GGAAAAATAA GAGAGTAAAA GAAAAAGAAC AAAAAATCTG GTCGGAGATT GGGATGATAG
2281 CGGGAGCATT TGCGCTGCTT GATGTGATCA TCCGCGGCAT TATGTTTGAA TTTCCGTTTA
2341 AAGAATGGGC TGCAAGCCTT GTGTTTTTGT TCATCATTAT CTTATATTAC TGCATCAGGG
2401 CTGCGGCATC CGGAATGCTC ATGCCGAGAA TAGACACCAA AGAAGAACTG CAAAAACGGG
2461 TGAAGCAGCA GCGAATAGAA TCAATTGCGG TCGCCTTTGC GGTAGTGGTG CTTACGATGT
2521 ACGACAGGGG GATTCCCCAT ACATTCTTCG CTTGGCTGAA AATGATTCTT CTTTTTATCG
2581 TCTGCGGCGG CGTTCTGTTT CTGCTTCGGT ATGTGATTGT GAAGCTGGCT TACAGAAGAG
2641 CGGTAAAAGA AGAAATAAAA AAGAAATCAT CTTTTTTGTT TGGAAAGCGA GGGAAGCGTT
2701 CACAGTTTCG GGCAGCTTTT TTTATAGGAA CATTGATTTG TATTCACTCT GCCAAGTTGT
2761 TTTGATAGAG TGATTGTGAT AATTTTAAAT GTAAGCGTTA ACAAAATTCT CCAGGAACGC
2821 GATTTCCAAT GAGGTTAAGa cctTTTTCAG AACGCTCGGT TGCCGCCGGG CGTTTTTTAT
2881 TTGAATAACT AATGTGAGTG TGAGAGTGTA AATCTGCGTA TGAGACGGCT CTGTGACTAT
2941 TGTGTATCGT CTCTATGGAT TCAATAGAAA AGGTAAGCAT GGTTAGTAAA GGAGAAGAAC
3001 TTTTCACTGG AGTTGTCCCA ATTTTAGTTG AACTAGATGG CGACGTGAAC GGTCATAAGT
3061 TCAGTGTCTC CGGCGAAGGT GAGGGTGATG CAACGTACGG TAAGTTAACT TTGAAGTTAA
3121 TATGTACAAC CGGCAAGCTG CCTGTTCCCT GGCCTACCCT GGTGACAACG TTAGGTTATG
3181 GGTTGATGTG CTTTGCTAGA TACCCAGATC ACATGAAAAG GCATGACTTC TTTAAATCTG
3241 CAATGCCAGA AGGTTACGTC CAAGAACGTA CTATTTTCTT TAAAGATGAC GGTAATTATA
3301 AAACTAGGGC TGAAGTTAAA TTCGAAGGTG ACACACTTGT AAATCGAATA GAGTTAAAGG
3361 GGATTGATTT CAAAGAGGAT GGTAATATTC TAGGCCATAA ACTTGAATAT AACTATAATT
3421 CACACAACGT TTACATTACC GCCGACAAGC AGAAGAATGG AATCAAAGCC AATTTTAAGA
3481 TTAGACACAA TATTGAGGAT GGTGGAGTAC AGCTTGCTGA TCATTACCAA CAAAATACCC
3541 CGATCGGTGA TGGACCAGTT TTGCTACCCG ATAACCATTA TCTGTCCTAT CAAAGCAAAT
3601 TGTCAAAAGA TCCTAACGAA AAAAGAGATC ACATGGTACT CTTGGAATTT GTAACAGCTG
3661 CTGGGATTAC ACATGGCATG GATGAACTAT ACAAATAATG ATACTAGAGC CAGGCATCAA
3721 ATAAAACGAA AGGCTCAGTC GAAAGACTGG GCCTTTCGTT TTATCTGTTG TTTGTCGGTG
3781 AACGCTCTCT ACTAGCCCAT TAGTTCAACA AACGAAAATT GGATAAAGTG GGATATTTTT
```

```
3841 AAAATATATA TTTATGTTAC AGTAATATTG ACTTTTAAAA AAGGATTGAT TCTAATGAAG
3901 AAAGCAGACA AGTAAGCCTC CGAACTGAGA CCGAACCGGT GAACGCTCTC TACTAGAGAT
3961 CTGTATAATA AAGAATAATT ATTAATCTGT AGACAAATTG TGAAAGGATG TACTTAAACG
4021 CTAACGGTCA GCTTTATTGA ACAGTAATTT AAGTATATGT CCAATCTAGG GTAAGTAAAT
4081 TGAGTATCAA TATAAACTTT ATATGAACAT AATCAACGAG GTGAAATCAT GAGCAATTTG
4141 ATTAACGGAA AAATACCAAA TCAAGCGATT CAAACATTAA AAATCGTAAA AGATTTATTT
4201 GGAAGTTCAA TAGTTGGAGT ATATCTATTT GGTTCAGCAG TAAATGGTGG TTTACGCATT
4261 AACAGCGATG TAGATGTTCT AGTCGTCGTG AATCATAGTT TACCTCAATT AACTCGAAAA
4321 AAACTAACAG AAAGACTAAT GACTATATCA GGAAAGATTG GAAATACGGA TTCTGTTAGA
4381 CCACTTGAAG TTACGGTTAT AAATAGGAGT GAAGTTGTCC CTTGGCAATA TCCTCCAAAA
4441 AGAGAATTTA TATACGGTGA GTGGCTCAGG GGTGAATTTG AGAATGGACA AATTCAGGAA
4501 CCAAGCTATG ATCCTGATTT GGCTATTGTT TTAGCACAAG CAAGAAAGAA TAGTATTTCT
4561 CTATTTGGTC CTGATTCTTC AAGTATACTT GTCTCCGTAC CTTTGACAGA TATTCGAAGA
4621 GCAATTAAGG ATTCTTTGCC AGAACTAATT GAGGGGATAA AAGGTGATGA GCGTAATGTA
4681 ATTTTAACCC TAGCTCGAAT GTGGCAAACA GTGACTACTG GTGAAATTAC CTCGAAAGAT
4741 GTCGCTGCAG AATGGGCTAT ACCTCTTTTA CCTAAAGAGC ATGTAACTTT ACTGGATATA
4801 GCTAGAAAAG GCTATCGGGG AGAGTGTGAT GATAAGTGGG AAGGACTATA TTCAAAGGTG
4861 AAAGCACTCG TTAAGTATAT GAAAAATTCT ATAGAAACTT CTCTCAATTA GGCTAATTTT
4921 ATTGCAATAA CAGGTGCTTA CTTTTTAGGT ACGAAGGAGT GTCAAGAATG TTTGCAAAAC
4981 GATTCAAAAC CTCTTTACTG CCGTTATTCG CTGGATTTTT ATTGCTGTTT CATTTGGTTC
5041 TGGCAGGACC GGCGGCTGCG AGTGCTGAAA CGGCGAACAA ATCGAATGAG CTTACAGCAC
5101 CGTCGATCAA AAGCGGAACC ATTCTTCATG CATGGAATTG GTCGTTCAAT ACGTTAAAAC
5161 ACAATATGAA GGATATTCAT GATGCAGGAT ATACAGCCAT TCAGACATCT CCGATTAACC
5221 AAGTAAAGGA AGGGAATCAA GGAGATAAAA GCATGTCGAA CTGGTACTGG CTGTATCAGC
5281 CGACATCGTA TCAAATTGGC AACCGTTACT TAGGTACTGA ACAAGAATTT AAAGAAATGT
5341 GTGCAGCCGC TGAAGAATAT GGCATAAAGG TCATTGTTGA CGCGGTCATC AATCATACCA
5401 CCAGTGATTA TGCCGCGATT TCCAATGAGG TTAAGAGTAT TCCAAACTGG ACACATGGAA
5461 ACACACAAAT TAAAAACTGG TCTGATCGAG GT
//
```

# Appendix E

# Primer Sequences

Sequences for all primers used during the project are shown in the table below. Physical DNA is available in the freezer boxes marked 'Haydn Primers I-V' stored at −20 °C. This information is also available in digital form at -

https://docs.google.com/spreadsheets/d/1cw5qsrBMo47zV6GzNeU1o2P9nLXBbZ2ZQIHnnJcSxoU

| Number | Name | Sequence |
|---|---|---|
| 3a | H1-spec_fw | AATATGCATCCTTTAACAAAAGAGTTTATTGATCTGTATAATAAAGAATAATTATTAATC |
| 4b | H1-spec_rv | GATTAATAATTATTCTTTATTATACAGATCAATAAACTCTTTTGTTAAAGGATGCATATT |
| 4a | spec-H2_fw | CAATAACAGGTGCTTACTTTTTTAATCAGACCTTTTCTTTTTGAATAAGT |
| 3b | spec-H2_rv | ACTTATTCAAAAAGAAAAGGTCTGATTAAAAAAGTAAGCACCTGTTATTG |
| 2a | H2-backbone_fw | ACTTTCTGGATGTTTTCAATTTGCTGGCGTTTTTCCATAG |
| 1b | H2-backbone_rv | CTATGGAAAAACGCCAGCAAATTGAAAACATCCAGAAAGT |
| 1a | backbone-H1_fw | CTTCAAAAAATCAAATAAGGAGTGTCAAGAATGATCGGTGAAATAGTTAAAAATCATC |
| 2b | backbone-H1_rv | GATGATTTTTAACTATTTCACCGATCATTCTTGACACTCCTTATTTGATTTTTTGAAG |
| s1 | pHK001-seq0 | GCGTATCACGAGGCCCTTT |
| s2 | pHK001-seq1 | CGATCCCTTGATCGCTTGTTC |

| | | |
|---|---|---|
| s3 | pHK001-seq2 | AGCAGCATCTCCACAAAGCG |
| s4 | pHK001-seq3 | GCGGACCAATTACGGCG |
| s5 | pHK001-seq4 | GAAAGCAACTAATCCTGCCGC |
| s6 | pHK001-seq5 | TATCACTTGGCGGTGTTACAC |
| s7 | pHK001-seq6 | GCAGCAATCTCCTTCACCTT |
| s8 | pHK001-seq7 | CCTCCATCGCTACTTTGCTACTAC |
| s9 | pHK001-seq8 | TTGCTGTGAGCCGGTGTAG |
| s10 | pHK001-seq9 | GGAGCACGAAGGAGGTGAAT |
| s11 | pHK002-seq0 | GTACTTAAACGCTAACGGTCAGC |
| s12 | pHK002-seq1 | ATACGGTGAGTGGCTCAGG |
| s13 | pHK002-seq2 | GGGGAGAGTGTGATGATAAGTGGG |
| 5 | pHK003_cat-H2_fw | GGAGTCGCATAAGGGAGAGCGGTTCTTCATTTTCATAAAAGGGC |
| 6 | pHK003_cat-H2_rv | GCCCTTTTATGAAAATGAAGAACCGCTCTCCCTTATGCGACTCC |
| 7 | pHK003_ars-venus_fw | CATCCTTCCACATAAGGTTTTAAACAATTTCAGAACGCTCGGTTGCC |
| 8 | pHK003_ars-venus_rv | GGCAACCGAGCGTTCTGAAATTGTTTAAAACCTTATGTGGAAGGATG |
| s14 | pHK003_seq0 | ACCCAGATCACATGAAAAGGCA |
| s15 | pHK003_seq1 | TGATACTAGAGCCAGGCATCA |
| s16 | pHK003_seq2 | CATTCTCTGGTATTTGGACTCCTG |
| s17 | pHK003_seq3 | GCGCCTATATCGCCGACAT |
| 9 | pHK003_turk-cat_fw | GTTTGTCGGTGAACGCTCTCTACTAGCCCATTAGTTCAAC |
| 10 | pHK003_turk-cat_rv | GTTGAACTAATGGGCTAGTAGAGAGCGTTCACCGACAAAC |
| 11 | pHK003_ars-turk_fw | CACATAAGGTTTTAAACAATTTCAGAACGCTCGGTTGCCG |
| 12 | pHK003_ars-turk_rv | CGGCAACCGAGCGTTCTGAAATTGTTTAAAACCTTATGTG |
| s18 | pHK003_seq4 | CAAGCTGACCCTGAAGTTCATC |
| s19 | pHK003_seq5 | GCTGAGCAAAGACCCCAAC |
| 13 | pHK004_yqcK-vnus_fw | TATAATACCTAAAAAATTTTTTCACTTGTACAGCTCG |
| 14 | pHK004_yqcK-vnus_rv | CGAGCTGTACAAGTGAAAAAATTTTTTAGGTATTATA |

| | | |
|---|---|---|
| 15 | pHK004_vnus-arsR0_fw | ACCTTTTCTATTGAATCCATTCATATCGCCTTCTTTAATC |
| 16 | pHK004_vnus-arsR0_rv | GATTAAAGAAGGCGATATGAATGGATTCAATAGAAAAGGT |
| 17 | pHK004_vnus-arsR30_fw | CTTTTCTATTGAATCCATCCGTAGCAGTTCTGATTT |
| 18 | pHK004_vnus-arsR30_rv | AAATCAGAACTGCTACGGATGGATTCAATAGAAAAG |
| 19 | pHK004_vnus-arsR60_fw | ACCTTTTCTATTGAATCCATAGCAAGAGCCTTAAATTTTT |
| 20 | pHK004_vnus-arsR60_rv | AAAAATTTAAGGCTCTTGCTATGGATTCAATAGAAAAGGT |
| s20 | pHK004_seq0 | GTAGTAGCAAAGTAGCGATGGAGG |
| 26 | pHK003_ver0 | CTACACCGGCTCACAGCAA |
| 27 | pHK003_ver1 | CAGGAGTCCAAATACCAGAGAATG |
| 28 | pHK003_ver2 | GATGAACTTCAGGGTCAGCTTG |
| 29 | pHK003_r2os1l1-turk_fw | TATTTACCCGTGGTTATCCAACAACTGAGTGGCAATCAATTTCAGAACGCTCGGTTGCCG |
| 30 | pHK003_ars-r2os1l1_rv | ATTGATTGCCACTCAGTTGTTGGATAACCACGGGTAAATAATTGTTTAAAACCTTATGTGGAAG |
| 31 | pHK003_r2os1l2-cat_fw | GTTGGTAAAAGGCACTGTGAATCTCTTCTCCGTAATAAACCTACTAGCCCATTAGTTCAAC |
| 32 | pHK003_turk-r2os1l2_rv | GTTTATTACGGAGAAGAGATTCACAGTGCCTTTTACCAACAGAGCGTTCACCGACAAAC |
| 33 | pHK003_r2os1l3-H2_fw | TTGTTATCTACTGTGAAATAGGAAGTGCCGAAAACTCCCTGGTTCTTCATTTTCATAAAAGGG |
| 34 | pHK003_cat-r2os1l3_rv | AGGGAGTTTTCGGCACTTCCTATTTCACAGTAGATAACAAGCTCTCCCTTATGCGACT |
| 21 | pHK005_yqcK-ven_fw | TCTTATATATTAATCAAAAAAATTTGATATTCACTTGTACAGCTCGTCCA |
| 22 | pHK005_yqcK-ven_rv | TGGACGAGCTGTACAAGTGAATATCAAATTTTTTTGATTAATATATAAGA |
| 23 | pHK005_ven-arsR_fw | CCTTCTTTAATCAATTTTTTTTGATGTATTTAAATACCTAAAAAATTTTTTTAGCAGC |
| 24 | pHK005_ven-arsR_rv | AAAAAATTGATTAAAGAAGGCGATATGAATGGATTCAATAGAAAAGGTAAGC |
| 25 | pHK005_ver0_rv | AAGGTGAAGGAGATTGCTGC |

| | | |
|---|---|---|
| 35 | PHK003.1_1L3-H2_fw | CTACTGTGAAATAGGAAGTGCCGAAAACTCCCTCAATC CATCCTTCCACATAAGG |
| 36 | pHK003.1_back-H1_rv | TTAACTATTTCACCGATCATTCTTGACACTCCTTATTT GATTTTTTG |
| 37 | pHK003.1_back-H1_fw | CAAAAAATCAAATAAGGAGTGTCAAGAATGATCGGTGA AATAGTTAAAAATC |
| 38 | pHK003.1_op-1L1_rv | GTTGTTGGATAACCACGGGTAAATAATTGTTTAAAACC TTATGTGGAAG |
| 39 | pHK003_ver3 | CTACACCGGCTCACAGCAA |
| 40 | pHK003_ver4 | GAACAAGCGATCAAGGGATCG |
| 41 | pHK003.2_H2-back_fw | TTAATCAGACCTTTTCTTTTTGAATAAG |
| 42 | pHK003.2_H2-back_rv | TCTTGACACTCCTTATTTGATTTTTTG |
| 43 | pHK003.2_H1-op_fw | ATGATCGGTGAAATAGTTAAAAATC |
| 44 | pHK003.2_H1-op_rv | ATTGTTTAAAACCTTATGTGGAAGG |
| 45 | pHK003.2_cat_fw | TACTAGCCCATTAGTTCAACAAAC |
| 46 | pHK003.2_cat_rv | AGCTCTCCCTTATGCGAC |
| 47 | pHK003.2_mTurk_fw | TTCAGAACGCTCGGTTGC |
| 48 | pHK003.2_mTurk_rv | CAGAGCGTTCACCGACAAAC |
| 49 | pHK003.2_H2-back_H1-op | CCGAAGTAAGTCTTCAAAAAATCAAATAAGGAGTGTCA AGAatgatcggtgaaatagttaaaaatcatccgaatga tttttgc |
| 50 | pHK003.2_H1-op_mTurk | GCCATTCAATCCATCCTTCCACATAAGGTTTTAAACAA Tttcagaacgctcggttgccgcc |
| 51 | pHK003.2_mTurk_cat | CGTTTTATCTGTTGTTTGTCGGTGAACGCTCTGtacta gcccattagttcaacaaacgaaaattggataaagtg |
| 52 | pHK003.2_cat_H2-back | CCTAATGCAGGAGTCGCATAAGGGAGAGCTttaatcag acctttttcttttgaataagttcttcattttcataaaa gg |
| 53 | pHK006_dif-cat_fw | ACTTCCTAGAATATATATTATGTAAACTATGAACTTTA ATAAAATTGATTTAGAC |
| 54 | pHK006_dif-cat_rv | AGTTTACATAATATATATTCTAGGAAGTTTATAAAAGC CAGTCATTAGG |
| 55 | pHK006_venus_fw | CAAATTAAAAACTGGTCTGATCGTCCAATTTTCGTTTG TTGAAC |
| 56 | pHK006_venus_rv | CTGAATTAGCCCTTCGTTCCGAACTGGTCTGATCGTTC |

| | | |
|---|---|---|
| 57 | PHK006_turq_fw | TCTGAACGATCAGACCAGTTCGGAACGAAGGGCTAATTCAG |
| 58 | PHK006_turq_rv | TATATATTCTAGGAAGTGTCTACAGATTAATAATTATTCTTTATTATAC |
| 59 | pHK006_cat_fw | AAAGAATAATTATTAATCTGTAGACACTTCCTAGAATATATATTATGTAAAC |
| 60 | pHK006_cat_rv | CATCATCGCTCATCCATGTCGAAGTTTACATAATATATATTCTAGGAAGTTTATAAAAG |
| 61 | pHK006_back_fw | CTTCCTAGAATATATATTATGTAAACTTCGACATGGATGAGCGATG |
| 62 | pHK006_back_rv | GTTCAACAAACGAAAATTGGACGATCAGACCAGTTTTTAATTTG |
| 63 | pHK006_ver0 | CCTTTTCATGTGATCTGGG |
| 64 | pHK006_ver1 | TTTGGCTTTTCCCCGGG |
| 65 | pHK007_back_fw | GTTTTTTCAAGCTTTGACATCACTTGCTGGCGTTTTTCCATAG |
| 66 | pHK007_back_rv | GAGAATTGTCGTAAACACACATTAGTCTTGACACTCCTTATTTGATTTTTTG |
| 67 | pHK007_lacA_fw | CAAAAAATCAAATAAGGAGTGTCAAGACTAATGTGTGTTTACGACAATTCTC |
| 68 | pHK007_lacA_rv | CTATGGAAAAACGCCAGCAAGTGATGTCAAAGCTTGAAAAAAC |
| 69 | pHK007_ver0 | GGTATCTTTATAGTCCTGTCG |
| 70 | pHK008_ars_fw | GAATTCCTCCTTTAATTGGTGTCTTCATATCGCCTTCTTTAATCAATTTTTTTTG |
| 71 | pHK008_ars_rv | GCCGCCGGGCGTTTTTTATTAATTGTTTAAAACCTTATGTGGAAG |
| 72 | pHK008_back_fw | CCACATAAGGTTTTAAACAATTAATAAAAAACGCCC |
| 73 | pHK008_back_rv | CAAAAAAAATTGATTAAAGAAGGCGATATGAAGACACCAATTAAAGGAGGAATTC |
| 74 | pHK009_ase_fw | CCGGGCGTTTTTTATAACAGGCCTCTAAAGAGACC |
| 75 | pHK009_ase_rv | GAATTCCTCCTTTAATTGGTGTTATGAGCTCCTCCTTCCC |
| 76 | pHK009_back_fw | GGGAAGGAGGAGCTCATAACACCAATTAAAGGAGGAATTC |
| 77 | pHK009_back_rv | TCTCTTTAGAGGCCTGTTATAAAAAACGCCCG |

| | | |
|---|---|---|
| 78 | pHK006.2_A_fw | TATTTACCCGTGGTTATCCAACAACTGAGTGGCAATCA ATACTTCCTAGAATATATATTATGTAAAC |
| 79 | pHK006.2_A_rv | GTTTATTACGGAGAAGAGATTCACAGTGCCTTTTACCA ACAGTTTACATAATATATATTCTAGGAAG |
| 80 | pHK006.2_B_fw | GTTGGTAAAAGGCACTGTGAATCTCTTCTCCGTAATAA ACTCGACATGGATGAGCGATG |
| 81 | pHK006.2_B_rv | GCCACTCAGTTGTTGGATAACCACGGGTAAATAGTCTA CAGATTAATAATTATTCTTTATTATAC |
| 82 | pHK010_bla_fw | CCTAACGAAAAAAGAGATCACATGGTACTCTTGGAATT TG |
| 83 | pHK010_bla_rv | CAAATTCCAAGAGTACCATGTGATCTCTTTTTTCGTTA GG |
| 84 | pHK010_ven_fw | GCTGCAATGATACCGCGGGACCCACGCTCACCGGC |
| 85 | pHK010_ven_rv | GCCGGTGAGCGTGGGTCCCGCGGTATCATTGCAGC |
| 86 | pHK006v4_ven_fw | CACACCAGGTCTCAGTTGTATCCCACTTTATCCAATTT TCG |
| 87 | pHK006v4_ven_rv | CACACCAGGTCTCACAGTCACAAATTAAAAACTGGTCT GATCG |
| 88 | pHK006v4_turq_fw | CACACCAGGTCTCAACTGCCGTCATCTTTATTATGCTG C |
| 89 | pHK006v4_turq_rv | CACACCAGGTCTCATGTGCATCCTTTCACAATTTGTCT ACAG |
| 90 | pHK006v4_cat_fw | CACACCAGGTCTCACACAACTTCCTAGAATATATATTA TGTAAACTCAGACAAGTAAGCCTCCTAAATTC |
| 91 | pHK006v4_cat_rv | CACACCAGGTCTCAGAGAAGTTTACATAATATATATTC TAGGAAGTACGATGCGTCCGGCG |
| 92 | pHK006v4_back_fw | CACACCAGGTCTCATCTCTCGACATGGATGAGCGATG |
| 93 | pHK006v4_back_rv | CACACCAGGTCTCACAACCGATCAGACCAGTTTTTAAT TTGTG |
| 94 | pHK006v5_preB_fw | TATTTACCCGTGGTTATCCAACAACTGAGTGGCAATCA ATCAACATCGAGGACGGC |
| 95 | pHK006v5_preB_rv | TAACAACTGTGGATTTTCTGATTGAACTCACTTACGGC GAATCTACACGACGGGGAG |
| 96 | pHK006v5_preS_fw | TATTTACCCGTGGTTATCCAACAACTGAGTGGCAATCA ATCGATCACATGGTCCTGC |

| | | |
|---|---|---|
| 97 | pHK006v5_preS_rv | TAACAACTGTGGATTTTCTGATTGAACTCACTTACGGC GAGTTGAACTACATGCACTCC |
| 98 | pHK006v5_B_fw | AGTTCAATCAGAAAATCCACAGTTGTTAGGTCTCAGTC AGACAAATTGTGAAAGGATGC |
| 99 | pHK006v5_B_rv | CACTCAGTTGTTGGATAACCACGGGTAAATAGGTCTCA ACTGACAACCGATCAGACCAG |
| 100 | pHK006v5_T_fw | GAGTTCAATCAGAAAATCCACAGTTGTTAGGTCTCAAG TCCATATACATTGCCCGTCGG |
| 101 | pHK006v5_T_rv | CTCAGTTGTTGGATAACCACGGGTAAATAGGTCTCATG ACATAAAGCTGACCGTTAGCG |
| 102 | pHK006v5_V_fw | TGAGTTCAATCAGAAAATCCACAGTTGTTAGGTCTCAC AGTGGAGGCTTACTTGTCTGC |
| 103 | pHK006v5_V_rv | CTCAGTTGTTGGATAACCACGGGTAAATAGGTCTCAGA CTTGATTATGCCGCGATTTCC |
| 104 | pHK006v5_seq0 | TTCTGCTGGTAGTGGTC |
| 105 | pHK006v5_seq1 | CGTTCATCCATAGTTGCC |
| 106 | pHK006v6_T_fw | GAGTTCAATCAGAAAATCCACAGTTGTTAGGTCTCATG ACCATATACATTGCCCGTCGG |
| 107 | pHK006v6_T_rv | CTCAGTTGTTGGATAACCACGGGTAAATAGGTCTCAAG TCATAAAGCTGACCGTTAGCG |
| 108 | pHK006v6_V_fw | TGAGTTCAATCAGAAAATCCACAGTTGTTAGGTCTCAG ACTGGAGGCTTACTTGTCTGC |
| 109 | pHK006v6_V_rv | CTCAGTTGTTGGATAACCACGGGTAAATAGGTCTCACA GTTGATTATGCCGCGATTTCC |
| 110 | pHK008v2_A_fw | TTTGGTCTCAAGTCCCTCTATCTGGTTACACAAGC |
| 111 | pHK008v2_A_rv | TTTGGTCTCATGACCAATTTTTTTTGATGTATTTATTA TACTCATG |
| 112 | pHK008v2_B_fw | TTTGGTCTCAGTCACAATTAAAGGAGGAATTCAAAATG G |
| 113 | pHK008v2_B_rv | TTTGGTCTCAGACTGCAACCGAGCGTTCTGAAC |
| 114 | pHK011_fw | CCAATTAAAGGAGGAATTCAAAATGGTTAGTAAAGGAG AAGAAC |
| 115 | pHK011_rv | GTTCTTCTCCTTTACTAACCATTTTGAATTCCTCCTTT AATTGG |
| 116 | pHK012_fw | CACCAATTAAAGGAGGAATTCAAAATGGTGAGCAAGGG CG |

| | | |
|---|---|---|
| 117 | pHK012_rv | CGCCCTTGCTCACCATTTTGAATTCCTCCTTTAATTGGTG |
| 118 | HK_amyE_fw | CGAACTGGTACTGGCTG |
| 119 | HK_amyE_rv | TCAAGGAAAACATGAGGCG |
| 120 | HK_rep_fw | GCCCATTAGTTCAACAAACG |
| 121 | HK_rep_rv | TTAAACGCTAACGGTCAGC |
| 122 | pHK008v3_B_fw | CATTCAATCCATCCTTCCAATGATGAATAAAAAACGCCCG |
| 123 | pHK008v3_B_rv | CATGAGTATAATAAATACATCAAAAAAAATTGGACACCAATTAAAGGAGGAATTCAAAATG |
| 124 | pHK008v3_P_fw | CATTTTGAATTCCTCCTTTAATTGGTGTCCAATTTTTTTTGATGTATTTATTATACTCATG |
| 125 | pHK008v3_P_rv | CGGGCGTTTTTTATTCATCATTGGAAGGATGGATTGAATGGC |
| 126 | HK_cat-lin_fw | GGTCTCAGTCATACTAGCCCATTAGTTCAAC |
| 127 | HK_cat-lin_rv | GGTCTCACAGTCTCCCTTATGCGACTCC |
| 128 | pHK006v6-lin_fw | GGTCTCAACTGTCTCTCGACATGGATGAGC |
| 129 | pHK006v6-lin_rv | GGTCTCATGACCATATACATTGCCCG |
| 130 | pHK006v7_preB-fr_fw | GAGTTCAATCAGAAAATCCACAGTTGTTAGGTCTCAGTCATACTAGCCCATTAGTTCAACAAACG |
| 131 | pHK_RR_preB_a1 | GTTGTTAGGTCTCAGGTATACTAGCCCATTAGTTCAACAAACG |
| 132 | pHK_RR_preB_a2 | GGGTAAATAGGTCTCAACCTACAACCGATCAGACCAGTTTTTAATTTG |
| 133 | pHK_RR_preB_b1 | CTGATCGGTTGTAGGTTGAGACCTATTTACCCGTGGTTATCC |
| 134 | pHK_RR_preB_b2 | CTAATGGGCTAGTATACCTGAGACCTAACAACTGTGGATTTTCTG |
| 135 | pHK_RR_preT_a1 | GTTGTTAGGTCTCATACCCATATACATTGCCCGTCGGTC |
| 136 | pHK_RR_preT_a2 | GGGTAAATAGGTCTCACCACATAAAGCTGACCGTTAGCGTTTAAG |
| 137 | pHK_RR_preT_b1 | GGTCAGCTTTATGTGGTGAGACCTATTTACCCGTGGTTATCC |
| 138 | pHK_RR_preT_b2 | GGCAATGTATATGGGTATGAGACCTAACAACTGTGGATTTTCTG |

| | | |
|---|---|---|
| 139 | pHK_RR_preV_a1 | GTTAGGTCTCAGTGGGGAGGCTTACTTGTCTGCTTTC |
| 140 | pHK_RR_preV_a2 | GGGTAAATAGGTCTCAAGGTTGATTATGCCGCGATTTC CAATG |
| 141 | pHK_RR_preV_b1 | GCGGCATAATCAACCTTGAGACCTATTTACCCGTGGTT ATC |
| 142 | pHK_RR_preV_b2 | CAAGTAAGCCTCCCCACTGAGACCTAACAACTGTGGAT TTTC |
| 143 | pHK_RBS_preV_fw | TCGCCGTAAGTGAGTTCAATCAGAAAATCCGTCTCATG AAAGTATTACATATGTAAGATTTAAATGC |
| 144 | pHK_RBS_preV_rv | GATTTTCTGATTGAACTCACTTACGGCGACGTCTCTAG GTAAGCATGGTTAGTAAAGGAGAAG |
| 145 | pHK018_pV_A_fw | ACCGAAAAGTCTCAGTTACAGGGATAGTTGTCTACTCT TATTAGCAGCAATCTCCTTCAC |
| 146 | pHK018_pV_A_rv | TAAGAGTAGACAACTATCCCTGTAACTGAGACTTTTCG GTGACACCAATTAAAGGAGGAATTCAAAATG |
| 157 | pHK020_ArsR_fw | AAAGAAGACAAATGGATGAAACGAAATCAGAACTGCTA CG |
| 158 | pHK020_ArsR_v1_rv | AAAGAAGACCCGTTCTTTTAGCAGCAATCTCCTTCAC |
| 159 | pHK020_ArsR_v2_rv | AAAGAAGACCCCGTATTTTAGCAGCAATCTCCTTCAC |
| 160 | PHK020_seq0 | GGTTGCCGTCATCTTTATTATG |
| 161 | pHK020_seq1 | CTGTTCAATAAAGCTGACCG |
| 162 | pHK020_spc-bla_v1_fw | AAAGAAGACAAGAACAGTGATGATACTAGAGCCAGG |
| 163 | pHK020_spc-bla_v1_rv | AAAGAAGACAATTCAGTGATACGCCTATTTTTATAGGT TAATG |
| 164 | pHK020_spc-bla_v2_fw | AAAGAAGACAATACGGAGCTGTACAAGTGATGATACTA G |
| 165 | pHK020_spc-bla_v2_rv | AAAGAAGACAAAAGCAGGTTAATGTCATGATAATAATG GTTTC |
| 166 | pHK020_yvfM_rv | TTAGAAGACAAAGGTTTAGCCCTTCGTTCCGC |
| 167 | pHK020_yvfM_v1_fw | AAAGAAGACAATGAATGTTGCTAGCCTTTATGGC |
| 168 | pHK020_yvfM_v2_fw | TTTGAAGACTTGCTTTCTCCTTTGGACGGCAG |
| 169 | pHK021_seq2 | GTTGCCGTCATCTTTATTATGC |
| 170 | pHK021_seq1 | GTTGTGTCTCTACTGTAATAAATTCG |
| 171 | pHK021_seq0 | GACACAACATTATCGTATTGCC |
| 172 | pHK023_seq1 | GCTGACCGTTAGCGTTTAAGTAC |
| 173 | pHK023_seq0 | CCCGTCGGTCTATTCAATTTAG |

| | | |
|---|---|---|
| 174 | pHK023v2_back_rv | AAAGAAGACAACGTAGCATAATAAAGATGACGGCAACC |
| 175 | pHK023v2_back_fw | AAAGAAGACAAGAACACGCTCTCTACTAGAGATCTG |
| 176 | pHK023v1_back_fw | AAAGAAGACTAGAACCGGTGAACGCTCTCTAC |
| 177 | pHK023v1_back_rv | ATAGAAGACAACGTATTAGCCCTTCGTTCCGC |
| 178 | pHK023v2_ven_rv | AAAGAAGACTTGTTCGGTCTCTGTTCCTGCTTTCTTCA TTAGAATCAATCC |
| 179 | pHK023v1_ven_rv | AAAGAAGACTTGTTCGGTCTCAGTTCGGAGGCTTACTT GTCTGC |
| 180 | pHK023v2_ven_fw | AAAGAAGACAATACGGGTCTCACCGTTCTCACAGTTGA TTATGCCG |
| 181 | pHK023v1_ven_fw | AAAGAAGACAATACGGGTCTCACCGTTGATTATGCCGC GATTTCC |
| 182 | pHK022_seq1 | GTCATTTTCATGTGCTGTAACC |
| 183 | pHK022_seq0 | GCTTACTTGTCTGCTTTCTTC |
| 184 | pHK022v2_e4R_rv | AAAGAAGACAATACGGGTCTCAGAACGGCAATACGATA ATGTTGTGTC |
| 185 | pHK022v1_e4R_rv | AAAGAAGACAATACGGGTCTCAGAACGTTGTGTCTCTA CTGTAATAAATTCG |
| 186 | pHK022v2_e4R_fw | AAAGAAGACTTATGGATTCAATAGAAAAGGTAAGCATG TCAGTGATTAAGCAGGTAATTAAG |
| 187 | pHK022v1_e4R_fw | AAAGAAGACAAATGGATTCAATAGAAAAGGTAAGCATG TCAGTGATTAAGCAGGTAATTAAGACCAAG |
| 188 | pHK022v2_back_fw | AAAGAAGACAAAGGTACAAACGAAAATTGGATAAAGTG G |
| 189 | pHK022v1_back_fw | AAAGAAGACAAAGGTGCCCATTAGTTCAACAAACG |
| 190 | pHK022v2_back_rv | ATTGAAGACAAGTTCTGATAATAATGGTTTCTTAGACG TCAG |
| 191 | pHK022v1_back_rv | AAAGAAGACAAGTTCAGGTGGCACTTTTCGGG |
| 192 | pHK022v2_amyEf_fw | AAAGAAGACTAGAACATGTTTGCAAAACGATTCAAAAC |
| 193 | pHK022v1_amyEf_fw | AAAGAAGACAAGAACTAAGGAGTGTCAAGAATGTTTGC |
| 194 | pHK022v2_amyEf_rv | AAAGAAGACAACGTAGGTCTCAACGGCCAGTTTTTAAT TTGTGTGTTTCC |
| 195 | pHK022v1_amyEf_rv | AAAGAAGACAACGTAGGTCTCAACGGCGATCAGACCAG TTTTTAATTTGTG |
| 196 | pHK022fix_aE(f)_r | CAGCCAGTACCAGTTCGACATGCTTTTATCTCCTTGAT TC |

| | | |
|---|---|---|
| 197 | pHK022fix_aE(f)_f | GAATCAAGGAGATAAAAGCATGTCGAACTGGTACTGGC TG |
| 198 | pHK022fix_aE(b)_r | CCTTAAACGCCTGTCGTCTGGTCCATTATTGATTTGAT AAAC |
| 199 | pHK022fix_aE(b)_f | GTTTATCAAATCAATAATGGACCAGACGACAGGCGTTT AAGG |
| 200 | pHK_Pars_ArsR_rv | TTTCGTCTCAGTCTAAAAAATTTTTTTAGCAGCAATCT CC |
| 201 | pHK_RBSars_rv | TTTCGTCTCACCATTCATATCGCCTTCTTTAATCAATT TTTTTT |
| 202 | pHK_RBSars_fw | AAACGTCTCAAGACAAAAAAAATTGATTAAAGAAGGCG ATATGA |
| 203 | pHK_Pars_rv | TTTCGTCTCTGTCTGATGTATTTATTATACTCATGATT TTTAAATTAATG |
| 204 | pHK_Pars_fw | TTTCGTCTCTCGTAATTGTTTAAAACCTTATGTGGAAG |
| 205 | pHK022_ggFixB_rv | ATTGAAGACTTATGCTTTTATCTCCTTGATTCC |
| 206 | pHK022_ggFixA_fw | AAAGAAGACTAGCATGTCGAACTGGTACTGGCTG |
| 207 | pHK022_ggFixB_fw | AAAGAAGACAAGACAGGCGTTTAAGGATG |
| 208 | pHK022_ggFixA_rv | AATGAAGACTATGTCGTCTGGTCCATTATTGATTTGAT AAAC |
| 209 | HK_pPen_fw | TTTCGTCTCACGTATCATCATTTCCTTCCGAAAAAACG GTTGCATTTAAATCTTACATA |
| 210 | HK_pPen_rv | TTTCGTCTCAGTCTTTGAAAGTATTACATATGTAAGAT TTAAATGCAACCGTTTTTTCG |
| 211 | pHK023_arsRfix_bb_int | TCGACGCTCAAGTCAG |
| 212 | pHK023_arsRfix_arsR_rv | CTTACCTTTTCTATTGAATCCATTTTGAATTCCTCCTT TAATTGGTGTC |
| 213 | pHK023_arsRfix_arsR_fw | GATTAAAGAAGGCGATATGAATGGATGAGACGAAATCA GAACTGCTACGG |
| 214 | pHK023_arsRfix_bb_rv | CCGTAGCAGTTCTGATTTCGTCTCATCCATTCATATCG CCTTCTTTAATC |
| 215 | pHK023_arsRfix_bb_fw | GACACCAATTAAAGGAGGAATTCAAAATGGATTCAATA GAAAAGGTAAG |
| 216 | pHK025_synth_rv | TTTGAAGACAACCATAGAGACGATACACAATAGTCACA G |
| 217 | pHK025_seq1 | GTACAGCTCGTCCATGC |

| | | |
|---|---|---|
| 218 | pHK025_seq0 | CTGTCCTTCCCCGAG |
| 219 | pHK025_mCherry_rv | AAAGAAGACAATACGCGAAGGTGAGCCAGTGTGAC |
| 220 | pHK025_mCherry_fw | AAAGAAGACAAATGGTGAGCAAGGGCGAG |
| 221 | pHK023_arsRfix_arsR_rv2 | TGAATCCATTTTGAATTCCTCCTTTAATTGGTGTCTAA AAAATTTTTTTAGCAGCAATCTCCTTCACC |
| 222 | pHK025_mCherry_int | AATGAAGACTTCCCATGGTTTTCTTCTGCATTAC |
| 223 | pHK001_ars_RHS_fw | TTTGGTCTCATACGTTAATGTCTAGCTATACGAACTTG AC |
| 224 | pHK001_arsC_RHS_fw | TTTGGTCTCATACGGGAATTTGCTGAAACAGGG |
| 225 | pHK001_arsB_RHS_fw | TTTGGTCTCATACGATGGTTTCAACGAAAATACTTTGG |
| 226 | pHK001_yqcK_RHS_fw | TTTGGTCTCATACGTGATATTACAACGAACTCTTGC |
| 227 | pHK001_arsR_RHS_fw | TTTGGTCTCATACGAAAAGGTGAAGGAGATTGCTG |
| 228 | pHK001_ars_LHS_rv | TTTGGTCTCTACCTCCTCCATCGCTACTTTGCTACTAC |
| 229 | pHK001_arsC_LHS_rv | TTTGGTCTCAACCTATTTTATTCTCCATATATTCCACC TCTAC |
| 230 | pHK001_arsB_LHS_rv | TTTGGTCTCTACCTCGTTTCATTACTTTATCCTCACC |
| 231 | pHK001_yqcK_LHS_rv | TTTGGTCTCAACCTTTAACCCCTACATGAACATATTTC |
| 232 | pHK001_arsR_LHS_rv | TTTGGTCTCTACCTGATTTCGTCTCATCCATTCATATC |
| 233 | pHK001_bbRHS_rv | TTTGGTCTCGGTTCCCAACGATCAAGGC |
| 234 | pHK001_bbLHS_fw | TTTGGTCTCTGAACCGGAGCTGAATGAAG |
| 235 | pHK001_dcat_rv | TTTGGTCTCGCGTAAGTTTACATAATATATATTCTAGG AAGTCTTCAACTAACGGGGCAG |
| 236 | pHK001_dcat_fw | TTTGGTCTCAAGGTACTTCCTAGAATATATATTATGTA AACTGCCCATTAGTTCAACAAACG |
| 237 | pHK023_bbsI_RHS_fw | AAAGAAGACAAATGGATTCAATAGAAAAGGTAAGCATG G |
| 238 | pHK023_bbsI_arsR_rv | TAAGAAGACATCCATTTTGAATTCCTCCTTTAATTGGT G |
| 239 | pHK023_bbsI_arsR_fw | TAAGAAGACAATGAATGGATGAGACGAAATCAGAAC |
| 240 | pHK023_bbsI_LHS_rv | TTGGAAGACTTTTCATATCGCCTTCTTTAATCAATTT |
| 241 | pHK_colE1_rv | TTTGAAGACTTGTCGGGTTTCGCCACCTC |
| 242 | pHK_colE1_fw | TTAGAAGACAACGACAGGACTATAAAGATACC |
| 243 | pHK_pNULL_rv | AAACGTCTCAGTCTTTCCTGTACGAGAGACGAAA |
| 244 | pHK_pNULL_fw | TTTCGTCTCTCGTACAGGAAAGACTGAGACGTTT |
| 245 | pHK025v1_mC_rv | AAAGAAGACAATACGGGTCTCTGAACCGAAGGTGAGCC AGTGTG |

| | | |
|---|---|---|
| 246 | pHK025_mCherry_ifw | AAAGAAGACGATGGGCTGGGAGG |
| 247 | pHK_amyE(locus)_rv | AACTGCTTCCAACAAAACC |
| 248 | pHK_amyE(locus)_fw | GCATTGTTCTCCATTCTCG |
| 249 | pHK026_seq1 | GAGGTTTTGAATCGTTTTGC |
| 250 | pHK026_seq0 | GAAGCGTTCACAGTTTCG |
| 251 | pHK026_amyE(us)_rv | AAAGAAGACAAGTTCCTGGAGAATTTTGTTAACGCTTAC |
| 252 | pHK026_amyE(us)_fw | AAAGAAGACTACCGACTGTGTGAACCCGACATC |
| 253 | pHK026_amyE(fr)_fw | ATAGAAGACTTTACGAAGGAGTGTCAAGAATGTTTGC |
| 254 | pHK026_amyE(fr)_rv | AAAGAAGACAAACCTCGATCAGACCAGTTTTTAATTTGTG |
| 255 | pHK026_mV_rv | AAAGAAGACATCGTACCTAAAAAGTAAGCACCTGTTATTG |
| 256 | pHK026_mV_fw | AAAGAAGACAAGAACGCGATTTCCAATGAGGTTAAG |
| 257 | pHK026_bb_rv | AAAGAAGACAATCGGGGAGAGGATAAACAATGATGACC |
| 258 | pHK026_bb_fw | AAAGAAGACAAAGGTCATTTTACCAATCCTGAAACGG |
| 259 | pHK025v2_amyE_rv | TTTGGTCTCATACGAAGGTTTTTATATTACAGCTCCAGATCC |
| 260 | pHK025v2_BsaI_synth_fw | TTTGGTCTCTACCTTTTTCAGAACGCTCGG |
| 261 | HK_mCherry_0 | CCTTGTAGATGAACTCGCC |
| 262 | HK_mCherry_1 | CGCATGAACTCCTTGATGATG |
| 263 | HK_mCherry_2 | CCCACGCCGAAACAAG |
| 264 | pHK_lacA_rv | TGCTTTTCATGATTTCATCC |
| 265 | pHK_Cot_seq1 | GTTTTTCCGCAGCTCATTG |
| 266 | pHK_Cot_seq0 | TTATAACCACTCGTTCACTCC |
| 267 | pHK026hr_cot_DS_rv | ATGAAGACAAATCGTTTTCTACAGCTTCACGCAC |
| 268 | pHK026hr_cot_DS_fw | AAGAAGACAAGTCACATAAGGGAGAGCGTTTTTC |
| 269 | pHK026hr_cot_US_rv | TTGAAGACTACAGTTGCCGGACAAGATACAATTC |
| 270 | pHK026hr_cot_US_fw | TTGAAGACATATGCAGTAAGGCCTCCGTTTTTTCAG |
| 271 | pHK026hr_lacA_ds_rv | TTGAAGACATATCGTTCCAGCCGTTTCAGGATTG |
| 272 | pHK026hr_lacA_ds_fw | TTGAAGACTTGTCATAGGCTGATGCTCCGC |
| 273 | pHK026hr_lacA_US_rv | TTGAAGACATCAGTTTAGCCCTTCGTTCCGC |
| 274 | pHK026hr_lacA_US_fw | TTGAAGACTAATGCTCTCCTTTGGACGGCAG |
| 275 | pHK026hr_I_rv | ATGAAGACTTTGACGCTTACTTGTCTGCTTTCTTC |
| 276 | pHK026hr_I_fw | TAGAAGACTAACTGGTTAAGacctTTTTCAGAACG |
| 277 | pHK026hr_BB_rv | TTGAAGACTTGCATGGAGAGGATAAACAATGATGACC |

| | | |
|---|---|---|
| 278 | pHK026hv_BB_fw | AAGAAGACAACGATTTTACCAATCCTGAAACGGC |
| 279 | pHK026hr_cot_US_fw2 | TTGAAGACTAATGCTCATGCTCTTTTTGCTTTTGTTC |
| 280 | pHK026hr_I_rv2 | ATGAAGACTATGACGAGGTTTTGAATCGTTTTGC |
| 281 | HK_R1_fw | AAACGTCTCAAGACACCAATTAAAGGAGGAATTCAAAA TGGTGAGACGAAA |
| 282 | HK_R1_rv | TTTCGTCTCACCATTTTGAATTCCTCCTTTAATTGGTG TCTTGAGACGTTT |
| 283 | HK_R2_fw | AAACGTCTCAAGACAAAGGAGGAAAAACAATGGTGAGA CGAAA |
| 284 | HK_R2_rv | TTTCGTCTCACCATTGTTTTTCCTCCTTTGTCTTGAGA CGTTT |
| 285 | HK_R3_fw | AAACGTCTCAAGACTAAACGGGGAAATAATGGAGGTGG CACGATGGTGAGACGAAA |
| 286 | HK_R3_rv | TTTCGTCTCACCATCGTGCCACCTCCATTATTTCCCCG TTTAGTCTTGAGACGTTT |
| 287 | HK_R4_fw | AAACGTCTCAAGACTATCTTATAAAAAACAAGGGGGGC TAAACATGGTGAGACGAAA |
| 288 | HK_R4_rv | TTTCGTCTCACCATGTTTAGCCCCCCTTGTTTTTTATA AGATAGTCTTGAGACGTTT |
| 289 | HK_R5_fw | AAACGTCTCAAGACGCATACTGTTTCGAAAGGAGGCGT GCTATATGGTGAGACGAAA |
| 290 | HK_R5_rv | TTTCGTCTCACCATATAGCACGCCTCCTTTCGAAACAG TATGCGTCTTGAGACGTTT |
| 291 | HK_R6_fw | AAACGTCTCAAGACACCAATTAAAGGAGGAATTCAAAT GGTGAGACGAAA |
| 292 | HK_R6_rv | TTTCGTCTCACCATTTGAATTCCTCCTTTAATTGGTGT CTTGAGACGTTT |
| 293 | HK_R7_fw | AAACGTCTCAAGACACCAATTAAAGGAGGAATTCAATG GTGAGACGAAA |
| 294 | HK_R7_rv | TTTCGTCTCACCATTGAATTCCTCCTTTAATTGGTGTC TTGAGACGTTT |
| 295 | HK_R8_fw | AAACGTCTCAAGACACCAATTAAAGGAGGAATTCATGG TGAGACGAAA |
| 296 | HK_R8_rv | TTTCGTCTCACCATGAATTCCTCCTTTAATTGGTGTCT TGAGACGTTT |

| | | |
|---|---|---|
| 297 | pHK027_mV_rv | AAAGAAGACATGAACCCTAAAAAGTAAGCACCTGTTATTG |
| 298 | pHK027_mV_fw | AAAGAAGACAACGTAGCGATTTCCAATGAGGTTAAG |
| 299 | pHK031_R_rv | TTTGGTCTCATACGACTCTCACACTCACATTAGTTATTC |
| 300 | HK_RBS-arsC_rv | TTTCGTCTCACCATATATTCCACCTCTACATCGAGTGTCTTGAGACGAAA |
| 301 | HK_RBS-arsC_fw | TTTCGTCTCAAGACACTCGATGTAGAGGTGGAATATATGGTGAGACGAAA |
| 302 | HK_RBS-arsB_rv | TTTCGTCTCACCATTACTTTATCCTCACCTTTTTTAGTCTTGAGACGAAA |
| 303 | HK_RBS-arsB_fw | TTTCGTCTCAAGACTAAAAAAGGTGAGGATAAAGTAATGGTGAGACGAAA |
| 304 | HK_RBS-arsR_rv | TTTCGTCTCACCATTCATATCGCCTTCTTTAATCAATTTTTTGTCTTGAGACGAAA |
| 305 | HK_RBS-arsR_fw | TTTCGTCTCAAGACAAAAAATTGATTAAAGAAGGCGATATGAATGGTGAGACGAAA |
| 306 | HK_RBS-yqcK_fw | TTTCGTCTCAAGACGATTAATATATAAGAGGGGGAATTAAAATGGTGAGACGAAA |
| 307 | HK_RBS-yqcK_rv | TTTCGTCTCACCATTTTAATTCCCCCTCTTATATATTAATCGTCTTGAGACGAAA |
| 308 | HK_arsR-stop_fw | TTTGCTCTCCTGGATTAGACGTAATGAGAACTGCTACGGAAATATGAAC |
| 309 | HK_arsR_fw | TTTGCTCTCATGGATGAAACGAAATCAGAACTGCTAC |
| 310 | HK_pArsPlus_rv | TTTCGTCTCATCCATTCATATCG |
| 311 | pHK029_L_fw | CGTAGGTCTCAATGGCGATC |
| 312 | pHK028_I_rv | TTAGGTCTCGCCATAGTAGAGAGCGTTCACCG |
| 313 | pHK028_I_fw | TTTGGTCTCTCGTACATACCCACGCCGAAAC |
| 314 | pHK029_R_rv | TTTGGTCTCATACGAGGAGTCGCATAAGGGAG |
| 315 | pHK028_L_fw | TTTGGTCTCAATGGAGTGTCAAGAATGTTTGC |
| 316 | pHK029_IR_rv | TTCGGTCTCTCCATAAAAGTAAGCACCTGTTATTG |
| 317 | pHK029_IL_fw | TAAGGTCTCACGTAGCGATTTCCAATGAGGTTAAG |
| 318 | pHK028_R_rv | TTTGGTCTCATACGCTGGAGAATTTTGTTAACGCTTAC |
| 319 | pHK030_spc_rv | TTTGGTCTCGCGGTATCATTGC |
| 320 | pHK030_bb_fw | TTTGGTCTCTACCGCGGGACCCACGCTCAC |
| 321 | pHK030_spc_fw | TTTGGTCTCAGAACCGGTGAACGCTC |

| | | |
|---|---|---|
| 322 | pHK030_mV_rv | TTCGGTCTCAGTTCGGAGGCTTAC |
| 323 | pHK030_bb_rv | TTTGGTCTCGCCATAGAGACGATACACAATAGTC |
| 324 | pHK030_mV_fw | TTTGGTCTCGATGGTTAGTAAAGGAGAAGAACTTTTCAC |
| 325 | HK_LP_rv | TTTGGTCTCTCCATAGAGACGATACACAATAGACAGAGCCGTCTCATACGAGAGACCTTT |
| 326 | HK_LP_fw | AAAGGTCTCTCGTATGAGACGGCTCTGTCTATTGTGTATCGTCTCTATGGAGAGACCAAA |
| 327 | HK_LPcomGA_rv | TTTGGTCTCTCCATGCTTACCTTTTCTATTGAATCCATAGAGACGATACACAATAGTCACAGAGCCGTCTCATACGAGAGACCTTT |
| 328 | HK_LPcomGA_fw | AAAGGTCTCTCGTATGAGACGGCTCTGTGACTATTGTGTATCGTCTCTATGGATTCAATAGAAAAGGTAAGCATGGAGAGACCAAA |
| 329 | pHK031_bb_fw | CGTAGGTCTCAAGTCCGATCAG |
| 330 | pHK031_R_rv | TTTGGTCTCATACGACTCTCACACTCACATTAGTTATTC |
| 331 | HK_mK2_seq0 | GGAAGGCACAGTGAATAACC |
| 332 | HK_mK2_seq1 | CTCATTGTCTGTGTTCCCTC |
| 333 | HK_mK2_seq2 | GAAGCGAGCACAGAAACC |
| 334 | HK_mS_seq0 | GGAAGGCAGTATGAATGGAC |
| 335 | HK_mS_seq1 | GAGGTCCACCTTTTGTAACC |
| 336 | HK_mS_seq2 | GTGATGCAGAAAAAGACGATG |
| 337 | pHK035_H1_fw | TTGAAGACTTATCGAGCCTATGACCTATTGCTGG |
| 338 | pHK035_H2_rv | AAGAAGACTAATGCTTCAATACAAGCCCCTTTGG |
| 339 | pHK035_H2_fw | AAGAAGACAATACGCAAAGCAATGAAAGAAGCAGG |
| 340 | pHK035_H1_rv | AAGAAGACAAACCTGACTATTTCAGTTATGGGAGTGG |
| 341 | pHK035_dcat_rv | tttGAAGACagCGTAAGTTTACATAATATATATTCTAGGAAGTaTTCAACTAACGGGGC |
| 342 | pHK035_dcat_fw | tttGAAGACaaAGGTACTTCCTAGAATATATATTATGTAAACTGCCCATTAG |
| 343 | HK_ase_rv2 | TCATCTGCATAAACAGCTCC |
| 344 | HK_ase_rv | CTGTCACCAGCTAAATAACTTCC |
| 345 | pHK_ase_fw2 | ACAAGGGTTCTATACTTCAATGG |
| 346 | HK_ase_fw | CCACTCCCATAACTGAAATAGTC |

| | | |
|---|---|---|
| 347 | pHK037_arsR_rv | TTGGTCTCAGACTCTAAAAAATTTTTTTAGCAGCAATC TCC |
| 348 | pHK037_arsR_fw | TTGGTCTCAATGGATGAAACGAAATCAGAACTGCTAC |
| 349 | pHK036_mV_fw | AAGGTCTCAAGTCACACCAATTAAAGGAGGAATTCAAA ATGGATTCAATAGAAAAGGTAAGCATGG |

Table E.1 All primers sequences used during the project.