# Expanding the regulatory repertoire available for synthetic genetic circuits in *S. cerevisiae*.

Revised version: 28th February, 2017

28th November, 2016

Tim Weenink

Supervisor:
Dr Tom Ellis

Imperial College London
Department of Bioengineering
Centre for Synthetic Biology and Innovation

*A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy*

# Declaration of originality

I hereby certify that the work presented in this thesis is entirely my own and in the instances where this is not the case, that the original contributors and sources are clearly attributed and acknowledged.

Tim Weenink

# Acknowledgements

Doing research can be challenging. In many cases experiments fail, things turn out to be different than expected and equipment breaks down. My case was no exception. However, the biggest challenge for me was writing this thesis. Four people have been instrumental in the creation of this thesis. Without their support it is certain that this thesis would not have reached the state it is today and I would like to thank each of them for their generous contributions.

Firstly, I thank my supervisor, Dr Tom Ellis, for his substantial contributions in terms of time, advice and manuscript editing. Your help gave me back the momentum that I needed when progress seemed glacial. You were always on the ball with guidance and instant replies. I have been stubborn and wrong at times, but this has not kept you from supporting me to the full, throughout the process of thesis writing and in the lab. I am very grateful that you have given me the opportunity to complete this work in your group.

Secondly, I thank my parents, for their understanding and caring. You have seen me struggle with the writing process from early on and you have seen the hurdles in the distance that I did not want to see myself. But never have you been anything less than fully supportive of the things I wanted to achieve. I could not wish for more loving, caring parents and I thank you for always being there for me.

Finally, I thank my girlfriend Evelyn Bosma, for a seemingly endless supply of cups of tea, motivational words and kindness, especially during those times when it seemed like I would never get to the end. We have not always been right by each others side during the years of lab work, but you have always been with me in spirit. You were right behind me to cheer for me and to listen to my rants when another experiment had failed. During the writing you have done everything in your power to help me succeed and it has made an incredible difference. I could not have wished for a better companion to share my successes and my sorrows with.

The writing of this thesis was also supported by the Fryske Akademy, who generously offered me a place to work and caffeinate while I was away from college.

Apart from the writing, a lot of effort has gone into the labwork that forms the core of this work. I have learned to manage my expectations when it comes to experimental results and it is with the help of many people that I have done so. But mostly my colleagues and lab-mates have made my time in the Ellis lab an unforgettable and wonderful experience. So here I would like to express my deep gratitude and appreciation to those most intimately involved:

- Ali Awan, for helping me stay punctual and motivated at times when the experiments were particularly punishing.
- Arturo Casini, for his uncanny ability to make me feel welcomed in the lab when I was new.

I would say something about your musical taste, but words simply cannot describe it.

- Ben Blount, for being a great and knowledgeable colleague, but more importantly for introducing me to the world of homebrewing. Cheers!
- Carlos Bricio, for drawing me into the world of Star Wars and being a great colleague.
- Dejana Jovicevic, for all the fun we have had in and outside the lab. But also for your professionalism in more serious matters, such as your excellent practical demonstrations of laboratory safety procedures.
- Felix Jonas, for tirelessly working to increase the social cohesion within our group.
- Robert Chen, for his commitment and enthusiasm for the project and for being an excellent travelling companion.
- The archery club, for providing a place to release the piled up tension. I also thank my fellow-committee members and friends (Richard, Ian, Diane, Milla and many more) who have made serving on the committee an absolute joy.

Naturally, I would like to thank all the other people that I have had the pleasure of working with in the lab, including Elena, Francesca, GeoBen, Georgios, Jonek, Marta, Maureen, Ollie and many more. There are undoubtedly people who deserve to be listed here whose name I have failed to mention and I apologise sincerely for those omissions. If you know how forgetful I am you may find it in your heart to forgive me.

Lastly, I owe many thanks to the department of BioEngineering for funding this work and hosting me at Imperial College. None of this work would have been possible without it and I would like to express my appreciation for making this experience a reality.

## Abstract

Complexity is arguably the biggest challenge to the field of synthetic biology today. As synthetic constructs include more and more parts, their performance becomes less predictable and more costly to host cells. In this thesis, we work towards the expansion of the regulatory repertoire available for *S. cerevisiae* with the aim of reducing the complexity of synthetic gene circuits and thus improving their performance.

Three projects contribute to achieving this goal. First, we note that no tool exists for yeast similar to the bacterial RBS Calculator that enables accurate tuning of expression levels. We address this by designing a system for tuning translation efficiency based on predictable hairpin structures placed in mRNA 5'UTRs. We characterise the relationship between folding strength and expression output and show that this facilitates predictable expression level tuning. We implement this system as a method for rapid library generation and characterise it with regards to both context and predictability.

Next, we implement transcriptional interference (TI) as a tool to augment existing regulatory interactions that may not possess sufficient regulatory power to implement the desired function. We demonstrate that TI performs as expected at the mRNA level, but observe that the implementation interferes with translational output. We test a variety of solutions relying on different molecular mechanisms within the host and conclude with a system for functionalising the RNA product produced.

In the third project, we implement a system for simplifying circuit designs by combining activation and repression functionality into a single transcription factor. This system is based on TAL-effectors fused to an activation domain that can be targeted to an upstream region of a promoter for activation and a downstream region for repression. In a systematic series of characterisation experiments we show the creation of a TAL-effector promoter pair that exhibits the desired functionality.

# Contents

# List of abbreviations

**AD** Activation Domain

**ATc** Anhydro Tetracycline

**AU** Arbitrary Units

**CAD** computer aided design

***E. coli*** *Escherichia coli*

**EFM** Evolutionary Failure Mode

**FSC** Forward Scatter

**gRNA** guide RNA

**IPTG** Isopropyl $\beta$-D-1-thiogalactopyranoside

**IRES** Internal Ribosome Entry Site

**MFE** Minimum Free Energy

**NMD** nonsense mediated decay

**ORF** open reading frame

**PAM** Protospacer Adjacent Motif

**PIC** preinitiation complex

**qRT-PCR** quantitative Reverse-Transcription PCR

**RBS** ribosome binding site

***S. cerevisiae*** *Saccharomyces cerevisiae*

**SSC** Side Scatter

**SGD** Saccharomyces Genome Database

**STAR** Simultaneous Transcription Activation and Repression

**TALE** Transcription Activator-Like Effector

**TBP** TATA Binding Protein

**TF** transcription factor

**TI** transcriptional interference

**TSS** Transcription Start Site

**TTS** Transcription Termination Site

**UAS** Upstream Activation Sequence

**URS** Upstream Repression Sequence

**UTR** untranslated region

**yeGFP** yeast enhanced Green Fluorescent Protein

**YTK** Yeast ToolKit

# 1. Introduction

*"What I cannot create, I do not understand."*
**Richard Feynman**

## 1.1 Synthetic Biology

Although Feynman's quote more likely referred to the generation of mathematical proofs, rather than the construction of physical objects, his sentiment echoes the general philosophy of modern synthetic biology very well. The term *Synthetic Biology* was coined decades ago, but its current form and interpretation were largely defined with the publication of two seminal papers in the year 2000. "Construction of a genetic toggle switch in *Escherichia coli*." by Gardner *et al.*[1] and "A synthetic oscillatory network of transcriptional regulators." by Elowitz *et al.*[2]. Although not explicitly self-identifying as synthetic biology at the time, these papers focused on the creation of genetic circuits with defined functions through bottom-up design and modelling, rather than the tweaking of pre-existing genetic circuits in nature.

This new engineering-style approach formed the foundations of a field that drew in people from disciplines not traditionally associated with molecular biology. Electrical and electronic engineering proved to be an especially relevant field to synthetic biology, where in the past many problems had been solved in this subject that showed strong parallels with problems and challenges being encountered in synthetic biology. The involvement of engineers from other fields, such as from mechanical engineering, also brought about a more formalised definition of challenges in the field of synthetic biology along with a more rigid approach to tackling them.

In the first decade of synthetic biology from 2000 to 2010, concepts that had always been present in molecular biology and metabolic engineering but had not been clearly and explicitly defined now began to be described and dissected in the context of an engineering problem. Rather than every project and each lab encountering the same problems over and over again and each finding bespoke solutions for them, efforts in synthetic biology now began to be made to solve these issues more generally, for example, by developing and implementing standardised workflows that allow a more systematic approach to solving common issues[3].

Perhaps the most important concept that emerged during the founding years of synthetic biology was the concept of forward engineering. Much of molecular biology and especially biotechnology has relied primarily on the reverse engineering of biological pathways. This entails dissection of a genetic circuit or pathway found in nature in order to gain a detailed understanding of its function before then modifying it to perform the desired function. **Forward engineering**, on the other hand, assumes a detailed understanding of a variety of fundamental biological parts (i.e. they must be **characterised**) and these are subsequently assembled to yield a genetic circuit that performs in accordance with a predefined design specification. In order for this forward engineering approach to work, ideally the parts must be interchangeable (i.e. they must be **modular**), they must not affect other parts in unintended ways (i.e. they must be **orthogonal**) and they must perform as expected, independent of the specific implementation (i.e. they must be **robust**). If the tested circuit or pathway does not perform as expected, one of the above conditions was not met and a new cycle of design, assembly and testing is performed to identify and rectify the issues and iterate towards a more optimal solution. This process is called the **design cycle**. Each of the mentioned concepts (highlighted in bold) is fundamental to the field of synthetic biology and in the following sections, I will elaborate on their significance.

### 1.1.1 Forward engineering and abstraction

To understand the rationale behind the foundational efforts underway in synthetic biology, it is useful to look at how electronics engineering moved from research in physics to the computers and information technologies of today. As transistor circuit technologies moved into the first semiconductor chips and both their capacity for computation and complexity of production increased exponentially, methods had to be devised to manage the design process, without also requiring an exponential increase in the number of humans employed in the design. By definition, the required circuits for semiconductor chips had no precedent, because nothing as complex as a silicon chip had ever existed in computers before. This meant that design of these circuits had to be done from the ground up, starting with the most basic components and working towards a set of design specifications in steps of increasing complexity. This process is an example of forward engineering and is fundamental to the design of such systems. The opposite to forward engineering, reverse engineering, is the process of deconstructing an existing design in order to gain an understanding of how it works and how it could be changed to meet the design criteria.

In contrast to electronic engineering, molecular biology has the advantage that complex and sophisticated systems are abundantly available in nature. Consequently, the focus in molecular biology has traditionally been on the reverse engineering approach, where existing genetic circuits and pathways are deconstructed or rewired for new functions. In the first decades of molecular biology, the reverse engineering approach was inevitably the method for biotechnology, since forward engineering is not possible without a thorough understanding of the basic building block components with which to make a system. However, as much more has become known about basic biology with the rise of genomics methods and systems biology, opportunities have arisen since the start of the new century for the application of the forward engineering approach in biology.

It is not surprising that in the year 2000 the point came where these opportunities were taken, because the forward engineering approach offers distinct advantages over reverse engineering. When relying on reverse engineering, one is limited to the circuits and systems that have come into existence through evolution. However, selective pressure in evolution rarely matches precisely the design specifications set for a particular project. In addition, circuits established in evolution are typically tightly integrated into other processes in the host, and these connections can affect the circuit in unintended and unpredictable ways. Identifying and decoupling these connections can be time consuming and may often be impossible because the host is dependent on proper operation of the circuit for its growth and survival. For these points and many other reasons, forward engineering can be an attractive alternative to the traditional approach of deconstructing biology; but how, in practice, does the forward engineering of synthetic biology work?

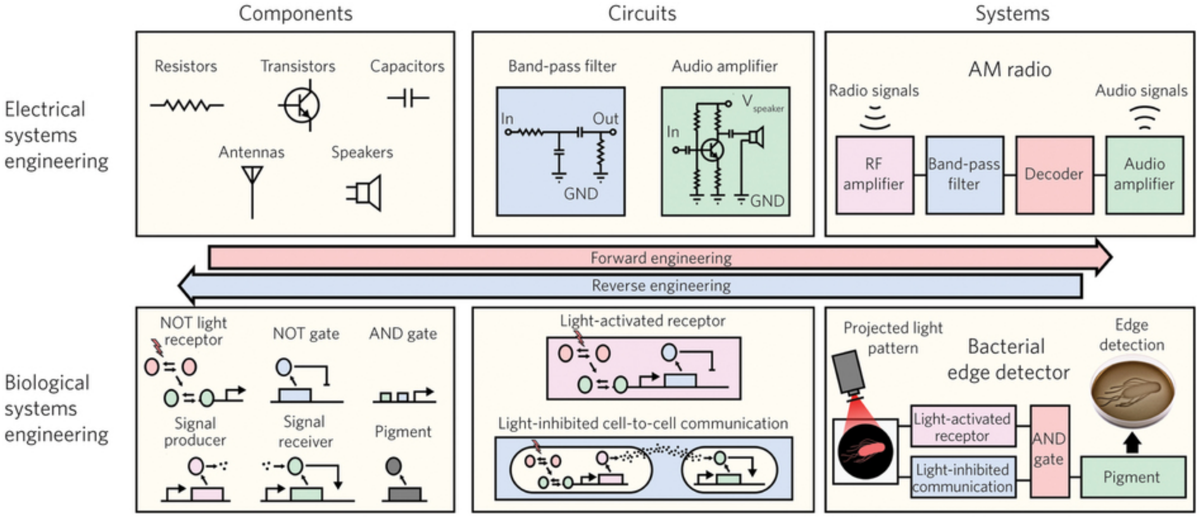
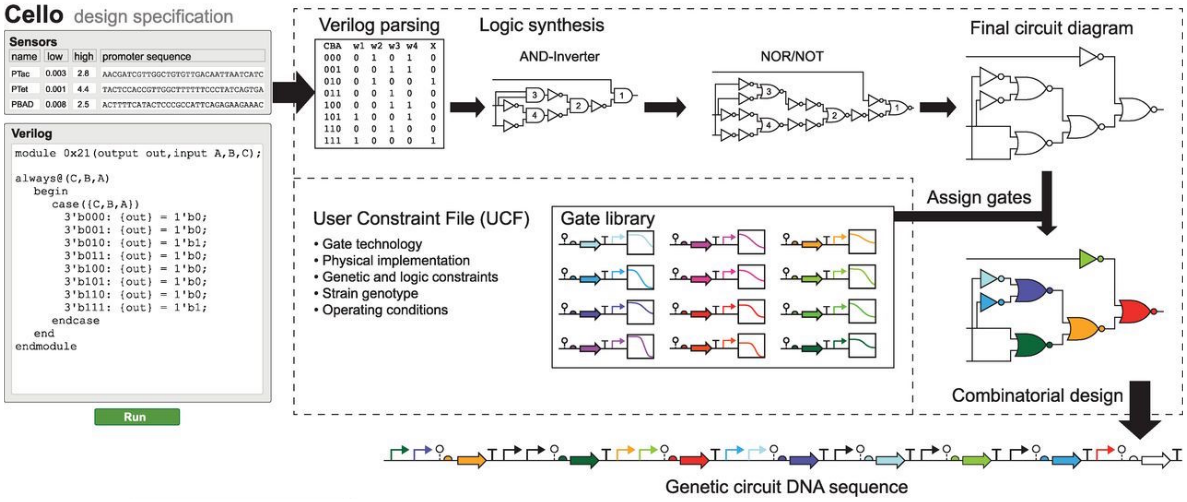


**Figure 1.1:** The parallels between electronic and biological systems engineering. Figure reproduced from Olson and Tabor, 2014[4].

To layout how the ideal synthetic biology method works, we return to the parallels between electronic and biological systems engineering. As noted earlier, the advances in transistor production technology led to a design complexity too overwhelming for traditional human design. The solution that emerged was the establishment of an abstraction hierarchy[4] with different levels that combine to give the full complex system. Semiconductors were used to build transistors at the base level, transistors were then arranged into circuits, and circuits were combined into systems. At every level, design specifications were formulated and components from a lower level were selected that could be combined to fulfil these required specifications. Crucially, however, no thorough understanding of the other levels in the hierarchy is required to complete this task. This allows complex problems to be solved without the requirement for one individual to master all the intricacies of every type of design challenge at all levels in the entire system. Instead, researchers and companies can specialize in being experts in components or design at one level of the system and improve these for all others to benefit. A further advantage is that abstraction also allows the more repetitive and mundane design challenges to be automated.

As shown in **Figure 1.1** on the preceding page, the parallels between electric engineering and (synthetic) biology allow a similar abstraction hierarchy to be used in order to reduce complexity in biological engineering. DNA sequences are defined into **parts**, such as promoters, open reading frames and terminators. Parts can be combined into **devices**, e.g. transcription units capable of producing a particular protein. Devices are coupled to form **modules**, such as one that can perform a logic operation based on inputs or convert the intensity of light at a particular wavelength into a signal that can serve as input for another module. Such modules can form a **circuit** that, in the example shown is capable of reading the intensity of light, performing some logic operations on the determined values and producing a pigment signal as the output. By spatially separating many instances of this circuit, a **system** is formed that acts as an edge detector in patterns of light. The terms of these various levels of abstraction: part, device, module and circuit, are used throughout this thesis.



| name | low | high | promoter sequence |
|---|---|---|---|
| PTac | 0.003 | 2.8 | AACGATCGTTGGCTGTGTTGACAATTAATCATC |
| PTet | 0.001 | 4.4 | TACTCCACCGTTGGCTTTTTTCCCTATCAGTGA |
| PBAD | 0.008 | 2.5 | ACTTTTCATACTCCCGCCATTCAGAGAAGAAAC |

```
module 0x21(output out,input A,B,C);

always@(C,B,A)
   begin
      case({C,B,A})
         3'b000: {out} = 1'b0;
         3'b001: {out} = 1'b0;
         3'b010: {out} = 1'b1;
         3'b011: {out} = 1'b0;
         3'b100: {out} = 1'b0;
         3'b101: {out} = 1'b0;
         3'b110: {out} = 1'b0;
         3'b111: {out} = 1'b1;
      endcase
   end
endmodule
```

| CBA | w1 | w2 | w3 | w4 | X |
|---|---|---|---|---|---|
| 000 | 0 | 1 | 0 | 1 | 0 |
| 001 | 0 | 0 | 1 | 1 | 0 |
| 010 | 0 | 1 | 0 | 0 | 1 |
| 011 | 0 | 0 | 1 | 0 | 0 |
| 100 | 0 | 0 | 1 | 1 | 0 |
| 101 | 1 | 0 | 0 | 1 | 0 |
| 110 | 0 | 0 | 1 | 0 | 0 |
| 111 | 1 | 0 | 0 | 0 | 1 |

**Figure 1.2:** The Cello environment for Computer Aided Design (CAD) and modelling in synthetic biology. Circuit behaviour is specified in the verilog language. This code is then converted to a circuit diagram by the parser. Finally, the biological parts are assigned to the circuit diagram, based on the user constraints file, which contains characterisation data of a large library of biological parts. Figure adapted from Nielsen *et al.* 2016[5].

In electronic engineering, low level design challenges were automated at a relatively early stage. However, due to the inherent complexity and stochasticity (i.e. noise) in molecular biology, this has not yet been achieved in synthetic biology. Instead, despite community efforts towards abstraction, standardisation and modularity computer aided design (CAD) in biology is still in its infancy. In order to make correct design decisions, a prediction must be made how the possible options will perform. These predictions are generally much easier to make for electronic circuits than for biological ones, as electronic circuits can be built to design without requiring host systems to run them, unlike biological circuits needing to run inside a living host cell. Computer-based design works well for electronic circuits, as efficient computer design thrives with rigid and quantitative datasets and models that aid in making design decisions successfully. These features are currently lacking in synthetic biology, and circuit design is therefore still at the stage of being somewhat bespoke rather than automated, done largely by expert researchers relying on intuition and their own qualitative experience

Unfortunately, intuition and qualitative experience generally perform poorly as complexity rises, and this results in long and costly development cycles for the realization of intended biological circuits. Ideally, mathematical modelling would be used to aid the design and simulation of circuits before they are constructed and would be an essential part of a CAD for synthetic biology. However, in many cases with biological systems, the parameters required for the creation of a quantitative model are unknown or crudely measured. This hinders the ability of those in synthetic biology to develop and employ CAD tools to reduce the cost and development time for the creation of biological circuits. Many different research groups have attempted to solve this major obstacle, but with mixed results that have resulted in a proliferation of the number of CAD tools, modelling approaches and biologically-oriented programming languages available today[6]. Perhaps the most advanced of the various attempts is the Cello CAD tool, which is shown in **Figure 1.2** on the previous page.

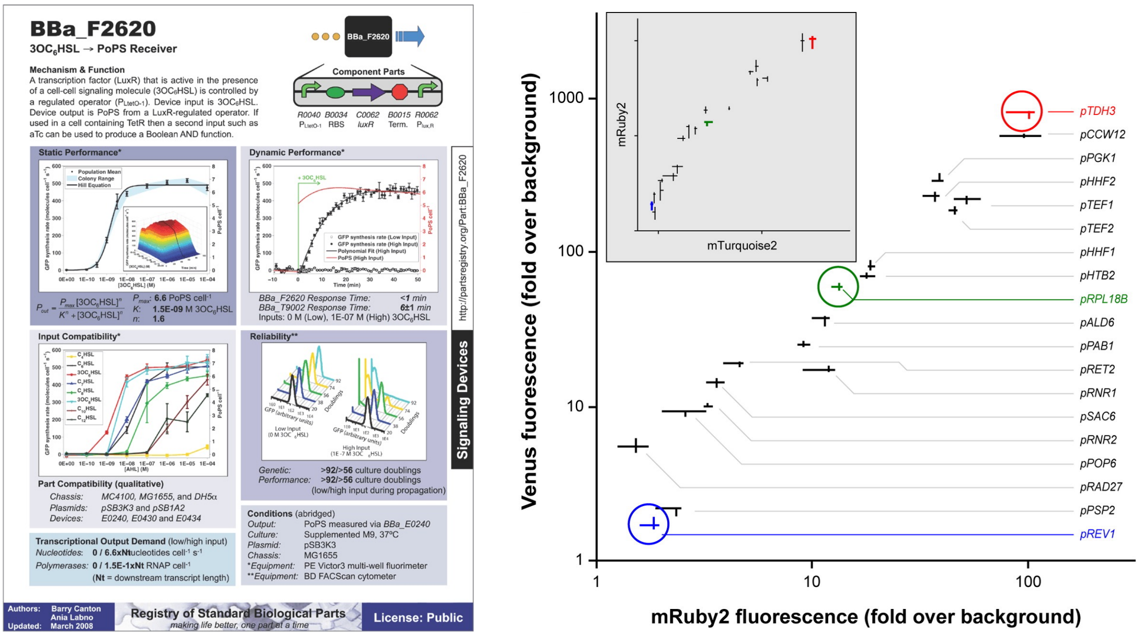
In Cello, biological circuit design specifications are converted into a computer readable format in the form of the Verilog programming language, a language originally developed for the design and verification of digital circuits. It autonomously converts the design specifications into a circuit diagram consisting of logic gates, connected in such a way that they are capable of performing the required operations. It then matches a set of pre-characterised biological logic gates to this circuit design, making sure that the characteristics of each individual gate are compatible with the other biological logic gates to which it is connected.

To demonstrate the power of this technique, 60 circuits were designed by the Cello CAD tool and then built to design with no further human input. Of these circuits, 45 showed the correct output states for all possible outcomes[5]. This is a very impressive achievement, that is not typically matched by expert human genetic circuit engineers and is currently the state-of-the-art. Sadly, not all types of circuits are amenable to the type of logic-based design required for coding in Verilog, but owing to its scalability this approach can be expected to see a widespread adoption in the field.

### 1.1.2 Standardisation, characterisation and modularity

Efforts like the Cello CAD approach discussed above are powerful illustrations of the potential for synthetic biology to turn biology into an engineering material. However, this approach and the many others preceding it typically rely heavily on libraries of standardised, characterised, orthogonal and robust parts. Standardisation is an important, yet somewhat controversial topic in synthetic biology, as standardisation can happen at many different levels and at a variety of scales. Generally, the more people involved in the creation of a standard, the less likely it is that a consensus will be reached about the specifics of the standard. This is because standardisation invariably leads to a reduction in flexibility.

Many standards for genetic circuit engineering, part characterisation and for quantifying and modelling molecular biology have been proposed, to varying degrees of success. One admirable but ultimately failed attempt at standardisation was the proposal for the creation of (machine-readable) datasheets for biological parts and devices[7], shown in panel **a** of **Figure 1.3** on the following page. While impressive, this approach has not been taken up widely as the variety of biological parts and devices people wish to work with are too diverse to be captured in a single



BBa_F2620

$3OC_6HSL$ → PoPS Receiver

Mechanism & Function

A transcription factor (LuxR) that is active in the presence of a cell-cell signaling molecule ($3OC_6HSL$) is controlled by a regulated operator ($P_{LtetO-1}$). Device input is $3OC_6HSL$. Device output is PoPS from a LuxR-regulated operator. If used in a cell containing TetR then a second input such as aTc can be used to produce a Boolean AND function.

Component Parts

| *R0040* | *B0034* | *C0062* | *B0015* | *R0062* |
|---|---|---|---|---|
| $P_{LtetO-1}$ | RBS | *luxR* | Term. | $P_{lux,R}$ |

Static Performance*

$$P_{out} = \frac{P_{max}[3OC_6HSL]^n}{K^n + [3OC_6HSL]^n}$$

$P_{max}$: **6.6** PoPS $cell^{-1}$
$K$: **1.5E-09** M $3OC_6HSL$
$n$: **1.6**

Dynamic Performance*

*BBa_F2620 Response Time:* **<1** *min*
*BBa_T9002 Response Time:* **6±1** *min*
Inputs: 0 M (Low), 1E-07 M (High) $3OC_6HSL$

Input Compatibility*

Part Compatibility (qualitative)

*Chassis:* *MC4100*, *MG1655*, and *DH5α*
*Plasmids:* *pSB3K3* and *pSB1A2*
*Devices:* *E0240*, *E0430* and *E0434*

Reliability**

*Genetic:* **>92/>56** culture doublings
*Performance:* **>92/>56** culture doublings (low/high input during propagation)

Transcriptional Output Demand (low/high input)

*Nucleotides:* **0 / 6.6xNt**ucleotides $cell^{-1}$ $s^{-1}$
*Polymerases:* **0 / 1.5E-1xNt** RNAP $cell^{-1}$
(**Nt** = downstream transcript length)

Conditions (abridged)

*Output:* PoPS measured via *BBa_E0240*
*Culture:* Supplemented M9, 37°C
*Plasmid:* pSB3K3
*Chassis:* MG1655
**Equipment:* PE Victor3 multi-well fluorimeter
***Equipment:* BD FACScan cytometer

http://partsregistry.org/Part:BBa_F2620

Signaling Devices

Authors: Barry Canton, Ania Labno
Updated: March 2008

Registry of Standard Biological Parts
*making life better, one part at a time*

License: Public

**(a)** A standardised datasheet for the characterisation of BioBrick format biological device BBa_F2620, as proposed by Canton, Labno and Endy in 2008[7].

**(b)** An example of part characterisation for inclusion in a modular DNA assembly kit (Yeast ToolKit). Fluorescent output is shown for 19 different promoters driving expression of mRuby2 (x-axis) or Venus (y-axis). Figure adapted from Lee *et al.* 2015[8].

**Figure 1.3:** Standardisation and characterisation in synthetic biology.

datasheet format. Furthermore, to characterise different parts and devices to a widely-usable degree also requires obtaining many different types of data, which is an effort seemingly too costly to be performed by individuals as parts of their project or even by a single lab specialising in parts and devices.

Another more successful effort to standardise biological parts has been led by the iGEM student competition, which has developed and populated a Registry of Standard Biological Parts. Rather than concentrate on the standardisation of data, the iGEM competition and registry focuses on parts, devices and circuits all adhering to a physical standard DNA format known as BioBricks. The iGEM parts registry, which is a freely-browsable online resource, is a popular repository with a wide variety of different data on BioBrick parts and devices, and the standard format of the DNA sequence enables their easy reuse by others in further designs, through the BioBrick DNA assembly format. While the amount of different parts and devices in the iGEM Registry is impressive, their characterisation over time has been *ad hoc*, and so data on part performance is not standardised nor machine-readable, making it challenging to use in a quantitative way. And although the registry is still in use by virtue of the popularity of the iGEM competition, the BioBricks method of DNA assembly around which the standard is based has long been superseded by more advanced assembly methods and is now virtually obsolete outside (and increasingly within) the competition. Many new more powerful methods for DNA assembly have since emerged that enable more complex constructs whilst being quicker and easier to perform. Unfortunately, even for these new assembly methods attempts at standardisation can be an issue. For example, when multiple groups individually decide to set standard formats for a DNA assembly method, multiple standards start to compete and end up fragmenting the field rather than uniting it. This has happened recently for the Golden Gate method of modular assembly of DNA parts for yeast. Simultaneous work by two groups to define a modular standard for this method led to competing yeast Golden Gate (yGG) and Yeast ToolKit (YTK) cloning standards[8,9].

Despite these examples of failed or suboptimal standards, the value of standardisation elsewhere in engineering is undeniable and so its pursuit in synthetic biology is likely to be worthwhile. Importantly, standardisation closely links to characterisation, perhaps the most essential requirement for automating the future design of devices and circuits. Without adequate data on the performance of DNA parts and devices, both humans and automation tools are unable to make informed decisions about circuit design. Ideally, characterisation should itself be standardised so that others can build upon data obtained elsewhere. Typically this is achieved by characterising a set of potential DNA parts or devices against an agreed standard. At the very least, these parts of devices should be characterised in a manner that makes them comparable among a set. This means that if one of the parts is used in a circuit, predictions can then be made on how the circuit behaviour would alter if one of the other parts from the set was used instead. As an example of a characterised part set, panel **b** of **Figure 1.3** on the previous page shows the result of a characterisation of yeast promoters available in the YTK cloning standard.

### 1.1.3 Orthogonality

Two further important engineering concepts that are often revealed in the characterisation process are orthogonality and robustness. When parts are orthogonal, they behave predictably and do not erratically affect and interact with other parts in the circuit (or elsewhere in the host cell in the case of synthetic biology). When a repressor protein is not orthogonal, it will bind and act on other parts of the network in undesirable ways, effectively causing 'short-circuits' that lead to sub-optimal performance. For repressor proteins, it is normally the binding of these to non-cognate repressor binding sites that is where loss of orthogonality is seen. This means that the proteins end up repressing promoters that they are not supposed to repress. In a circuit such as one built from multiple logic gates that rely on different repressors, this behaviour can substantially interfere with the intended operation, just as an electronic circuit will behave erratically when water or metal contaminants cause electrical connections to go between unintended places. Identifying orthogonality issues before the parts or devices are implemented in a design can save time and effort trying to understand and debug failures and unanticipated performance problems. The effects of non-orthogonality can manifest itself in subtle ways that are particularly challenging to troubleshoot. Characterisation of orthogonality is therefore often a crucial part of the characterisation process, especially for sets of comparable parts and devices.

An example of the results for characterisation of orthogonality is shown in **Figure 1.4** on the following page for a set of repressors designed to bind different promoters in bacteria. A perfect diagonal red boxes indicates that each repressor only interacts and represses the promoter with its cognate repressor binding site. Instances in this characterisation experiment show where particular repressors are not perfectly orthogonal (highlighted by the white arrow and circle). In these cases the repressor targets one or more promoters that it was not designed to repress and this would cause a short-circuit if these promoters were used in a genetic circuit also containing the non-orthogonal repressor.

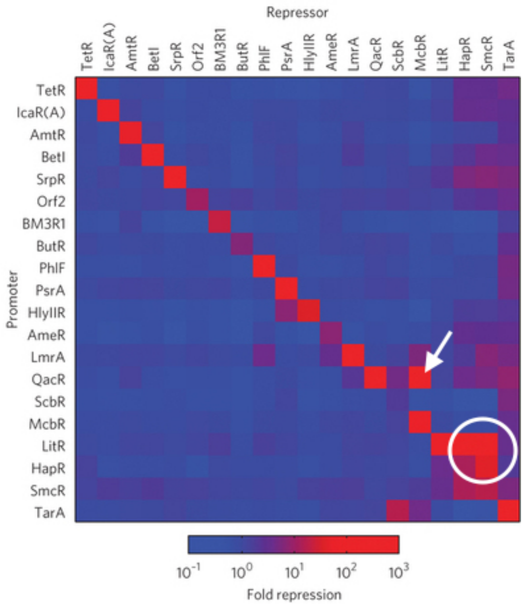


**Figure 1.4:** Characterisation of orthogonality for a set of 16 prokaryotic repressors and a set of 16 promoters that can be bound and repressed by these repressors. Effective repression of each promoter is shown in red shading, while blue indicates a lack of repression. The white arrow and circle indicate instances of non-orthogonality. Figure adapted from Stanton *et al.* 2014[10].

### 1.1.4 Robustness

After a part has been characterised, it is important that it performs as expected when it is implemented in a larger system and used over a period of time. In other words, the performance of a part must be constant and must not depend on its implementation. If this is the case, the part is said to be robust. Frequently, parts isolated from nature show performance in synthetic biology circuits that demonstrate that they are not robust. The key metric of a promoter part, the strength of mRNA expression that it directs (its output), in some cases can be heavily influenced by its surrounding sequence, also known as its 'context'. Many small DNA parts like promoters, are said to be 'context-dependent' and their performance will be altered by what local DNA sequence surrounds them. This is a challenge as the surrounding sequence typically changes when different parts are combined to make different devices and circuits. A common solution to this problem is to locally fix the surrounding sequence of the part and include these so-called insulator sequences in between the DNA that encodes the parts that are context-dependent in the DNA construct encoding a circuit.

A powerful example of this approach is shown in **Figure 1.5** on the next page which is based on experiments done in bacteria. It shows the strengths of fluorescent protein expression measured for 22 ribosome binding site (RBS) parts in *Escherichia coli* (*E. coli*) in the context of 14 different genetic contexts, where the promoter parts are changed by the Open Reading Frame (ORF) encoding the fluorescent protein remains the same[11]. These were tested in a standard design format (panel a), where the 14 different promoters define the mRNA expression levels and the 22 RBS sequences define how efficiently these mRNAs are translated to make the measurable fluorescent protein inside the cell. As the colours in the characterisation table indicate, there is a lack of correlation between the protein expression output and the parts used;

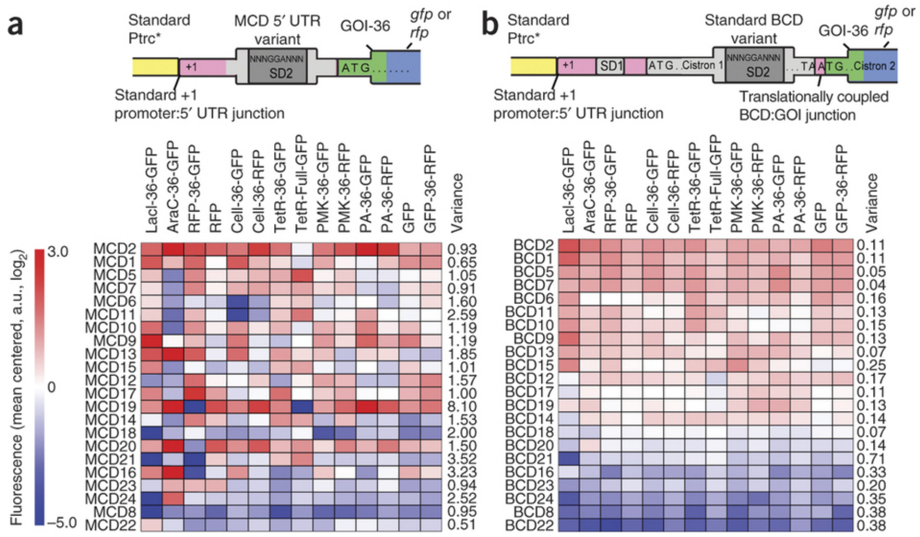


**Figure 1.5:** Robustness of and insulator sequences in 5'UTR sequences in prokaryotes. Ribosome Binding Sites of various strengths were tested in different genetic contexts. Panel **a** shows the monocistronic (uninsulated) design. Panel **b** shows the bicistronic design (insulated). High amounts of gene expression are shown in the grid as red colour, and low amounts as blue colour. White indicates middle amounts of measured gene expression from the constructed devices. Figure adapted from Mutalik *et al.* 2013[11].

an RBS that gives strong expression with one promoter, gives weak expression with another. To solve this, an insulator approach was taken (panel b). Here a so-called 'bicistronic design' is used, where the 22 RBS parts are insulated by a dummy 'RBS-ORF' upstream of the main RBS and ORF. This allows the upstream promoters to be varied, changing the sequence in that part of the DNA construct but not affecting the RBS performance. In the characterisation table, much less variability is now seen with this design. Although some variability remains, it is clear that the insulated RBS parts behave more predictably when paired with promoters of different strength and so can be said to be significantly more robust than the non-insulated versions. Although the highest expression strengths were observed in the non-insulated design, the low degree of context-dependency and high degree of predictability make the design with insulation a much more attractive option in the context of synthetic biology taking an engineering approach.

The above is only one example of where context-dependency means that parts and devices are not robust. Many more cases have been identified where DNA context effects like these play a role in (the absence of) robustness and this is an ongoing area of research in the field[12]. A further, more subtle and largely overlooked aspect of robustness is retroactivity[13]. An element is said to be retroactive when concentration changes of this element affect the activity of elements located upstream in the regulation cascade. Take for example a repressor protein binding to an operator sequence. If the number of operator sequences in the system increases, they will sequester more and more repressor molecules from the total available pool. This can effectively reduce the repression strength of the repressor independently of how many repressor molecules are being produced. In recent ground-breaking work in yeast, researchers showed that retroactivity can be mitigated by introducing a fast intermediate step (such as a phosphorylation cascade) that can quickly regenerate repressor molecules lost to sequestration[14].

### 1.1.5 The design cycle

Despite the variety of best practices devised to tackle the challenges in forward design of circuits and systems for synthetic biology, the first iteration of a design frequently does not meet the specifications when constructed and implemented in vivo. When this happens, a hypothesis is formed on the cause of the failure, possibly supported by characterisation data and additional troubleshooting experiments. Based on this hypothesis a new design is created and a new round of DNA assembly and data collection is started. This cycle of designing, building and testing is called the engineering design cycle. It is visualised in **Figure 1.6**.

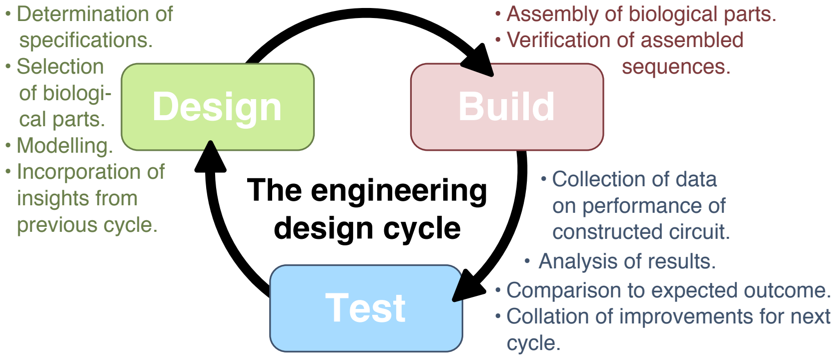

• Determination of specifications.
• Selection of biological parts.
• Modelling.
• Incorporation of insights from previous cycle.
Design
• Assembly of biological parts.
• Verification of assembled sequences.
Build
The engineering design cycle
• Collection of data on performance of constructed circuit.
• Analysis of results.
• Comparison to expected outcome.
• Collation of improvements for next cycle.
Test

**Figure 1.6:** A diagrammatic representation of the engineering design cycle in the context of synthetic biology.

The engineering design cycle is commonly referred to within synthetic biology projects and often includes a 4th part of the cycle called **Learn**, where tested designs are explored to understand their behaviour before a new design begins. Here, between the stages of testing and formulating a new design, the different ways that a circuit can fail are investigated. In an extensive review of circuit design in synthetic biology two years ago, all of the so-called 'failure modes' that were identified related to issues discussed above, such as the lack of robustness and orthogonality of circuit parts[15]. The only additional factor in this review that has not been covered previously, is the issue of recombination, where the physical DNA encoding the circuit parts is mutated and rearranged, leading to its performance being modified or lost, for example, by the synthetic DNA being removed from the cell entirely.

Recombination as a failure-mode is especially relevant when circuits and systems encoded in synthetic DNA are used within yeast cells. This is because yeast has a very efficient system for homologous recombination, where two regions of DNA that are closely matched in sequence and normally more than 30 bp in length are often brought together and recombined to remove or invert the DNA between them. The current best solution to the problem of recombination is to aim to reduce its likelihood by paying attention to the DNA sequences used in a design. A reduction in the number and length of sequences within a circuit design that are homologous to one another, should improve genetic stability against recombination. Avoiding encoding direct or inverse repeats of DNA is especially important. Recently, a useful online tool that estimates the genetic instability that could arise from the inclusion of homologous sequences has become available[16]. We describe and use this tool in **section 5.3.5** on page 192.

### 1.1.6 New enabling technologies in synthetic biology

As synthetic biology has progressed into its second decade, many advances from within the subject and from molecular biology and microbiology have greatly enabled device, circuit and system design and implementation on increasingly more complex scales. For the content of this thesis - the design and optimisation of genetic circuits - the most important advances have been in methods that dramatically improve DNA cloning and new proteins that have been discovered and engineered to enable scalable and orthogonal gene regulation.

#### High Throughput Cloning

A decade ago, one of the limiting factors in synthetic biology research was the time and effort required to build any circuit or system composed of many parts. Fortunately, since 2008, a variety of new methods for DNA cloning have come into practice that now allow new designs to be built and tested rapidly, speeding up the engineering design cycle. The first of these, Gibson Assembly, removed the need to use restriction enzyme sites to put regions of DNA together and allowed multiple parts to be assembled together in one-pot and one go. While revolutionary at the time, this DNA assembly method is now quickly being superseded in many synthetic biology labs by Golden Gate-based cloning approaches.

Golden Gate cloning goes back to using restriction enzymes and ligases to link DNA parts together, but uses type IIS restriction enzymes that cut outside their recognition sequence. This allows one enzyme to cut many parts but leave them with overhang bases of DNA that can only reassemble in a defined order. This means that many parts can be combined in a single reaction with great efficiency, so long as those parts adhere to the required standard, which is to be absent of the Type IIS enzyme recognition sites within the part sequence. As the speed and affordability of DNA synthesis has increased, this has enabled more groups to move to DNA assembly by this approach.

The method is also inherently modular, and so reinforces the principles of synthetic biology. Groups adopting this method often end up benefiting from being able to define and standardise their modular parts and re-use them between projects and between labs. This has led to Golden Gate assembly *kits* emerging for different uses that can be the starting point for assembling a vast number of possible DNA constructs. The aforementioned Yeast Tool Kit (YTK), which is used within this thesis, is an example of a kit that enables synthetic biology device and circuit construction in *Saccharomyces cerevisiae* (*S. cerevisiae*).

The kit contains a variety of genetic parts in modular, standardised format that can be assembled together by Golden Gate cloning in *E. coli* to make plasmids that encode yeast genes built from characterised promoters, terminators and ORFs. These plasmids can then be further combined by a second round of Golden Gate assembly to make multi-gene cassettes encoding devices, circuits or pathways on plasmids that can be added into yeast cells for testing. Importantly at all steps, the Golden Gate reactions are *one-pot* so that many parts come together in a defined order at one time. And also the reactions can be *combinatorial* so that multiple parts of the same type (e.g. ten different promoters) can be added into a single reaction so that a *library* of different constructs is made in the one-pot reaction.

### Orthogonal Gene Regulation

Perhaps the biggest barrier to realising complex circuits and systems in synthetic biology a decade ago was a lack of orthogonal gene regulators[6]. To enable the action of one gene to control the next in a circuit (i.e. to act as a wire), transcription factors are used as protein products that then modulate the promoters of downstream genes. Through the first decade of synthetic biology, the field relied on a very small number of reliable natural transcription factors with specific cognate promoters, e.g. the Tet Repressor (TetR) and the Lac Inhibitor (LacI). The number was so small that circuits with more than 6 different regulated promoters were never seen. The field initially attempted to solve this problem by using the only type of transcription factor at the time that was known to be able to be recoded at the protein level and made to predictably bind different promoter sequences - zinc finger protein[17]. Unfortunately very little progress was made with these as they proved difficult to reliably design and very complex to construct and express in microbial cells.

From 2010, a new type of transcription factor, the TAL-Effector, appeared to be the ideal solution. TAL-Effectors were revealed to have a repetitive, highly-modular protein sequence where the different modules could be arranged to bind almost any DNA sequence from 10 to 30 bp in length with very high strength and specificity. This for the first time offered a scalable way to make an almost unlimited number of orthogonal regulators, simply by designing TAL-Effectors to bind different DNA sequences within promoters and then fusing domains to these that either activate or repress gene expression. While construction of TAL-Effectors from the different modules that bind the different base pairs of DNA was not straightforward due to the repetitive nature of the proteins, it was immediately aided by Golden Gate cloning kits that enabled labs with little or no expertise in modular assembly to use TAL-Effectors in their research.

One of the first applications of TAL-Effectors was to fuse them to nuclease domains so that they could programmably cut genomic DNA at very specific positions inside cells when expressed[18]. However, within a year of TAL-Effector Nucleases first being put to use, CRISPR/Cas9 technology emerged which offers a simpler solution. Rather than redesign and assemble a new modular protein every time a different DNA sequence needs to be targeted, the CRISPR approach relies simply on expressing one protein - the bacterial nuclease Cas9 - and then co-expressing an RNA molecule known as a guide RNA (gRNA) that contains a short sequence that guides the Cas9 protein to the DNA site that is to be cut.

Within just 5 years, CRISPR/Cas9 has revolutionised many different areas of science and biotechnology, including synthetic biology. For genetic circuit construction, Cas9 can be mutated to create a form that binds DNA but does not cut it (dCas9) and this can be converted into an RNA-guided transcription factor by fusing repression or activation domains to the Cas9 protein[19]. As with TAL-Effectors, dCas9 offers the opportunity to have scalable, orthogonal regulators, simply by altering the sequence recognition element (the gRNA). Unfortunately in eukaryote cells like yeast, guide RNAs have to be expressed from promoters that specialise in RNA expression and cannot be expressed from the set of well-characterised mRNA-expressing promoters that are typically used in synthetic biology. This has also limited the use of RNA-based gene regulators (e.g. riboregulators and sRNAs) in circuit design in eukaryotic synthetic biology, despite these becoming a major tool used for synthetic biology circuits in bacteria[20].

### 1.1.7 Yeast as a model organism

Many of the above examples and advances in synthetic biology have been made in *E. coli*. However, for synthetic biology to live up to its full potential it is essential to apply the developed methodologies in more complex organisms such as eukaryotes. For using synthetic biology in biotechnology, bacterial systems have their limitations. Bacteria are limited in both fermentation capability and the complexity of the compounds and materials that they can produce when compared to eukaryotes like fungi, human cells and plants. It is therefore not surprising that there has been significant interest throughout the history of synthetic biology in applying the engineering approaches pioneered in *E. coli* to mammalian cells, plant cells and yeast. Mammalian systems have particularly led the way in genetic circuit work in eukaryotes, especially in terms of making cell lines specialised to act as health biosensors and produce therapeutics when needed. Despite yeast being easier to work with than any other eukaryote, more synthetic genetic circuits have been reported in mammalian cell line work than in yeast. The focus with yeast synthetic biology, especially in *S. cerevisiae*, has instead largely been on metabolic engineering applications, such as the engineering of heterologous pathways producing biofuels or speciality chemicals. This makes sense given that biotechnology uses yeast for these purposes already. However, yeast are also potentially a very useful test bed for genetic circuit synthetic biology research in eukaryotes.

While working with mammalian cells towards sensors and logic systems has obvious applications in health sensing and actuation, the actual day-to-day work of engineering mammalian cells is slow and made difficult by the complexity of mammalian cells and their genome. Yeast, on the other hand, is much simpler, and this brings many advantages:

- it has a short reproductive cycle
- it is cost-effective to propagate at lab scale and at industrial scale
- it is genetically tractable
- it is less complex in terms of the cell
- it is easier to maintain
- it is very well studied
- it is amenable to high density growth
- it is unicellular (in standard laboratory conditions)
- it has few and short introns
- it lacks native RNAi regulation
- it has a smaller genome (by over 2 orders of magnitude)
- it is widely adopted for biofuel, enzyme and pharmaceutical production
- it is less susceptible to infection during culturing

These properties make yeast highly attractive as the host organism for designing, testing and implementing novel synthetic circuits, and in 2012, myself and other members of our research group reviewed the available tools for synthetic biology research and applications in *S. cerevisiae*[21]. While we noted in that review that *S. cerevisiae* synthetic biology had developed a variety of different approaches and genetic parts that could help enable genetic circuits to be built to design, it was clear that tools for precisely controlling gene expression and programming regulation lacked significantly behind those in *E. coli*. Most notably, in *S. cerevisiae* almost all efforts to regulate expression were rather simplistic, focusing solely on changing and controlling the promoters in order to tune and switch on and off a protein-expression phenotype. Parts that modulate translation efficiency or interact with RNA post-transcription were not readily available or seldom-used, unlike in *E. coli*. Interestingly, the simplicity of how regulation has been engineered in *S. cerevisiae* synthetic biology is in complete contrast to the inherent complexity of eukaryotic gene expression. As we will see in the next section, gene expression from DNA to a protein is a multi-layered process in yeast, with many potential points of intervention. These could provide many new opportunities for more accurate control of gene expression and more complex regulation of genes in yeast synthetic biology in the future.

Note that while the term yeast refers to a wide selection of unicellular organisms in the fungi kingdom, unless noted otherwise in this thesis we refer exclusively to the organism *S. cerevisiae* when using the term 'yeast'.

## 1.2 Yeast gene expression

In our effort to expand the regulatory repertoire available for synthetic biology in yeast, we can intervene, interface and interact with a wide variety of processes that naturally occur in the cell. Yeast cells, just like any other living biological system, appear to be incredibly complex and highly networked systems, and despite decades of research we still cannot fully capture the details of all of the processes of regulation and simulate how a full cell would behave. However, despite the full picture for yeast gene regulation still being incomplete, the cellular processes that are most essential to the expression of genes in yeast are reasonably well known and these offer many opportunities for control via synthetic biology. In the following sections our understanding of these processes are described in order to lay the foundational knowledge that is essential for developing new ways to regulate and control gene expression in yeast. This is done by following the standard flow of information in gene expression, i.e. via the central dogma of molecular biology, which is shown in **Figure 1.7**. By describing each step in this process in more detail, the important enzymes, recognition sequences and interactions that are involved can be introduced in the context of how they could be controlled in synthetic biology.

In the first section we identify the important elements in transcription: the process of reading the DNA sequence to generate a copy in RNA. In the next section we discuss the modifications that the generated RNA undergoes to become a mature template for protein synthesis, i.e. post-transcriptional modifications. And in the third section we review the creation of protein from RNA: the process of translation. Many of the topics covered in these sections provide the direction of experiments for this thesis.

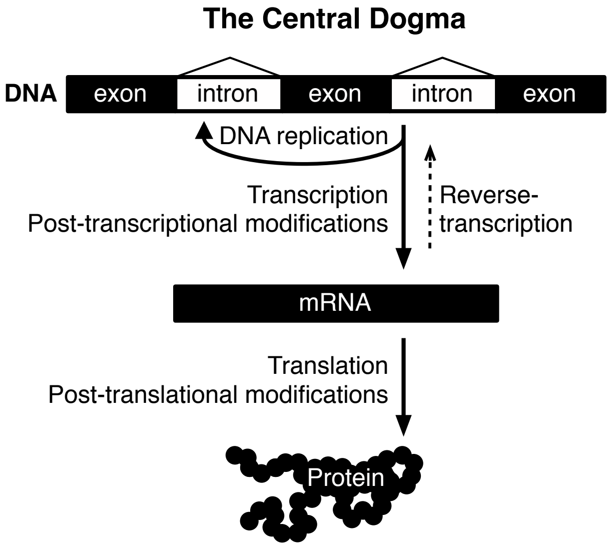


**Figure 1.7:** The central dogma in (yeast) molecular biology.

## 1.3 Transcription in yeast

The first step to creating a protein in a cell is the creation of an RNA copy of the gene's DNA sequence. The process to do this is itself directed by sequences on the DNA that flank the protein-encoding region, also known as the open reading frame (ORF). The part of the gene sequence that initiates transcription is called the promoter, and the part that terminates the copying process is called the terminator (or polyadenylation sequence). Together, a promoter, an open reading frame and a terminator make up the simplest form of a transcriptional unit in yeast. A visual representation of a basic transcription unit is shown in **Figure 1.8**. In many cases a transcriptional unit is effectively the same thing as a gene.

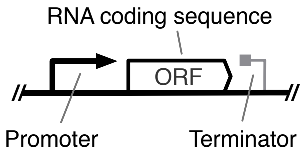


**Figure 1.8:** A typical transcription unit.

While the role of the DNA sequence of the ORF is typically seen to simply be to encode the information of the amino-acid sequence of the protein, the transcription process is itself mostly impartial to the nature of the RNA being produced. It is therefore possible that the RNA produced from a transcriptional unit is not protein encoding and instead produces non-coding RNA such as ribosomal RNA or a tRNA. Typically, though, the transcribed RNA will contain an ORF and this is the type of transcription unit of most importance in gene expression.

### 1.3.1 RNA polymerase type I, II and III.

The synthesis of the RNA copy of DNA is performed by the enzyme RNA polymerase. Three types of RNA polymerases are known to exist in yeast: type I, II and III. RNA pol I is a polymerase that exclusively transcribes the 35S ribosomal RNA from a cluster of approximately 150 repeats of transcription units[22]. This is a highly specialised RNA polymerase that carries out a very specific and important role in the cell and only binds and acts on its own cognate promoter. This specialisation and the fact that no protein encoding (i.e. messenger) RNA is produced from this promoter/polymerase system makes it difficult to exploit or control in synthetic biology, especially considering the options available with the other polymerases.

The most well-studied type of polymerase is Pol II which is responsible for transcription of all protein-encoding RNAs. In the nucleus, Pol II synthesises pre-messenger RNA which ultimately becomes the mRNA used for protein production. It also synthesises a number of small, stable RNA genes involved in mRNA splicing and in the modification of rRNAs, but primarily focuses on protein-encoding genes. The vast majority of genes in the yeast genome are transcribed by Pol II, and thus a large number of promoters have been identified and characterised in detail that are bound by this polymerase. This makes it particularly favourable for applications in regulation of synthetic circuits, especially as the vast majority of our knowledge of gene regulation in yeast is based on genes transcribed by Pol II.

The final polymerase, RNA polymerase III, transcribes a large set of genes encoding critical small untranslated RNAs like tRNAs, the 5S rRNA, U6 snRNA and RPR1 RNA[23]. Although the distinction can sometimes be difficult to make, Pol III promoters (and their terminators) typically contain certain elements that fundamentally distinguish them from their more well-known Pol II counterparts. Except for the 5S RNA gene, all known yeast Pol III transcribed genes share two control sequence motifs, the A and B blocks. These motifs are typically found *inside* the transcribed region, which is atypical for Pol II promoters where the promoter information lies upstream of the transcribed region[24]. Pol III termination sequences are also significantly less intricate than Pol II terminators, consisting simply of a stretch of six or more thymines[24]. Some examples of Pol III promoters with their control sequence configurations are illustrated in **Figure 1.9**.

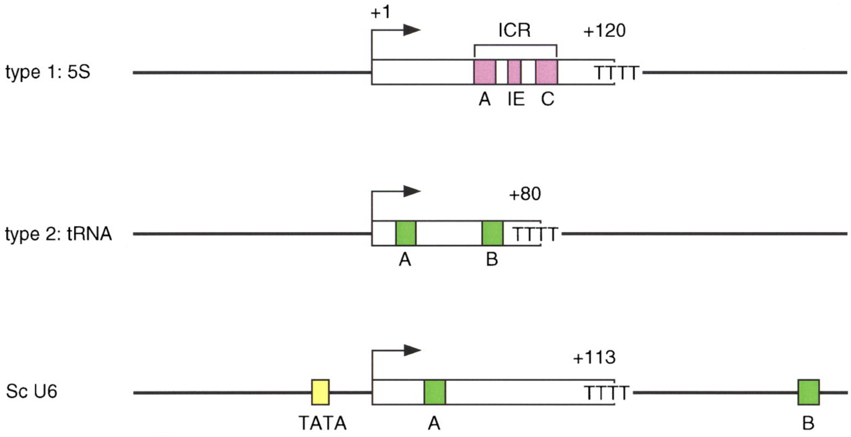


**Figure 1.9:** Variants of Pol III promoters and associated control sequences. Figure adapted from Schramm *et al.* 2002[25].

In terms of synthetic biology, it is the fact that the control sequence motifs of Pol III promoters are within the transcribed region that makes engineering regulation a difficult task. The A block control sequence is typically located 20 nucleotides downstream of the transcription start site, with the B block located a variable distance further downstream[23]. The sequences for the A and B block are therefore part of the resulting RNA, which means that any RNA transcribed from these promoters is constrained to also begin with these sequences. Nature does however offer a method to get around this, as many Pol III transcribed RNAs are co-transcriptionally cleaved. Such co-transcriptional cleavage acts to remove the A and B block from the mature RNA product and is seen for the exemplar SNR52 transcript shown in **Figure 1.10**[23]. By employing this co-transcriptional cleavage system, it therefore becomes possible to use Pol III promoters in synthetic biology systems that produce RNA, such as in guide RNA (gRNA) production in CRISPR/Cas9 applications. Indeed, one of the simplest ways to reliably express gRNAs in yeast is to insert the DNA encoding their sequence in place of the DNA region matching the mature RNA sequence for SNR52, as further detailed in **subsection 4.3.5** on page 147.

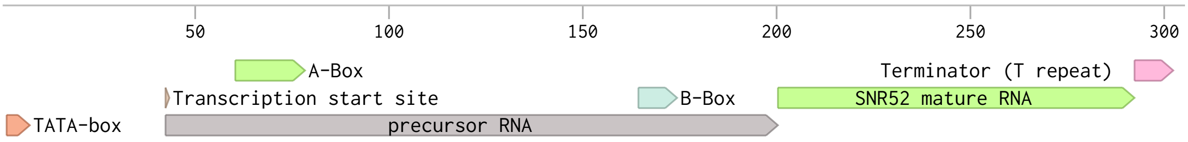
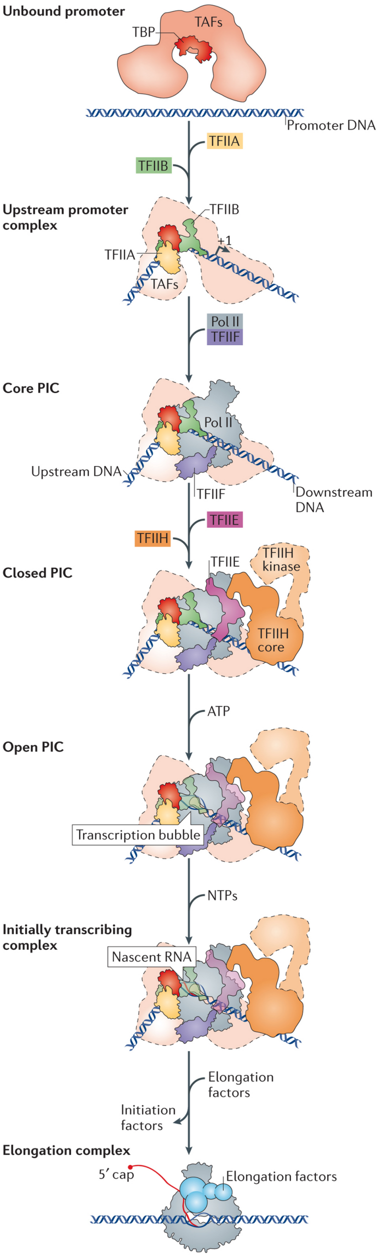


**Figure 1.10:** Diagram outlining the features of the transcriptional unit of the SNR52 RNA.

### 1.3.2 Pol II-promoter architecture

Given that almost all gene regulation and protein production in yeast requires Pol II transcription, it is unsurprising that Class-II promoters have been the subject of intense study over the past decades. These promoters are found throughout the yeast genome and give rise to all of yeasts' mRNA transcripts and thus define all of it proteins. Research in this area has led to Class II promoters in yeast being broadly categorised into two subclasses: TATA-less promoters and TATA-box containing promoters. TATA-less promoters form the biggest subclass, with an estimated 76% of all Pol II promoters belonging to this category[27]. Analyses have shown that promoters in this category tend to be associated with housekeeping genes and functions. This suggests that these genes are constitutively expressed at a low to moderate level throughout the yeast cell cycle and in many different conditions. TATA-box containing promoters, on the other hand, tend to be associated with genes involved in metabolism and response to biotic stimuli. They have been found to be highly regulated, with expression levels spanning the entire range of possible expression strengths. This regulation has been a major source of interest, making this type of promoter the most intensively studied one of the two subclasses.



Nature Reviews | Molecular Cell Biology

**Figure 1.11:** Diagram illustrating the steps of assembly of the preinitiation complex (PIC) on a TATA-box containing promoter. Figure reproduced from Sainsbury *et al.* 2015[26].

### TATA-box containing promoters

The TATA-box is a conserved region found in the promoters of many organisms, including bacteria. In *S. cerevisiae* the consensus sequence has been determined to be TATAWAWR (T-A-T-A-A/T-A-A/T-A/G)[28]. It recruits the TATA Binding Protein (TBP) transcription factor, which forces a sharp bend in the DNA helix when bound and helps to recruit the general transcription machinery and RNA polymerase II.

**Core promoter region and pre-initiation complex** The exact mechanism by which the general transcription machinery assembles onto the promoter is complex, and still contested despite decades of research. One theory is that TBP, several general transcription factors and Pol II pre-assemble into an RNA polymerase II holoenzyme complex. This complex is then directly recruited to the promoter through TBP-DNA interactions. In a competing theory, the TBP binds first, followed by each of the individual transcription factors and RNA polymerase II.

Assembly through either of these mechanisms leads to a protein complex called the preinitiation complex (PIC). The transcription factors contained in this complex each serve specific functions necessary for the initiation of transcription. For example, TFIIH possesses ATPase and helicase activity that can create negative superhelical tension in the DNA. This tension causes the DNA to melt and form an opening called the transcription bubble. Here, single stranded DNA is exposed for RNA polymerase II to use as a template for RNA synthesis. Other transcription factors contained in the PIC are TFIIA, TFIIB, TFIID, TFIIE and TFIIF. A diagram illustrating the steps of assembly of the PIC is shown in **Figure 1.11** on the previous page.

The promoter sequence that is minimally required for the assembly of the PIC is referred to as the core promoter region. It consists of the TATA-box and the transcription start site in addition to potentially unidentified binding sites that help to recruit the PIC. The boundaries of this region are not clearly defined, since the surrounding sequence plays an important role in determining the efficiency of the TATA-box[29]. Generally, A/T rich regions are found around strong core promoter regions. There is evidence that this aids in keeping the promoter region free of nucleosomes, which increases transcription rates by keeping the promoter accessible to transcription factors[30,31].

**The Transcription Start Site and the Initiator element** The location of the TATA-box in promoters in *S. cerevisiae* is almost always upstream of the start codon of the corresponding open reading frame, as highlighted by the analysis shown in **Figure 1.12**. This indicates that this motif is not likely to be functional when placed downstream of the start codon (i.e. it has a neutral function) or that it actually would have a negative effect when placed in this position, potentially interfering with transcription.

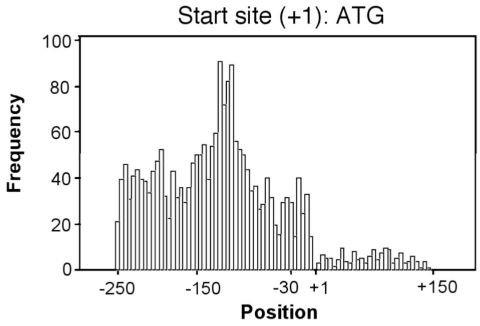


**Figure 1.12:** TATA-box location relative to the start codon of genes in *S. cerevisiae* that have TATA-box containing promoters. Figure adapted from Yang *et al.* 2007[27].

The analysis also shows that the peak occurrence of a TATA-box is at position -135 from the start codon (i.e. 135 bp upstream of the ATG codon), but in general the TATA-box position can vary considerably in relation to the start codon. In metazoan genes, transcription is thought to initiate strictly at a focused initiation site roughly 30 bases downstream of the TATA-box, called the initiator (INR) element[32]. In *S. cerevisiae*, however, transcription initiation does not necessarily occur at a single focussed site. Although the INR element (sequence: YYANWYY) has been found in some yeast promoters[27], there is some controversy as to whether it is involved in transcription initiation, with some authors refuting its existence in the first place[29]. What is clear, however, is that transcription initiation in yeast is often diffuse and located significantly further downstream of the TATA-box, compared to what is seen in metazoans. In fact, it has been shown that in *S. cerevisiae*, Transcription Start Sites (TSSs) located closer than 50 bp from the TATA-box perform worse than those located between 50 and 140 bases downstream[29]. The dispersed nature of transcription initiation in yeast is also reflected in the fact that a wide variety of TSSs have been reported. They include RRY$\underline{\text{R}}$R, TC$\underline{\text{R}}$A[33], Y$\underline{\text{A}}$WR[34] and A(A$_{rich}$)$_5$NY$\underline{\text{A}}$WNN(A$_{rich}$)$_6$[35] (underlined bases indicate the first base of the transcript).

In yeast, it has been well established that Pol II does not interact with the TSS when initially bound in the PIC. This in contrast to the case in metazoans where the footprint of the PIC overlaps with the TSS. Therefore, in yeast, Pol II has to scan down the promoter sequence in order to find the TSS[36]. Depending on the strength of the first TSS, Pol II will continue scanning looking for sites occurring further down if none suitable are found. Whether Pol II stays attached to the PIC during the scanning process currently remains unclear. Once transcription begins, Pol II is untethered and a new Pol II enzyme can be loaded into the PIC to enable a new round of transcription.

**Activation and UAS elements** The strength of the basal expression driven by core promoters is determined by two features of the core promoter sequence: (i) the degree to which the binding sites that direct the PIC to the core promoter resemble their consensus sequences, and (ii) the local DNA sequences surrounding these binding sites (which influence the access and stability of any binding). While these features can modulate the basal expression level from a promoter, they do not directly allow any specific regulation. Instead, different transcription factor binding sites are responsible for promoter regulation in yeast. Typically, the sites controlling regulation fall into two classes, namely Upstream Activation Sequences (UASs) and Upstream Repression Sequences (URSs).

Unlike in bacteria, where sequences modulating the activity of promoters are generally found in the direct vicinity of the core promoter region, in eukaryotes regulatory sequences can be hundreds or even up to several thousand base pairs away from the core promoter. They act as binding sites for proteins that directly or indirectly modulate the activity seen at the core promoter region, for example by helping to recruit Pol II and other factors.

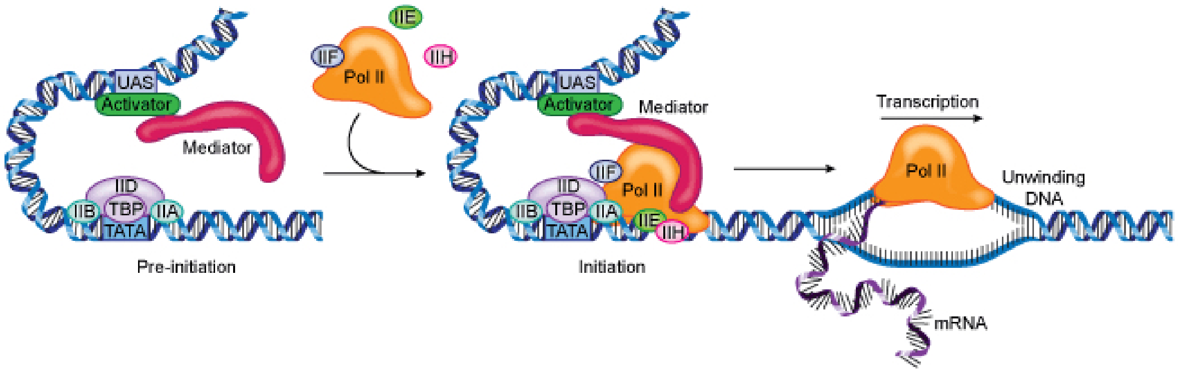


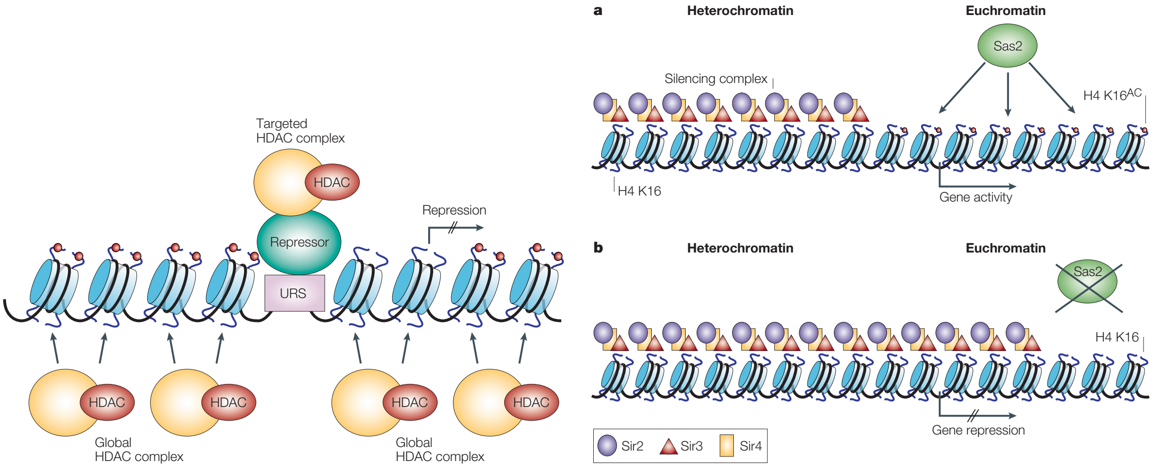
**Figure 1.13:** Process of transcription activation by an upstream activation sequence and bound transcription factor, facilitated by the mediator protein. Figure from mutagenix 2016[37].

It has been postulated that the majority of regulatory processes in yeast happen through activating interactions, rather than repression, and activation of gene expression from TATA-box containing promoters is generally directed by a UAS element[38]. The UAS is bound by specialised transcription factors that recognise the regulatory DNA sequences and then recruit proteins to aid in the assembly of the PIC, make it more active, or to stimulate transcription via other methods. Because activation sites can be located a considerable distance upstream of the core promoter, the DNA often needs to physically bend and loop around to allow the activating transcription factors bound to the UAS elements to exert their effect on the PIC. This process is typically facilitated by proteins called mediators, which interact with both the UAS-bound transcription factors and the PIC. **Figure 1.13** on the previous page shows a visual representation of this process.

Although this is a simple concept in principle, countless variations on this exist, allowing for intricate regulation patterns. For example, the binding of multiple transcription factors (TFs), of the same or of different kinds, can additively increase the strength of the output of a promoter. However, the binding of additional transcription factors may also change the stiffness of the DNA, and either positively or negatively affect the action from UAS elements that require DNA looping for their activity. Furthermore, phosphorylation modifications on a TF can modulate its effectiveness and some transcription factors may stimulate or inhibit the binding of other transcription factors, allowing for integration of signals from different sources. Even the physical orientation of the TF binding sites on the DNA double-helix has also been shown to be important. TFs that have binding sites out of phase with respect to the helical periodicity of the DNA have been shown to perform better at upregulating gene expression than those spaced in phase with the DNA helix[39]. Clearly, the position in 3D space in which the TF is presented to the PIC plays an important role in regulation processes. These mentioned examples are not exhaustive but illustrate the breadth of solutions possible for implementing regulation via the manipulation and combination of upstream activating sequences. For synthetic regulation, promoter UAS engineering is a powerful yet complex opportunity.

**Repression and URS elements** Many of the principles described above for promoter activation are also valid for repression. Sites have been identified upstream of the core promoter region that can be bound by repressor transcription factors that reduce the activity of a promoter. These sequences are called Upstream Repression Sequences (URSs). A variety of mechanisms is known by which repressors exert their function[40] and this has resulted in a distinction between repressors that act at short range, medium range and long range. Each type will be discussed briefly below.

In bacteria, repression is often seen as competitive binding to a binding site that would otherwise be bound by an activator. In eukaryotes this type of repression is not very common. Instead, repressors often bind close to a UAS and act to quench the activating TF that binds there through direct interactions[41]. Another direct and short range method for repression is through steric hindrance at the PIC. While this is not commonly seen in natural yeast promoters, it has been used as a strategy in synthetic promoter design. When DNA binding domains are targeted to be within the core promoter region, they repress expression, presumably by acting



**(a)** Local repression of a promoter through histone modification.

**(b)** Spreading of chromatin over longer distances along the genome, facilitated by the SIR family of silencing factors.

**Figure 1.14:** Diagrammatic illustration of medium and long-range repression via histone modifications and chromatin condensation. Figures reproduced from Kurdistani and Grunstein, 2003[50].

as a physical block preventing PIC assembly or Pol II scanning[42,43].

Another short range repression type occurs through interactions with the basal transcription machinery such as the TBP[44]. This interaction is more than steric hindrance; for example, Mot1 repression is dependent on ATP hydrolysis[45,46]. This type of interaction prevents the transcription machinery from initiating transcription. Ume6, Sin3 and Mxi1 have also been reported to belong to this class of repressors[47–49]. Interestingly, recent work has shown that the action of one of these repressor proteins, Mxi1, can be redirected to other sequences by synthetically tethering these to generic DNA binding domains[19]. This prospect allows for a host of opportunities for advanced synthetic regulation based on the specific properties of this repressor tethered to a DNA binding domain of choice.

Repression at medium distance is conferred by repressors that act not on the specific DNA sequence or TFs that bind it, but instead act on the chromatin that surrounds and packages the DNA inside the yeast nucleus. As in all eukaryotes, yeast DNA is wrapped into nucleosomes consisting of histone proteins and these offer a whole new level of sites for gene regulation control. Typically, the binding of histones to a core promoter region strongly represses transcription as the nucleosome positioning blocks access to the sequences where the PIC can assemble. Medium range repressors typically modify the nucleosomes bound around the promoter of a gene. Hir1 and Hir2 are Histone Regulatory genes that form a regulatory complex that promotes histone deposition onto DNA[51]. Histones occupy 147 basepairs of DNA each, which accounts for the increased distance over which these repressors act. A large class of repressors seems to be a combination of short range repression by TBP interaction and medium range repression through histone interactions. To what degree each of the different repression mechanisms contribute to repression is a matter of debate and seems to depend on the specific binding site[40].

Long range repression also involves histone modification. On top of promoting local condensation of the chromatin (like medium distance repressors) they also establish compacted chromatin that then spreads linearly along the DNA. This process is achieved through histone deacetylation by a family of Silent Information Regulator (SIR) proteins. The spreading considerably widens the effect of the repression, potentially affecting neighbouring genes. This effect is visualised in **Figure 1.14** on the preceding page.

**Promoter engineering** Although our understanding of yeast promoter architecture and functions are incomplete, the knowledge available today allows researchers to make rational modifications to existing promoters and even construct novel promoters from scratch. In both cases, this is especially relevant to synthetic biology, as promoters designed based on prior knowledge are likely to be more predictable and thus easier to implement as standardised modular parts.

An example of a modified yeast promoter that was developed and used in a synthetic biology study is given by the work performed by Ellis *et al.* on the GAL1 promoter[43]. This study is particularly relevant to the work in this thesis. The GAL1 promoter in *S. cerevisiae* has traditionally been a showcase of a tightly regulated promoter which modulates downstream gene expression in response to carbon-source availability. Nearly all aspects of yeast promoter architecture that are described above are united in this promoter and it can be considered canonical in many ways. It contains:

- A canonical TATA-box (TATATAAA).
- A separation of 90 base pairs between the TATA-box and TSS.
- A TSS that is canonical to the A$(A_{rich})_5$NY$\underline{A}$WNN$(A_{rich})_6$ motif.
- A UAS bound by the transcription factor Gal4p that activates in response to galactose.
- Two URS sites, each bound by transcription factor Mig1p and repress in response to dextrose.

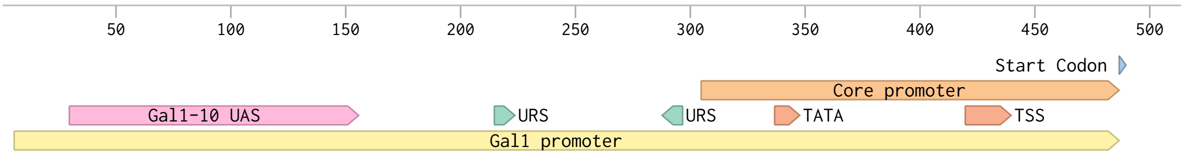


**Figure 1.15:** Organisation of sites on the yeast Gal1 promoter. The top ruler denotes the bp length of the DNA sequence of the promoter. TATA: the location of the TATA-box. TSS: the location of the transcription start site. Gal1-10 UAS: region containing 5 Gal4p binding sites, two of which are canonical to CGG$N_5$(T/A)$N_5$CCG[52] and the remainder containing a single mismatch to this sequence. URS: Mig1p binding site. Note that Gal1-10 is a bidirectional promoter and only the section involved in Gal1 transcription is shown.

These features are visualised in **Figure 1.15** and together they ensure very tight regulation of this promoter. It is normally in an off-state in rich media containing dextrose, but when the alternative carbon source, galactose, becomes prevalent it is activated by transcription factors. When galactose is sensed by the cell, the promoter is induced by Gal4p and turns on to direct one of the strongest levels of transcription known in yeast. Since activation of this promoter is so strong, any transcription from it is costly, so therefore it is also repressed tightly in the off-state, when dextrose is available. The Mig1p repressor responsible for this process induces silencing through chromatin in the presence of dextrose. Derepression is a stochastic process and can take 3 hours in some cells and 10 hours in others. Good regulation of GAL1 is essential for *S. cerevisiae* as the cell needs to have very accurate control of gene expression in response to changes in the carbon source in order to allow competitive growth and survival in different conditions. The detailed information available for the regulatory sequences in the GAL1 promoter allows for targeted edits to be made with the aim of adding functionality required for synthetic biology.

In the cited work, the native GAL1 promoter from yeast was modified to create two libraries of externally-inducible promoters. These were created by targeted mutagenesis of the core promoter region between the TATA box and the TSS. In addition to the promoters naturally having repression by dextrose and induction via galactose, further regulation was engineered into these promoters to enable external induction by small chemical molecules that are not otherwise recognised by yeast. To do this, the native sequences in the GAL1 core promoter between the TATA-box and TSS were directly replaced with sequences that are tightly bound by two widely used bacterial transcription factors, TetR and LacI. The LacI repressor binds tightly to its operator site on the DNA. When Isopropyl $\beta$-D-1-thiogalactopyranoside (IPTG) is introduced into the media, however, conformational changes cause it to be released from the DNA. Similarly, TetR is unable to bind to its operator sequence in the presence of Anhydro Tetracycline (ATc). Repression by each of these repressors can therefore be regulated externally using IPTG and ATc. This is also shown in **Figure 1.16**.

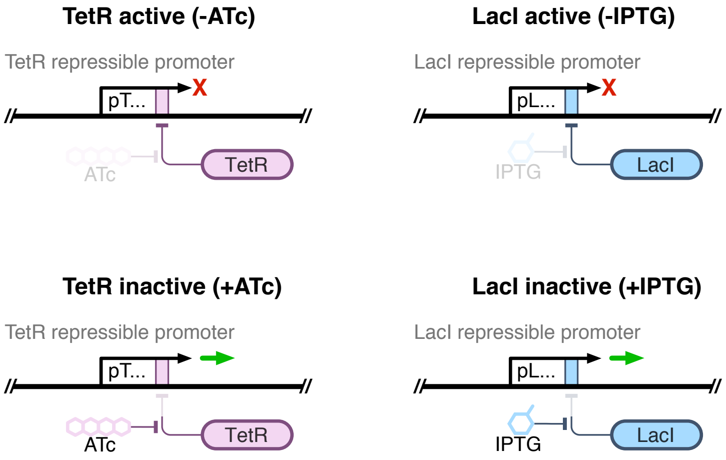
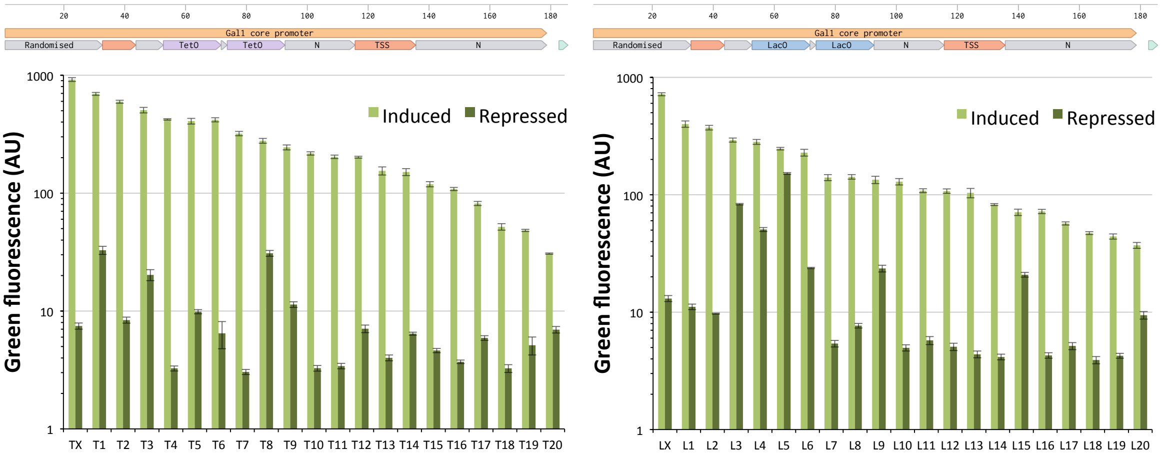


**Figure 1.16:** Repression of GAL1-based promoters by binding of TetR and LacI to their operator sites engineering into the core promoter region. Repression is relieved by addition of ATc and IPTG small molecule inducers.

Addition of the consensus operator sequences of these two repressor proteins to the GAL1 core promoter region does not significantly affect the output of the promoter when in its ON state with galactose present. However, if the TetR and LacI proteins are themselves heterologously expressed in yeast (e.g. from a constitutive promoter), they then confer specific repression of modified GAL1 promoters in all conditions by blocking the PIC through steric hindrance. It is assumed that neither TetR, LacI or their inducer molecules confer any other regulatory action in yeast, and so this design allows direct and specific control of the gene expression from this strong engineered GAL1 promoter. The two repressors bind orthogonal operator sites which allows each to specifically repress its own promoter individually and their inducers to only activate expression of the desired promoter.

With this initial promoter engineering producing two new modified GAL1 promoters linked to external regulation, these were next used as the basis for generating two libraries of promoters with different strengths of expression in their on- and off-states. All DNA sequence in the core regions of these promoters that did not encode an operator, the TATA box or the TSS were randomised by re-synthesising the DNA with degenerate bases in these locations. The resultant promoter libraries contained a vast diversity of different sequences between the defined sites and the base pair changes in these mutated sequences acted to alter the efficiency of the adjacent sites in defining the promoter outputs. After measuring the gene expression output in induced and repressed conditions for hundreds of variants of the promoters generated by the mutagenesis, a set of 20 promoters with a wide range of induced expression strengths were identified that were repressed at different degrees by the presence of the two repressors. These promoters were named pT1 to pT20 and pL1 to pL20 for the TetR and LacI repressible libraries, respectively. In addition, the pre-mutation versions of these repressible promoters, named pTX and pLX, were also included in the libraries. The libraries and the configuration of their core promoter regions are shown in **Figure 1.17**.

Members of the two libraries described above are used extensively throughout the work in this thesis. However, many more successful attempts have been made to modify and engineer regulated promoters in synthetic biology which are now offering many new tools. Recent work



**(c)** Library of TetR repressible promoters.

**(d)** Library of LacI repressible promoters.

**Figure 1.17:** Mutations in the core promoter region of engineered yeast GAL1 promoters containing TetR and LacI repression sites leads to the generation of externally inducible promoter libraries[43]. Grey regions in the core promoter area of the GAL1 promoters indicate sequences randomised by synthesis to generate mutation. Orange regions indicate core promoter elements left unmutated: TATA-box and Transcription Start Site (TSS) and purple and blue regions indicate the repressor binding sites. Characterisation of the strength of expression from the 20 promoters from each library using a green fluorescent protein output determines their on and off characteristics. The two libraries were sorted by expression strength: the strongest being pT1 and pL1 and the weakest being pT20 and pL20 for the respective libraries. pTX and pLX are the unmutated promoters with no modifications except for the repressor binding sites replacing the native yeast sequence.

by Redden *et al.* used design principles, combined with cell-sorting for expression output to generate libraries of synthetic minimal TATA-box containing promoters[53]. Earlier work from the same group examined how the output of synthetic yeast promoters could be tuned simply by small changes to the base sequence which affect the promoter nucleosome architecture[30]. Similar work using extensive mutation and screening of the core sequence of a few strong promoters in yeast has also enabled us to move towards a sequence-to-output understanding[29]. For example, it is now known that A- and T-rich sequences in a yeast core promoter promote stronger levels of transcription. This is especially true when located in the region 75 bp upstream and 50 bp downstream of the main transcription start site[31].

### TATA-less promoters

While TATA-less promoters make up the majority of those seen in the *S. cerevisiae* genome, our understanding of how the sites and sequences within these contribute to their function is much less understood than for TATA-box containing promoters. This is largely because most studies of yeast promoters and almost all attempts to re-engineer them focus on regulation and most TATA-less promoters are unregulated. It is perhaps easiest to visualise TATA-less promoters as regions of DNA that have evolved to enable good unregulated access for Pol II. Either by local active recruitment or by diffusion, these promoters are bound by the PIC and begin transcription in most conditions.

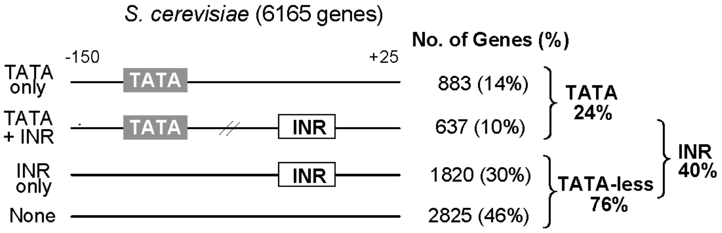


**Figure 1.18:** Prevalance of the the TATA-box and INR motifs in yeast promoters. Figure from Yang *et al.* 2007[27].

In yeast, transcription initiation at TATA-less promoters is thought to be much less defined than for TATA-box containing promoters and often occurs at multiple start sites within a window[29,54,55]. This is in contrast to what is seen in metazoans, where initiation occurs at a fixed distance from the other core promoter elements at TATA-less promoters . In yeast, only 40% of all promoters have the metazoan initiator element known as INR which may aid in defining the start-site[27]. The INR element is thought to possibly recruit the PIC to the promoter by itself, acting as both a TATA-box and TSS in one (see **Figure 1.18**). Interestingly, even in TATA-less promoters the TBP is also known to be involved in the PIC[56]. Another potentially important conserved DNA element, the GA element (GAE), is also possibly a key site in TATA-less promoters as this has been shown to generally not co-occur in promoters with the TATA box[57].

Work by myself and other members of the Ellis lab prior to this thesis investigated one particular TATA-less promoter in the context of yeast synthetic biology. The PFY1 promoter that drives constitutive expression of the cytoskeleton component profilin was selected as the basis for a promoter library due to its reliable expression in a wide variety of conditions. This promoter is relatively short in sequence (less than 400 bp) and only contains two sites of note: a REB1p binding site, followed by an A/T rich run that is thought to force a kink in the DNA secondary structure. Downstream of these two sites lies a core promoter region where presumably the PIC constitutively binds and initiates transcription. It is thought that the two upstream sites simply act to prevent nucleosomes being positioned unfavourably in core region and blocking the PIC[58].

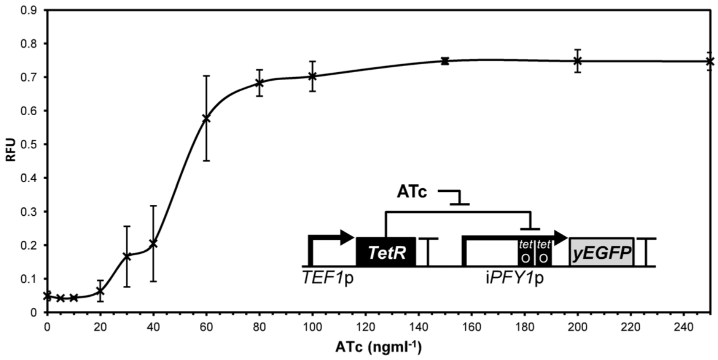


**Figure 1.19:** Dose-response of the TetR-repressible iPFY1 promoter in response to a range of ATc concentrations. Figure from Blount *et al.* 2012[42].

By extensively mutating the bases in the core region of this small promoter, we were able to generate a large library of 48 characterised variants of pPFY1 with constitutive expression levels ranging from very weak to medium in strength. As with the above GAL1 promoter engineering, the mutation of bases away from the main binding sites presumably alters expression output by changing the efficiency of binding and progression of the PIC. As well as randomising bases in this core sequence we also introduced the operator sequences for TetR (as had been done for pGAL1). This enabled us to turn this constitutive TATA-less promoter into a regulated promoter, which we called iPFY1 promoter. This promoter can be switched from OFF to ON by addition of increasing amounts of ATc to cells that are co-expressing the TetR protein[42]. This is shown in **Figure 1.19**.

### 1.3.3 Terminator sequences

While often overlooked compared to the promoter region, the terminator sequence at the distal end of a transcriptional unit is also important for defining the output of expression from a gene. The Transcription Termination Site (TTS), also known as the poly(A)-signal, determines the end of transcription, directing the release RNA polymerase from the DNA and promoting the attachment of a series of adenine residues (poly(A)) to the pre-mRNA. These adenines are not directly coded for on the genome, rather they are added during a separate event at the end of normal transcription in order to aid in the maturation of the nascent DNA into a usable mRNA that directs translation. The two functions of the terminator - to end transcription and promote polyadenylation - are complex and interlinked processes that may be a challenge to deconvolute at the sequence level[59]. There are two competing models for RNA polymerase II release that probably both tell parts of the same story[59,60]. Without going into full details, the evidence suggests that Pol II release at a terminator is an indirect result of the RNA modification processes such as polyadenylation that are initiated by the terminator sequences.

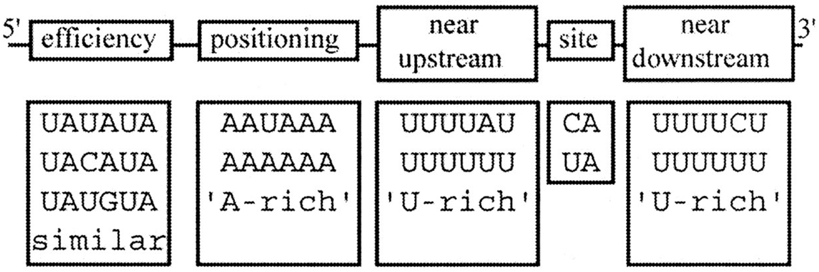


**Figure 1.20:** Consensus sequences of modules required for an efficient transcriptional terminator in yeast. Figure replicated from Graber et al.[62].

Despite the inherent complexity of TTS, the sequences that regulate 3' end formation have been well characterized by several studies[61,62]. The code has been shown to be relatively degenerate, with single nucleotide substitutions usually unable to completely abolish termination. Nevertheless, a consensus sequence has been derived what typically constitutes a yeast terminator. This consensus is built from 5 different modules that are either A-rich or T-rich at the DNA level (and therefore U-rich at the RNA level). These modules are shown in **Figure 1.20** and are thought to be mostly involved in directing the processing of nascent transcript into the mRNA, and not necessarily the actual process of transcription termination at the DNA/RNA interface. Recent work on designing and mutating these modules has led to the creation of libraries of synthetic terminators for yeast[63,64]. These can be used to fine-tune gene expression, but actually act post-transcriptionally. The different terminator sequences are thought to modulate the efficiency of the RNA being processed into a mature mRNA, the lifetime of that mRNA, how easily it is translocated to the cytosol and how efficiently it is translated into its protein product.

Interestingly, many of the more commonly used terminators in yeast research and biotechnology, such as the CYC1 and ADH1 terminators, act as efficient poly-(A) sites in both directions. That is to say, they will end transcription when placed in either the forward or reverse orientation. Bidirectional terminators are often found at locations between convergent genes and seem to contain two unidirectional recognition sequences to ensure all transcription from both strands is terminated at these sites[65,66]. Unidirectional terminators are much less well used in synthetic biology but studies have uncovered some like GCN4 and PHO5 which only terminate transcription from one DNA strand[67] which could potentially be exploited in future synthetic circuits with bidirectional transcription.

## 1.4 Co-transcriptional processes in yeast

The central dogma of DNA making RNA to make protein often overlooks the myriad of complex steps that cells take in this process. In eukaryotes, there is a large number of post-transcriptional steps, before a mature mRNA is presented to the ribosome in the cytosol for translation to begin. While full discussion of all of these processes is beyond the scope of this thesis, it is worth briefly mentioning the different operations that occur as each provide potential ways to modulate gene expression in yeast if fully understood. A full and excellent review of our understanding of these process was published by Hocine *et al.* in 2010[68] and is summarised in **Figure 1.21** on the next page.

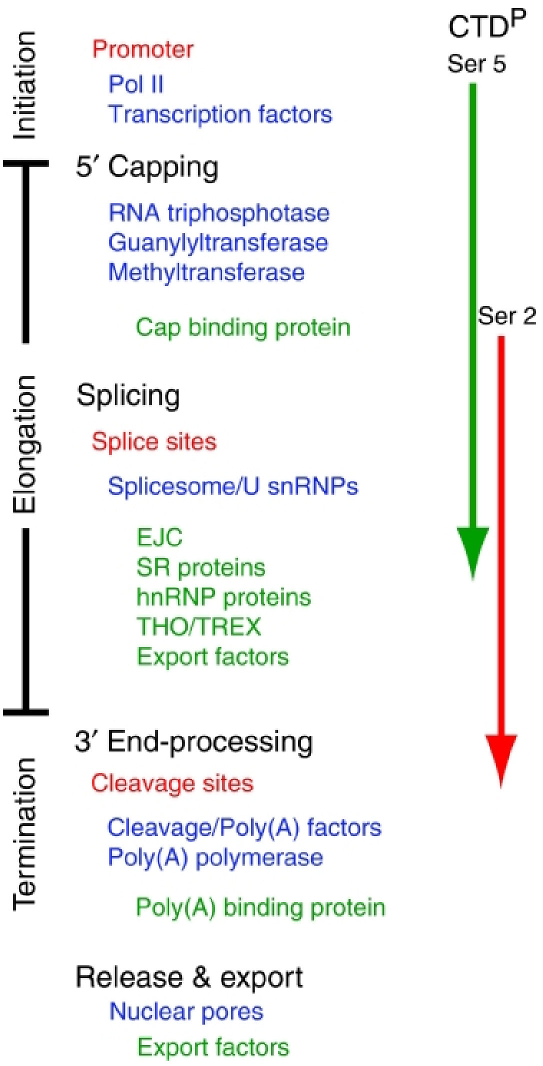


**Figure 1.21:** Summary of the post-transcriptional processes involved from transcription through to translation. Figure reproduced from Hocine *et al.* 2010[68].

### 1.4.1 Pol II CTD Phosphorylation

A key player in co-transcriptional processes in eukaryotes is a domain of Pol II known as the C-terminal domain (CTD) which is thought to be a long chain of amino acids that can be differentially phosphorylated at multiple sites during transcription[69]. The CTD can be considered to be a tail behind the polymerase that is modified by phosphorylation by transcription-associated factors in order to store a memory of signals and pass these on to other factors elsewhere during the process of producing the RNA[70]. The CTD coordinates the roles of many different co- and post-transcriptional processes in the yeast nucleus and is known to be bound by over 100 different yeast proteins in its phosphorylated state. The phosphorylation marks that are placed on the CTD are essential for the proper progression of transcription and are required for coordinating RNA processing events and for altering the states of the histone proteins in the nucleosomes of the DNA too. Effectively, the CTD is surveying transcription as it occurs and only recruiting the factors that mature the RNA into to an mRNA if the transcription progresses as expected.

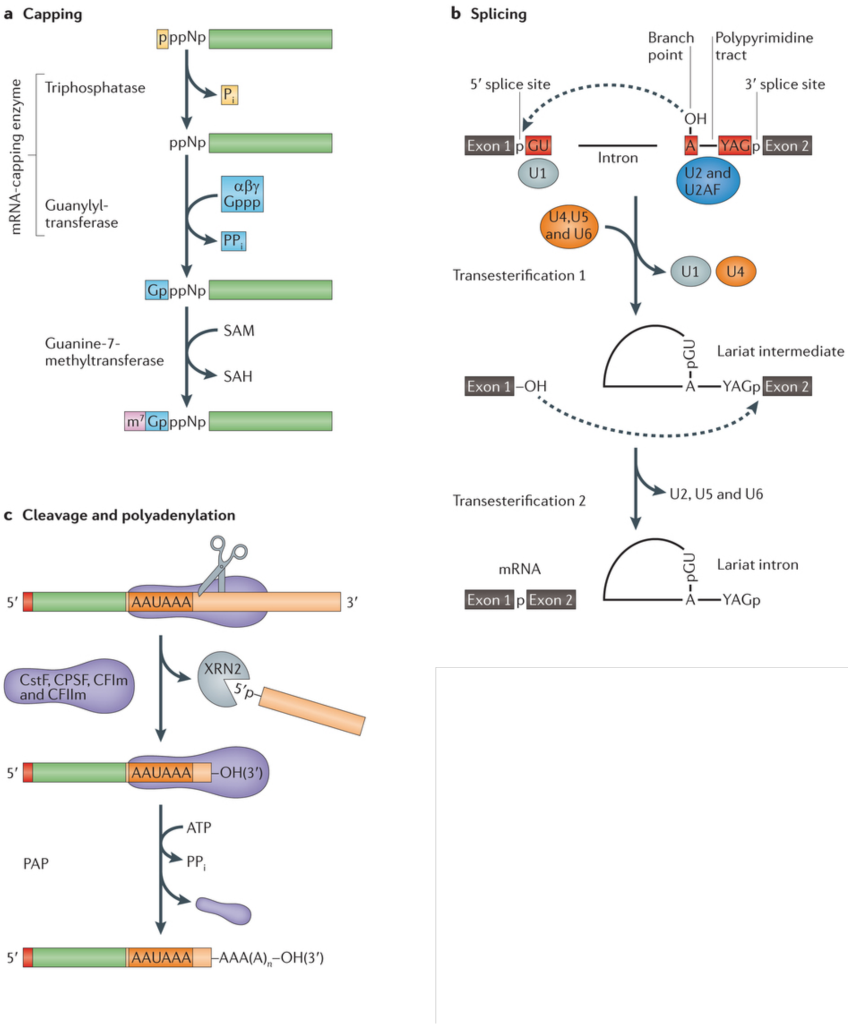


**Figure 1.22:** Diagrams illustrating the key steps of the three main co-transcriptional processes in yeast, capping, splicing and polyadenylation. Figure reproduced from Bentley *et al.* 2014[71].

### 1.4.2 5’ Capping

The first of the three main co-transcriptional processes directing by the CTD state is 5’ capping of the RNA, which is illustrated along with splicing and polyadenylation in **Figure 1.22**.

The RNA is capped at its 5’ end by three enzymes: RNA triphosphatase, guanylyltransferase and 7-methyltransferase. These act on early in transcription, usually after Pol II has transcribed the first 25-30 nucleotides of the RNA. They covalently attach GTP to the RNA, resulting in a methylated GpppN 5’ cap. The enzymes involved are typically bound to Pol II as transcription begins, so perform the modification early on in transcription. As the nascent transcript gets longer and becomes hundreds of bases in length, co-transcriptional phosphorylation of the polymerase and its CTD causes these enzymes to disassociate from the main complex. It is currently thought that the main transcription complex is held back within a 100 bases of the promoter until capping occurs and once capping is successful, the polymerase switches into an elongating mode[68].

Capping effectively stabilises the transcript and aids in many ways to ensure it matures into an mRNA and is effectively translated. A recent excellent review by Ramanathan *et al.*

(2016) covers these roles in extensive detail[72]. Briefly, however, the main immediate effects it has in the nucleus are to prevent degradation of the RNA by the many exonucleases present in a cell[73] and to promote excision of any 5' proximal introns by the splicing machinery (see below). It then plays a major role in nuclear export, where cap-binding proteins ensure that the processed mRNA can be fed out from the nucleus into the cytosol[74]. The cap then finally plays a key role in translation initiation in the cytosol (covered further below). Protein complexes involved in translation initiation recognize the cap before translation and the cap also helps circular mRNAs by interacting with polyA-binding protein (PAB1). Circularisation of mRNAs may aid their nuclear export and are also thought to promote in translation reinitiation to enhance further protein synthesis from a bound ribosome.

### 1.4.3 Intron splicing

Co-transcriptional RNA splicing is a complex process to remove the introns from nascent RNAs to create an mRNA consisting of the exons of a gene[75]. While splicing is common to all eukaryotes, it occurs far less frequently in *S. cerevisiae* than almost any other eukaryotic cell. In yeast only 283 of the around 6000 genes contain any introns at all, and only a handful of these have been shown to be essential[76]. Most of these are comparatively small in length, except for longer introns in ribosomal encoding genes[77].

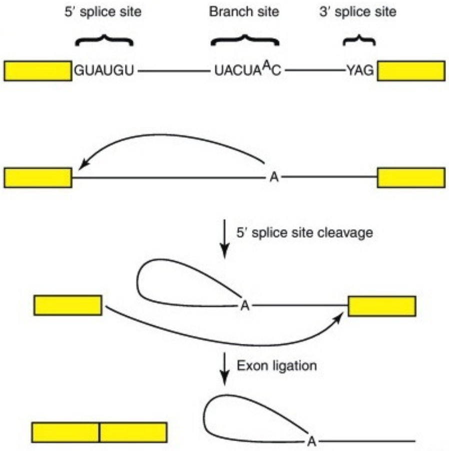


**Figure 1.23:** Key sequence motifs involved in the mechanism of yeast RNA splicing. Figure reproduced from Semlow *et al.* 2012[78].

Despite the lack of introns in yeast, splicing in this model organism has been extensively studied and many of the proteins and *cis* acting elements involved have now been elucidated[79–81]. Splicing of the nascent RNA occurs during transcription and typically coincides with the RNA region that encodes the intron exiting from Pol II complex as it passes down the DNA[82]. In yeast, there are three key sequences involved in splice site recognition which are illustrated in **Figure 1.23**. Firstly, there is a 5' recognition sequence: GUAUGU. The cleavage site is just prior to the first guanine residue of this site, making this guanine the first residue of the intron. This residue is then covalently attached to the last adenine in the branch point sequence: UACUAAC. The attachment results in the formation of a lariat structure. Finally, the end of the first exon is covalently attached to the beginning of the second exon at the end of the 3' recognition sequence with consensus YAG. The RNA lariat is thus removed from the mRNA.

There are other sequences that also have an effect on splicing recognition and efficiency. In higher eukaryotes and fission yeast (*S. pombe*) there is a Poly(Y) tract between the branch point and the 3' splice site. In budding yeast this tract also exists, but seems to only consist of uracil bases[83]. Additionally, it has been found that exon sequences themselves can also influence splicing efficiency to some extent[84]. However, these features seem to be of secondary importance compared to the strong influence of the three main splicing recognition sites.

Interestingly in yeast, the distribution of intron lengths is bimodal as shown in **Figure 1.24**. There is a peak around 100 bp for non-ribosomal protein genes and a second peak around 400 bp for the ribosomal protein genes[77]. Ribosomal protein genes represent a significant fraction of the 287 known intron-containing genes in yeast and account for the vast majority of mRNA produced by intron-containing genes[80,85]. Experiments in higher eukaryotes have suggested that the efficiency of splicing of introns larger than 250 bp is significantly reduced[86]. Given the high rate of transcription of ribosomal genes, this does not seem to be a strict limit for yeast.

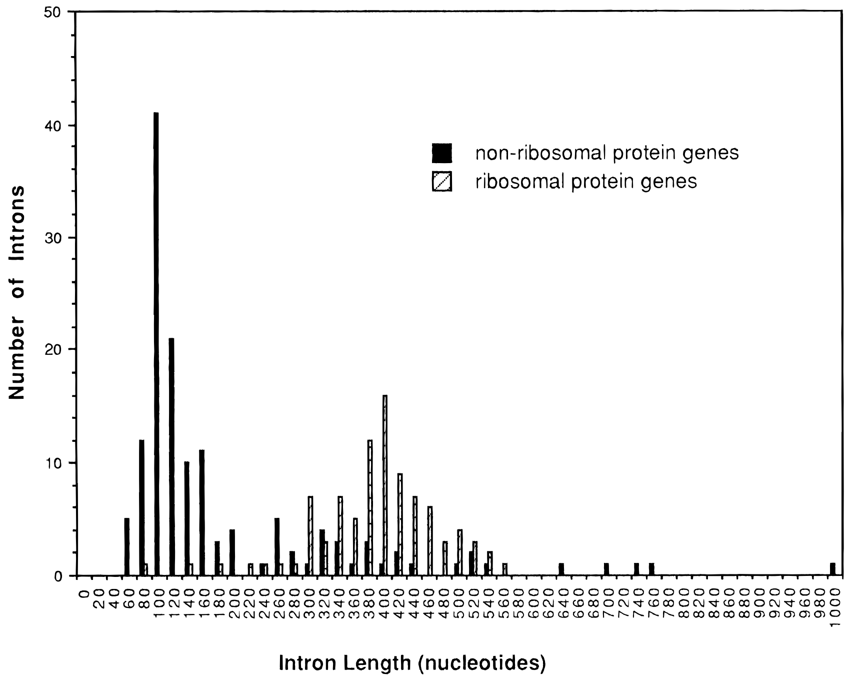


**Figure 1.24:** Distribution of lengths of all known introns in *S. cerevisiae*. Data from Spingola *et al.* 1999[77].

In higher eukaryotes, multiple long introns and small exons are the rule rather than the exception, which is the converse for yeast. When intron length exceeds 250 bp, intron recognition shifts from recognition across the intron to recognition across the exon[87]. In *S. cerevisiae*, only a small minority of intron-containing genes have more than one intron, so recognition across the exon is unlikely to occur like it does in higher eukaryotes. Potentially increasing the efficiency of long intron splicing could be achieved by optimizing the secondary structure of the pre-mRNA. It has been shown that mRNA folding is a significant factor in 3' splice site recognition[88,89], and it has also been suggested that long intron recognition depends on pre-mRNA folding that causes the splice sites to become spatially adjoined[90]. Placing inverted repeats in the mRNA exon sequence near the splice sites may therefore increase the splicing efficiency by promoting RNA folding that effectively reduces the distance between the 5' and 3' recognition sequences.

### 1.4.4 Polyadenylation

The final major co-transcriptional RNA processing occurs during termination of transcription, by polyadenylation of the RNA 3' end. In many ways, the functions of polyadenylation are very similar to that of 5' capping: it prevents enzymatic degradation from nucleases, promotes nuclear export[91], promotes pseudo-circularisation of the RNA and aids in translation[92]. The motifs of the terminator sequence described earlier act to recruit the polyadenylation enzymes during transcriptional termination and also aid in the binding of the poly(A) tail by various poly(A) binding proteins involved in mRNA transport and degradation. It is thought that poly(A) tail length in *S. cerevisiae* is controlled by the motif sequences in the mRNA which control how much de-adenylation occurs after the adenylation is added[93]. Sequences that promote de-adenylation can lead to mRNAs with quicker degradation rates due to shorter poly(A) tails.

The role of the poly(A) tail in translation is also important. Research in *S. cerevisiae* has shown that the poly(A) tail can act as an independent 'translational promoter', delivering ribosomes to uncapped mRNAs even if their 5' end is blocked. Normally when mRNAs compete for ribosome binding, neither the cap structure nor the poly(A) tail alone is enough to drive efficient translation and instead the 5' cap structure and the poly(A) tail act synergistically to initiate translation at the start codon nearest the 5' end[94]. Interestingly, the fact that the poly(A) tail on its own can function as an independent initiator of translation means that it is possible for it to recruit ribosomes to internal initiation sites within an mRNA[92]. Alternative and internal translation initiation sites are rarely seen in yeast mRNAs.

## 1.5 Translation in yeast

A mature mRNA translocated into the cytosol is translated into its protein product by the ribosome. While the standard model of translation where codons are matched to amino acids is widely known, there are also many other factors involved in ensuring efficient and accurate translation. Sequences within the mRNA play a role in determining how these other factors play out and can direct the rate and efficiency of translation and thus define the final expression level of the protein produced. Perhaps the key part of determining the strength of translation of an mRNA is how efficient and rapid the initiation is. The current model of translation initiation in yeast is called the ribosome scanning model. In this process, translation initiation factors form a complex at the 5' cap with the small ribosome submit and then scan along the mRNA to identify the first readily accessible AUG start codon[95]. Once this is found, the large ribosomal subunit is brought to the mRNA to form the full ribosome and translation then begins.

The process of translation initiation is mediated by a complex of initiation factors. In eukaryotes this complex is called eukaryotic Initiation Factor 4F (eIF4F). This complex consists of the proteins eIF4A, eIF4E and eIF4G, and some of these proteins are present in most eukaryotes as one or more homologs. eIF4E binds the 5' cap of the mRNA and is believed to be involved with RNA looping[96], binding together with eIF4G, which acts as a scaffold protein interacting with a variety of other initiation factors and also to the poly(A)-binding proteins that bind the mRNA poly(A) tail. eIF4A is DEAD-box RNA helicase that aids in the unwinding of mRNA secondary structure as the ribosome small subunit scans the mRNA for the first AUG codon[97,98]. Its activity is known to be increased through its interactions with eIF4G and eIF4B[99,100], and specifically in *S. cerevisiae* domains of eIF4G are known to strongly bias the RNA unwinding activity to target duplexes with 5' overhangs[101].

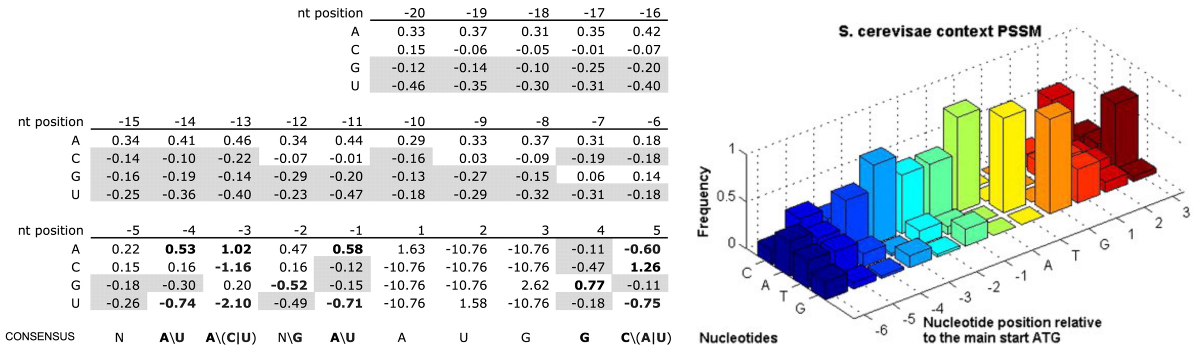
Many features of the mRNA can play a role in determining the efficiency of this process and thus the rate of translation initiation. As mentioned above, both the 5' cap and poly(A) tail are important components that begin this process. The details of how other sequences within the mRNA affect translation are briefly given below.

### 1.5.1 5'UTR and the Kozak Sequence

The sequence of the mRNA between the 5' cap and the first translated codon is known as the 5' untranslated region (UTR). As in prokaryotes, the 5'UTR sequence in yeast is thought to play a role in the efficiency of translation initiation, although how this modulates initiation is not as straightforward as the simple base-pairing interactions seen at bacterial ribosome binding sites. In yeast, the 5'UTR sequences of genes tend to be quite short compared to those seen in other eukaryotes and the median 5'UTR in *S. cerevisiae* genes is 68 bases in length. Interestingly, shorter 5'UTRs tend to be found in housekeeping genes, while longer 5'UTRs tend to be associated with more regulated genes hinting at a role in modulation of gene expression[102]. It has been shown that 5'UTR length is strongly correlated with gene expression, with shorter 5'UTRs typically giving stronger expression levels[103].

Early on in the study of 5'UTR sequences, it was identified that RNA structure within this region could pose a problem for translation initiation by blocking ribosome scanning. Work by Vega-Laso and others identified that RNA folds with base-pairing strength of -28 kcal/mol (Gibbs Free Energy) and lower, pose serious problems for translation[104]. This led to work investigating the RNA secondary structure features within yeast mRNA 5'UTRs to look at how sequences affect translational activity[105]. Research in this area has now produced genome-wide measurements of RNA secondary structures in yeast, such as the Parallel Analysis of RNA Structure (PARS) Scores that were published in 2010[106].

The fact that 5'UTR mRNA structures can act as roadblocks to inhibit translation initiation is presumably why the eIF4F complex includes the RNA helicase factor eIF4A. Indeed, it has been shown that the requirement for eIF4A in translation is in direct proportion to the degree of mRNA 5'UTR secondary structure[97]. Interestingly, in mammalian cell lines, it has been shown that over-expression of eIF4E reduces inhibition from mRNA secondary structure[107]. Presumably, increased levels of this factor promote the eIF4F complex to either unwind any structures with greater efficiency or enable the ribosome scanning to somehow skip past unfavourable motifs. Achieving similar cell engineering to enable yeast to be able to scan through secondary structures is complicated by the fact that any change in the expression of the factors involved in eIF4F disrupts the complex by stoichiometric imbalance and reduces global translation efficiency. However, in yeast a complementary RNA helicase to eIF4A exists, called Ded1, which is considered to be an alternative translation-initiation factor that also helps the 40S ribosome scan the mRNA[108,109]. In experiments with a cold-sensitive Ded1 mutant, it has been shown that its inactivation in living cells substantially reduces the expression of >600 mRNAs, and so presumably its role is to also aid in 5'UTR structure unwinding along with eIF4A[110].



| nt position | -20 | -19 | -18 | -17 | -16 |
|---|---|---|---|---|---|
| A | 0.33 | 0.37 | 0.31 | 0.35 | 0.42 |
| C | 0.15 | -0.06 | -0.05 | -0.01 | -0.07 |
| G | -0.12 | -0.14 | -0.10 | -0.25 | -0.20 |
| U | -0.46 | -0.35 | -0.30 | -0.31 | -0.40 |

| nt position | -15 | -14 | -13 | -12 | -11 | -10 | -9 | -8 | -7 | -6 |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.34 | 0.41 | 0.46 | 0.34 | 0.44 | 0.29 | 0.33 | 0.37 | 0.31 | 0.18 |
| C | -0.14 | -0.10 | -0.22 | -0.07 | -0.01 | -0.16 | 0.03 | -0.09 | -0.19 | -0.18 |
| G | -0.16 | -0.19 | -0.14 | -0.29 | -0.20 | -0.13 | -0.27 | -0.15 | 0.06 | 0.14 |
| U | -0.25 | -0.36 | -0.40 | -0.23 | -0.47 | -0.18 | -0.29 | -0.32 | -0.31 | -0.18 |

| nt position | -5 | -4 | -3 | -2 | -1 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.22 | **0.53** | **1.02** | 0.47 | **0.58** | 1.63 | -10.76 | -10.76 | -0.11 | **-0.60** |
| C | 0.15 | 0.16 | **-1.16** | 0.16 | -0.12 | -10.76 | -10.76 | -10.76 | -0.47 | **1.26** |
| G | -0.18 | -0.30 | 0.20 | **-0.52** | -0.15 | -10.76 | -10.76 | 2.62 | **0.77** | -0.11 |
| U | -0.26 | **-0.74** | **-2.10** | -0.49 | **-0.71** | -10.76 | 1.58 | -10.76 | -0.18 | **-0.75** |
| CONSENSUS | N | **A\U** | **A**\(**C**\|**U**) | N**G** | **A\U** | A | U | G | **G** | **C**\(**A**\|**U**) |

**(a)** Robbins-Pianka[111]

S. cerevisae context PSSM
Frequency
1
0.5
0
C
A
T
G
Nucleotides
-6
-5
-4
-3
-2
-1
A
T
G
1
2
3
Nucleotide position relative to the main start ATG

**(b)** Ben-Yehezkel[112]

**Figure 1.25:** Effect on expression output of different RNA bases in the 5'UTR sequences immediately upstream and downstream of the AUG start codon in yeast mRNAs.

While the most important features for a yeast 5'UTR to promote translation are (a) to be short and (b) to be relatively free of secondary structures in order to enable efficient ribosome scanning, a third component also plays a key role. It has long been recognised that having an A-rich RNA sequence in the mRNA bases just before start codon also aids in efficient translation initiation[113]. This sequence, immediately upstream of the AUG codon, is often referred to as the Kozak sequence and equivalents of this are found throughout eukaryotes. While it is tempting to compare Kozak sequences to prokaryotic RBS sequences, the two act by different mechanisms. In eukaryotes, the A-rich sequence of the Kozak site aids the assembly of the full ribosome to the start codon but not by direct base-pairing, but instead by indirect actions. Many studies have shown that its sequence can significantly alter the expression level of the protein produced. Three recent studies have examined this systematically[111,112,114] and snapshots of the results from two of these are shown in **Figure 1.25**. Based on these results and those of others, it is now typical to include consensus sequences for high translation strength around the AUG codon in most gene designs in yeast synthetic biology.

### 1.5.2 Open Reading Frames and IRES Sequences

The main role of the Open Reading Frame in an mRNA is obvious: it encodes the polypeptide sequence of the protein. Open Reading Frames are bounded by the start and stop codons of the genes which are some of the least ambiguous features in genomes and so are easy to identify using bioinformatics tools. In the vast majority of cases the start and stop codons of a gene can be predicted with single base pair resolution. This contrasts with the difficulties of predicting promoter boundaries, transcription start sites and poly(A) sites in genes.

Aside from encoding the amino acid order of the protein, the open reading frame region of an mRNA rarely encodes any other information that determines the resulting gene expression strength. The degeneracy of the genetic code means that different RNA sequences within mRNAs can lead to the same protein being produced and this has proven useful in synthetic biology for watermarking/barcoding of synthetic genes[115]. Quite why different codons that encode the same amino acid are preferred and enriched in many yeast genes is subject to continual debate. Research on codon optimisation typically looks at the rarity of different tRNAs in yeast and assumes that the most efficient codon to use for strong protein expression will correlate with the cell's most abundant tRNA for that amino acid[116]. However, the picture may be much more complex than that in reality. In some instances it may be preferable to use a rarer codon to potentially pause the ribosome (while it awaits the rare tRNA to arrive) at a boundary between protein domains. This pause could aid in enhanced folding of the domain that has just been translated, ahead of further polypeptide being produced. The use of rarer codons could also give stronger expression in cases where these codon choices help the mRNA to avoid having secondary RNA structures due to internal local base-pairing. Interestingly, in yeast and other organisms rare codons are found to naturally occur more often towards the start of open reading frames (i.e. towards the 5' end). The latest research in this area indicates that this may naturally act to slow translation at the start in order to effectively regulate ribosome density along the mRNA to avoid gaps or roadblocks appearing[117].

An interesting set of sequences found on mRNAs chiefly from viruses that infect eukaryotic cells are the Internal Ribosome Entry Site (IRES) sequences. These are not naturally found in yeast but offer a potentially useful part for yeast synthetic biology. IRES sequences are secondary structures that can be placed within an mRNA that act to load a ribosome directly to an AUG start codon, without it first scanning from the 5' cap, as shown in **Figure 1.26**. A variety of RNA sequences encoding complex secondary structures have been shown in different organisms to allow expression from a second alternative ORF downstream from the standard ORF at the 5'end[118]. In yeast, using an IRES sequence from a hepatitis C virus (HCV) RNA allowed researchers to express two proteins from a single mRNA by encoding a gene with a bicistronic design[119]. However, recent work on analysing and identifying IRES sequences has provided contradictory evidence, showing that the HCV IRES does not work in yeast[120].

The use of bicistronic and polycistronic designs is prevalent in bacterial synthetic biology but rare in eukaryotic work. However, the possibilities offered by bicistronic designs are interesting and indeed we applied IRESs in **subsection 4.3.3** on page 136.

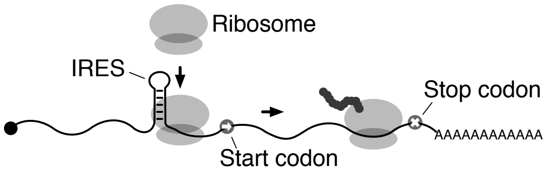


**Figure 1.26:** Diagram illustrating internal ribosome loading to an mRNA ORF by an Internal Ribosome Entry Site (IRES).

### 1.5.3 The 3'UTR and mRNA looping

In yeast, the 3' untranslated region (3'UTR) of an mRNA generally consists of a short stretch of sequence followed by poly(A) tail. Analysis of yeast gene expression has shown that there is no correlation between the lengths of 3'UTRs and the strength of expression levels, in contrast to what is seen for 5'UTRs[103]. As described in the previous sections, the polyadenylation of mRNAs is important for their stability, their nuclear export and aids in translation initiation and ribosome recycling. Sequences within the 3'UTRs have also been identified which play a role in stabilizing mRNA and preventing (or in some cases enhancing) its degradation in the cytosol[121,122]. Puf1p is one protein in yeast that is known to bind to 3'UTRs and in combination with other RNA-binding proteins with similar RNA binding motifs can control mRNA turnover and stability[123]. Recoding or designing 3'UTRs is one potential way for synthetic biology to alter the dynamics of a gene circuit in yeast, by controlling the speed of turnover of expressed mRNAs.

The other main role of the 3'UTR in translation is to enable rapid reinitation following a first round of protein synthesis. Motif sequences typically found within an mRNA 3'UTR act to aid connections between the 3' and 5' ends of the mRNA and cause mRNA looping, which is also known as pseudo-circularisation[124,125]. The 3'UTR sequences promote the binding of poly(A) binding proteins (PABPs) to the poly(A) tail and these interact with the eIF4F complex as shown in **Figure 1.27**. The so-called gene loops that form are proposed to enhance transcriptional directionality[126] and lead to quick reinitiation to drive strong gene expression.

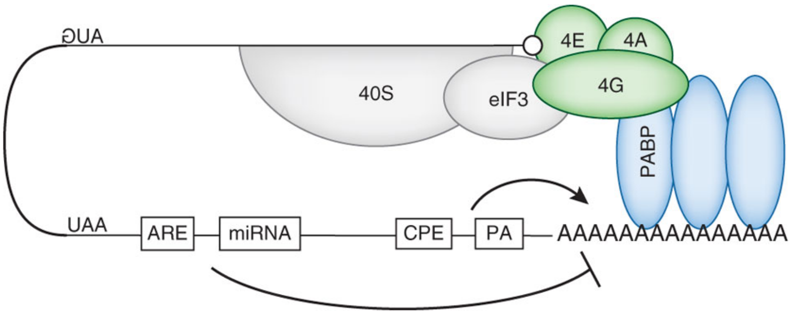


**Figure 1.27:** Sequence features and proteins involved in yeast mRNA looping during translation initiation[127].

### 1.5.4 mRNA surveillance and decay

The final feature associated with mRNAs and their translation which can contribute to the overall level of gene expression from a gene is their decay. As the last section mentioned, the binding of specific proteins such as the PUF proteins to the 3'UTR of an mRNA can be used to control the degradation rate of the mRNA and thus its turnover[128,129]. Altering the half-life of an mRNA can be used to tune the resultant gene expression strength and has been a strategy employed in yeast synthetic biology[63,64].

As well as programmed 3'UTR-directed mRNA degradation, mRNA decay can also be directed by the process of mRNA surveillance[130]. In this process, the ribosome is continually being monitored by a variety of factors during its normal roles of ribosome scanning and translation. If it is not efficiently performing these roles, for example if RNA or polypeptide roadblocks prevent scanning or translation elongation, then factors begin to trigger RNA degradation in order to rescue the ribosome and remove the mRNA causing aberrant translation[131]. Decay of mRNAs due to the surveillance mechanism can occur by the three mechanisms illustrated in **Figure 1.28** on the next page.

The first mechanism is called Nonsense-Mediated Decay and occurs when a premature stop codon is encountered in the mRNA. In other words, this mechanism determines if the 3'UTR on an mRNA molecule is exceptionally large. The ribosome then disassociates in a process where it is replaced on the mRNA by factors including Upf1 that promote the formation of a complex that then removes the mRNA 5' cap. This then leads to the mRNA being degraded by nucleases that attack the RNA from the 5' end. The second mechanism is called Non-Stop Decay and

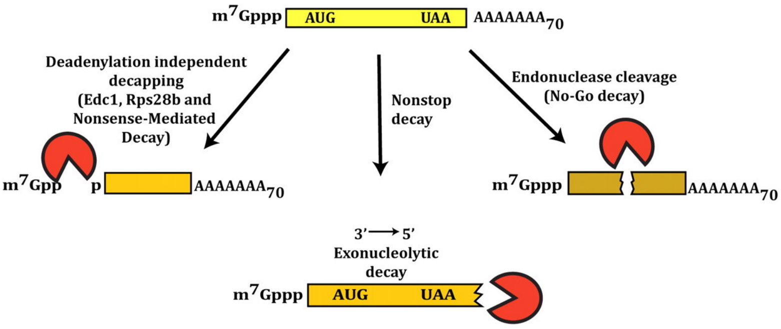

m7Gppp
AUG
UAA
AAAAAAA70
Deadenylation independent decapping (Edc1, Rps28b and Nonsense-Mediated Decay)
Nonstop decay
Endonuclease cleavage (No-Go decay)
m7Gpp
p
AAAAAAA70
m7Gppp
AAAAAAA70
3'⟶5'
Exonucleolytic decay
m7Gppp
AUG
UAA

**Figure 1.28:** The three main mechanisms of mRNA decay used in the RNA surveillance processes. Figure adapted from Parker[130].

occurs when the ribosome reaches the end of the mRNA without translation being terminated (e.g. when the stop codon is skipped or has mutated). As the ribosome reaches the poly(A) tail and is still translating it recruits Ski proteins to the mRNA that act to degrade the RNA from the 3' end via 3' to 5' exonucleases. The final mechanism, called No-Go Decay, occurs when a ribosome stalls during translation of the mRNA. To rescue and recycle the ribosome factors Hbs1 and Dom34 enter the ribosome at the site where the next tRNA would normally bind and help release the nascent peptide from the ribosome[132]. Their presence leads to cutting of the mRNA at the stall site, although the two proteins themselves do not appear to catalyse this cleavage reaction[133]. The ribosome and tRNAs can then be released from the cut mRNA and recycled. The rest of the split mRNA is then digested by exonucleases from the cleavage point.

### 1.5.5 Noise in gene expression

The previous sections all discussed the events during gene expression in yeast and how the efficiency of these are defined by the various sequence features encoded into the gene. Typically when describing how sequence determines gene expression, the focus is on the resulting protein expression strength. However, another measure of gene expression - its noise - is also important to consider, especially for attempts in synthetic biology and systems biology to use mathematical approaches to predict and model device, circuit and systems behaviour.

Noise is routinely seen in experimental data as variation in the measured expression of genetic devices and results from both intrinsic and extrinsic properties. The way protein expression itself is measured (e.g. the detection of a fluorescent protein by a flow cytometer) can contribute to the noise the data, as can the diversity in the states of the thousands of different yeast cells in the population (for example, in terms of them being at different points in their cell cycles and different sizes and ages). Ultimately however, noise is seen even at the single-cell level. This is because most of the processes in gene expression are inherently stochastic rather than deterministic.

Yeast has proved to be an exemplar organism for the study of stochasticity in gene expression and so our knowledge of what features of a gene sequence confer gene expression noise is quite advanced. Using synthetic promoter libraries similar to those described before Blake *et al.* reported an increase in the strength of expression noise when they increased gene expression in yeast, and were able to show that this effect was most pronounced when it was coupled to a stochastic transcriptional state[134]. This matched the work of others, who have used *S. cerevisiae* as a model organism, to determine that the majority of gene expression noise depends on the promoter used. Specifically, the noise comes from the type of promoter and not from the rate of expression it drives[135–138]. It seems that promoters that are slow to activate due to the need for chromatin remodelling generate much of the stochasticity in eukaryotic gene expression. It is thought that this is one of the key differences between eukaryotes and prokaryotes, where in prokaryotes the transition rates of promoters from OFF to ON states are comparatively fast[139].

As synthetic biologists attempt to improve constructed circuits with various regulatory mechanisms, it will be important to also consider how stochastic effects may play a role. Very recent impressive work in generating a robust oscillator system in *E. coli* has shown that altering the noise within a circuit can be part of a solution to make it more robust[140]. For work in *S. cerevisiae* it is likely that the mechanism via which promoters within a circuit are regulated will play the largest role in determining the noise in the circuit performance. This suggests a need for new synthetic ways to impose regulation where highly stochastic regulation can either be avoided or fine-tuned.

## 1.6 Enabling complexity in synthetic genetic circuits

Despite the substantial amount of knowledge now available on the molecular biology of the many processes in yeast gene expression, progress in yeast synthetic biology towards new gene regulatory circuits has not been as fast as would be expected[6,21]. This is even taking into account the considerable advances that have been made within synthetic biology, allowing the production of orthogonal regulators, promoter libraries and rapid modular DNA cloning kits. Six years ago the major challenges for progress of synthetic biology in all organisms were laid down. The so-called *Five Hard Truths for Synthetic Biology* were identified as poor characterisation of parts, unpredictability of circuits, unmanageable complexity, non-orthogonality and stochasticity[141]. Many of these challenges are related. For example, better characterisation of parts will allow potential difficulties to be avoided, by allowing better modelling of the circuit or system and therefore better predictability. It will also allow orthogonality issues to be prevented, by selectively choosing parts that are compatible. Admittedly, this may be a somewhat simplistic view, since it is often difficult to predict every parameter that requires characterisation and how they may change when combined in different contexts.

It is arguable that since 2010, orthogonality has been solved for synthetic biology, due to our new ability to program modular DNA-binding of repressors and transcription factors through dCas9/CRISPR and TAL-Effector proteins. Characterisation has become more extensive too, at least in *E. coli* synthetic biology, and this has allowed complex circuits to be produced and afford a much greater understanding of how device and circuit behaviour can be adjusted through various mechanisms.

Complexity and stochasticity are arguably the biggest and most fundamental challenges to synthetic biology. As shown in **Figure 1.29** on the following page, even in the most basic regulatory motif there are many factors that determine the precise output of a system, its performance

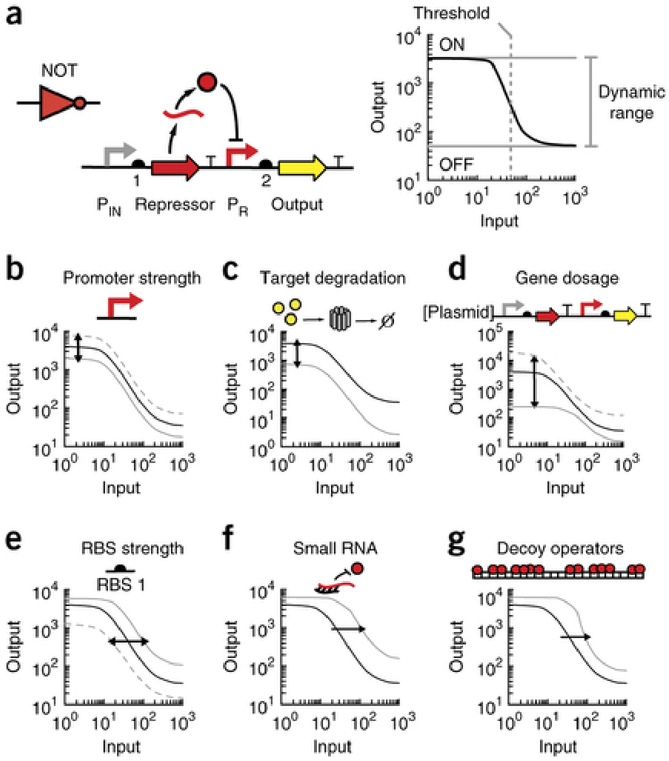


**Figure 1.29:** Different methods available in *E. coli* to adjust the performance of one of the most basic synthetic gene regulatory motifs: the inverter. Figure adapted from Brophy and Voight[15].

and dynamic range. This exemplar figure actually only shows a subset of the relevant factors, ignoring the degradation rates of the repressor protein, protein folding rates, terminator strength, codon optimisation and many more features. Given the fact that a typical circuit will consist of multiple regulatory motifs and not just a simple inverter it is not difficult to see that the degrees of freedom in the system quickly become very large and the corresponding rise in complexity can become daunting, especially for attempts to predict performance mathematically.

Modelling frequently does not offer a solution, because in all but a few cases the parameters for relevant cellular processes (degradation rates of mRNA and proteins, folding speeds of proteins, exact dose response curves for the repressor, precise concentrations of the relevant molecular species, etc.) are not known with enough precision, if known at all. Finding ways to limit the total number of parts in a system can therefore be an attractive strategy to reduce the complexity in a design while still making it achieve the required performance.

The burden of gene expression is also a consideration that makes it important to consider ways to limit the parts of a synthetic biology system *in vivo*. By adding new DNA constructs to cells in order to direct the expression of multiple proteins, the cells themselves are given extra work compared to wild-type cells and therefore have reduced fitness[142]. Importantly, it seems to be the case in most organisms that the major cost of gene expression is the process of translating mRNAs into functional mature proteins. This requires significantly more resources than the step of transcription and also occurs at a much slower time-scale. Therefore, the standard assumed approach in synthetic biology of increasing circuit complexity by simply connecting more (orthogonal) regulator proteins to different promoters is not ideal. As the circuit increases in size it can quickly become problematic for the cell due to the loading effects, the limited number of orthogonal regulators, and the time that it takes to process each step of gene expression within the circuit[143]. Systems built with more interacting protein components can also suffer from retroactivity (described in **subsection 1.1.4** on page 16) and require insulation themselves[13].

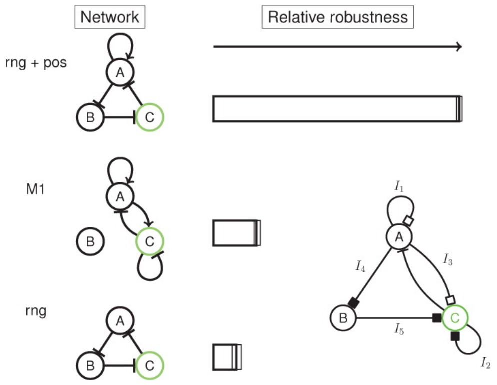


**Figure 1.30:** The predicted robustness emerging from 2, 3 and 4 node networks that define oscillator systems. Figure adapted from Woods et al.[144].

In reality, there are often many ways to design a circuit of interacting regulators so that it performs a desired task, and different designs can result in less complexity and requiring less protein components. For example, when designing a circuit that performs as an oscillator, it is possible that this can be achieved by either using just two or three regulators as shown in **Figure 1.30**. The least robust version of the network uses three repressors in series: the well-known repressilator motif. This can be improved when the regulators are able to be activated and repressed simultaneously by one another and even when only two are used[144]. Thus, this demonstrates that by taking into account more complex regulation at the parts level (e.g. promoters that can be activated or repressed by different transcription factors) it is possible to improve circuit performance while simultaneously decreasing the number of components involved. Considering the many different ways that gene expression is modulated naturally in yeast and the countless opportunities for adding different modes of regulation, it is surprising that we are still stuck with genetic circuits that simply follow the standard approach of a single protein transcription factor regulating its cognate promoter which then expresses another transcription factor. To scale complexity in yeast synthetic biology, we need to expand upon the mechanisms that are already available to allow more diverse, robust and complex functionality.

## 1.7 Aims of this thesis

The aim of this thesis is to research and develop parts and tools for yeast synthetic biology that will allow richer functionality and increased robustness of synthetic circuits while limiting their complexity in terms of the number of genes that need to be expressed to achieve the required functionality. Ideally, this will reduce the metabolic load on the host and increase predictability. These newly developed methods for regulation should be designed to be compatible with the widely adopted transcription/translation based circuits in use in much of synthetic biology. For this reason phosphorylation-based computation via signalling pathways has been avoided, despite this offering some fundamental advantages over traditional regulation[145].

**In the first results chapter** we look at ways to tune gene expression strength and note that while many successful attempts have been made to fine-tune promoter strength in yeast, the predictability of promoter strength remains poor. We also note that there is no yeast equivalent to the RBS Calculator existing for prokaryotes, that allows predictable tuning of expression and is extensively used in prokaryote synthetic biology. Currently, the only way to tune expression strengths easily in yeast, is to make use of pre-characterised promoter libraries or promoters that can be regulated externally through the addition of small molecules. These approaches may be incompatible with specific regulated promoters that are required for a desired functionality in a circuit, and we therefore set out to create a generalised method to predictably adjust expression strengths from any promoter in yeast. This approach is based on the predictability of base-pairing and secondary structure in mRNA and will serve as a fundamental tool for yeast synthetic biology.

**In the second results chapter**, we note that the fundamental functionality of a true synthetic bistable switch has not been created in *S. cerevisiae* yeast. We set out to develop a new regulatory mechanism that would aid in the development of bistable switches without leakiness, and that would also be useful in the future for other genetic circuits. Taking inspiration from natural regulation seen in yeast and other organisms we explore doing this through transcriptional interference. In this method, transcriptional interference reinforces mutually repressive regulation already present in a bistable switch, which reduces the need for additional genes to reduce leaky expression from repressed promoters. This design should lower the complexity of the genetic circuit while increasing robustness.

**Finally, in the third results chapter**, we worked on a generalised method for reducing complexity in genetic circuits that could also be efficiently applied to the bistable switch design. Using orthogonal, modular transcription factors, we set out to rationally design and build transcriptional regulators that can act as both activators and repressors, depending on the position of their binding site on the promoter. This theoretically eliminates the need for inverter motifs, when a positive signal needs to be converted to a negative signal. This elimination reduces the complexity of the design of a circuit and thus could also reduce the subsequent need for tuning and troubleshooting, speeding up the development process of complex gene regulatory circuits.

# 2. Materials & Methods

## 2.1 DNA assembly with conventional restriction enzyme cloning

When this work was started, all DNA cloning was primarily done with conventional restriction enzyme cloning. As new approaches and technologies emerged during the project lifetime, this was reflected in the approach taken for construct assembly. Here, we first describe the workflow for the standard conventional DNA cloning used at the start of the project.

The workflow was centred around a collection of modular yeast plasmids that we refer to as the pRS system[146]. This collection contains a variety of shuttle vectors that contain elements necessary for propagation in both *E. coli* and *S. cerevisiae*. For cloning in *E. coli* the vectors contain a marker for ampicillin selection and an origin of replication for high copy propagation in addition to a multiple cloning site for flexible incorporation of foreign DNA. For propagation in yeast, the system contains a choice of selectable markers and the option of non-integrative propagation at low copy number using the CEN-ARS autonomously replicative element. Integration into the chromosome is also made possible by using the selective markers, which double as flanking sequences for homologous recombination into defined sites within the genome. This is made possible by the accompanying yeast strains used within the project that have been engineered to be auxotrophic for certain nutrients, while retaining some homology to the plasmid-provided genes that rescue this phenotype. The specifics of this mechanism are further detailed in the section about yeast cloning (**subsection 2.1.2** on page 53).

### 2.1.1 Cloning in *E. coli*

All plasmid construct assembly was done in *E. coli*, before transformation of the finished plasmids into yeast. Plasmids built to contain the CEN-ARS replicon propagate episomally in yeast, while plasmids built without any yeast replicon parts are digested into linear form for integration into the yeast genome by homologous recombination.

#### Strain, media and culturing

The *E. coli* strain that was used for all plasmid construction steps was the DH10$\beta$ strain from Invitrogen/Thermo Fisher[147]. This strain was cultured in liquid LB-broth Miller (Cat No 1.10285.0500, Merck Millipore) or grown on solid LB-agar plates (Cat No 1.10283.0500, Merck Millipore), both prepared according to the manufacturer's specifications.
When applicable, antibiotics were added to the medium. These were used at the following

concentrations: Ampicillin: 100 $\mu$g/ml, Kanamycin: 50 $\mu$g/ml, Spectinomycin: 50 $\mu$g/ml, Chloramphenicol: 33 $\mu$g/ml, Tetracycline: 15 $\mu$g/ml.

Strains were typically grown overnight from picked colonies in 5 ml of media in 14 ml round bottom snap-cap tubes (Cat No 734-0446, VWR) in a rotary shaking incubator (MaxQ 6000, Thermo Scientific) at 37 °C at 250 rpm. This then provided the necessary cell volume for DNA plasmid preparation.

### DNA manipulation

DNA purification was performed using the Qiagen MiniPrep, PCR-Purification and Gel Extraction kits following the manufacturer's protocols. All DNA samples from these kits were eluted into MilliQ water with the minimal elution volumes (50, 30 and 30 $\mu$l respectively) to maximize DNA concentration. DNA concentrations were assessed using a Nanodrop1000 spectrophotometer (Thermo Scientific).

All restriction digests and ligations were done using enzymes from New England Biolabs, according to the manufacturer's specifications. Incubation time was typically one hour for both types of reactions. Blunt-end ligations were incubated for 2 hours. For standard ligations, T4 DNA ligase was used and the reactions were performed at room temperature. The majority of the ligations were single inserts into a vector and for these reactions an insert:vector ratio of 3:1 was used. Total DNA concentration for ligations was kept between 1 and 10 $\mu$g/ml. Reactions were always heat inactivated for 20 minutes at 65 °C (or 80 °C if required), unless restriction products were gel extracted.

### Electroporation

*E. coli* electrocompetent cells were transformed using a BioRad Micropulser electroporator according to the following protocol. Take electrocomptent cells from -80 °C and thaw on ice. Subsequently add 2 $\mu$l DNA to the cell suspension, mix and incubate on ice for 5 minutes. Transfer to a chilled BioRad electroporation cuvette with 1 mm gap width and tap to spread the suspension evenly in the cuvette and eliminate any bubbles. Set the electroporator to "Bacteria" and "Ec1". Wipe the cuvette and place in chamber slide. Apply electric pulse and directly add 0.5 ml of LB medium. Record the time constant. A time constant between 2.5 and 5 ms indicates a technically successful transformation. Then incubate the mixture in a microcentrifuge tube for 30 minutes at 37 °C for ampicillin selection and for 1 hour for any other antibiotic. Following this, spread 10 to 100% of the cell suspension on appropriate selective plates.

Electrocompetent cells were prepared using the following protocol which was supplied by Dr Benjamin Blount (Imperial College London):

#### Day One

1. Inoculate a 5 ml LB culture with a single colony of the *E. coli* strain and incubate O/N at 37 °C.
2. Incubate a 2 litre conical flask containing 500 ml of LB at 37 °C O/N.
3. Store 500 ml sterile dd$H_2O$ at 4 °C O/N.

**Table 2.1:** Different variants of pRS vectors available for the introduction of foreign DNA into yeast.

| Name | Auxotrophic marker | Propagation |
|---|---|---|
| pRS403 | HIS3 | genomic integration |
| pRS404 | TRP1 | genomic integration |
| pRS405 | LEU2 | genomic integration |
| pRS406 | URA3 | genomic integration |
| pRS413 | HIS3 | extrachromosomal (CEN6-ARS4) |
| pRS414 | TRP1 | extrachromosomal (CEN6-ARS4) |
| pRS415 | LEU2 | extrachromosomal (CEN6-ARS4) |
| pRS416 | URA3 | extrachromosomal (CEN6-ARS4) |

**Day Two**

1. Use the O/N culture to inoculate the 37 °C LB broth 1:100 and incubate shaking at 37 °C.
2. Pre-chill sterile 20% (v/v) glycerol and the sterile $ddH_2O$ using ice. Label microtubes and store in the -80 °C freezer. Pre-chill a centrifuge rotor to 4 °C.
3. When the OD600 of the culture reaches approximately 0.5, transfer to 50 ml Falcon tubes (ensure that there is no more than 40 ml/tube) and chill on ice for 30 minutes. In order to maintain the efficiency of the cells, it is vital that once the mid-exponential culture is chilled, the cells remain at low temperature.
4. Centrifuge the tubes in the rotor pre-chilled to 4 °C at 4000 rpm for 15 minutes.
5. Discard the supernatant and, on ice, re-suspend the cells in the equivalent volume of pre-chilled water.
6. Centrifuge as before.
7. Discard the supernatant, on ice re-suspend cells in pre-chilled 20% glycerol (volume is not important but ideally just enough to re-suspend the cells e.g. 2ml/tube) and pool all of the cells into one 50 ml Falcon tube.
8. Centrifuge as before.
9. Discard the supernatant and, on ice, re-suspend the cells in approximately 3 ml pre-chilled 20% glycerol.
10. Transfer the cells into the pre-chilled microtubes in 50 $\mu$l aliquots and store immediately at -80 °C.

### 2.1.2 Cloning in *S. cerevisiae*

The pRS vector system allows for flexible transformation of yeast[146]. It is a modular system with markers for rescuing various auxotrophies, allowing selection of yeast that have acquired foreign DNA. With multiple markers available, multiple plasmids can be incorporated into yeast, allowing for further flexibility and complexity. **Table 2.1** shows the available variants of the pRS plasmid series.

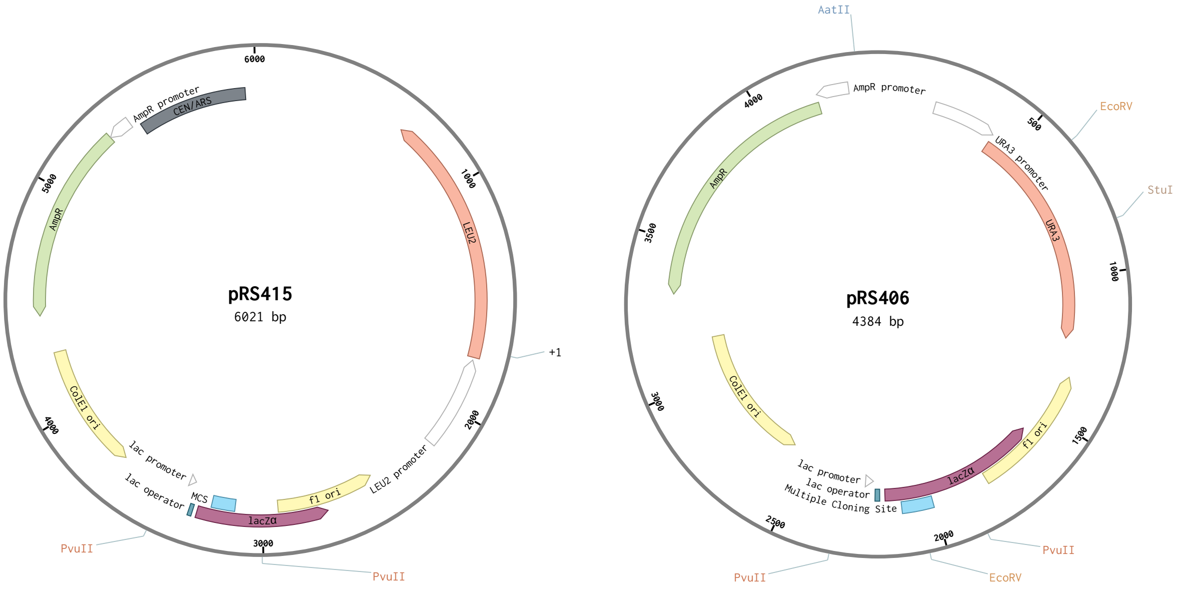


**Figure 2.1:** Examples of pRS plasmids. Important restriction sites are highlighted. PvuII was used to introduce the cargo DNA into the plasmid. '+1' indicates the BstEII restriction site that was used to linearise the plasmid for genomic integration of pRS405. pRS415 (shown) does not require linearisation, since the pRS41x series contains a CEN6-ARS4 sequence for extrachromosomal propagation. In pRS406, the AatII restriction site is shown that was used for piggyback integration into a pRS plasmid already present on the genome. Additionally, the EcoRV and StuI sites are shown that were used for genomic integration into the URA3 locus (depending on identical sites being present elsewhere).

Schematic representations of two examples of pRS plasmids are shown in **Figure 2.1**. This figure shows that apart from the CEN6-ARS4 and marker sequences, the backbone of the vectors is identical between the two versions. This allows the cargo, that was always inserted at the PvuII sites, to be exchanged between different versions of pRS plasmids, without concerns about context effects. This also allows the plasmids to be integrated into a genomic locus that has previously had a pRS vector integrated into it (so-called piggyback integration). This can be advantageous when the restriction site that would normally be used to linearise the plasmid for integration, is also present in the cargo DNA.

As mentioned previously, homologous recombination is the central process used for introduction of foreign DNA into the yeast genome. **Figure 2.2** on the following page shows a diagram of how this process takes place. Typically, the auxotrophic marker serves both for selection and as a homology flank to direct homologous recombination. It is directed to the homologous locus on the genome, where an inactivated version of the same gene resides. The specific way the plasmid is linearised, results in a process called 'ends-in homologous recombination'. After the single-crossover integration event, this results in two hybrid copies of the auxotrophic marker existing on the genome. Both contain part of the sequence originally present on the genome and part of the sequence that was originally part of the plasmid. Only one will be a functional copy, since the other version will still contain the disruption that was present on the genome.

The 'ends-in' type of homologous recombination results in the reconstitution of the original sequence present on the genome (i.e. the disrupted auxotrophic marker). This leads to an

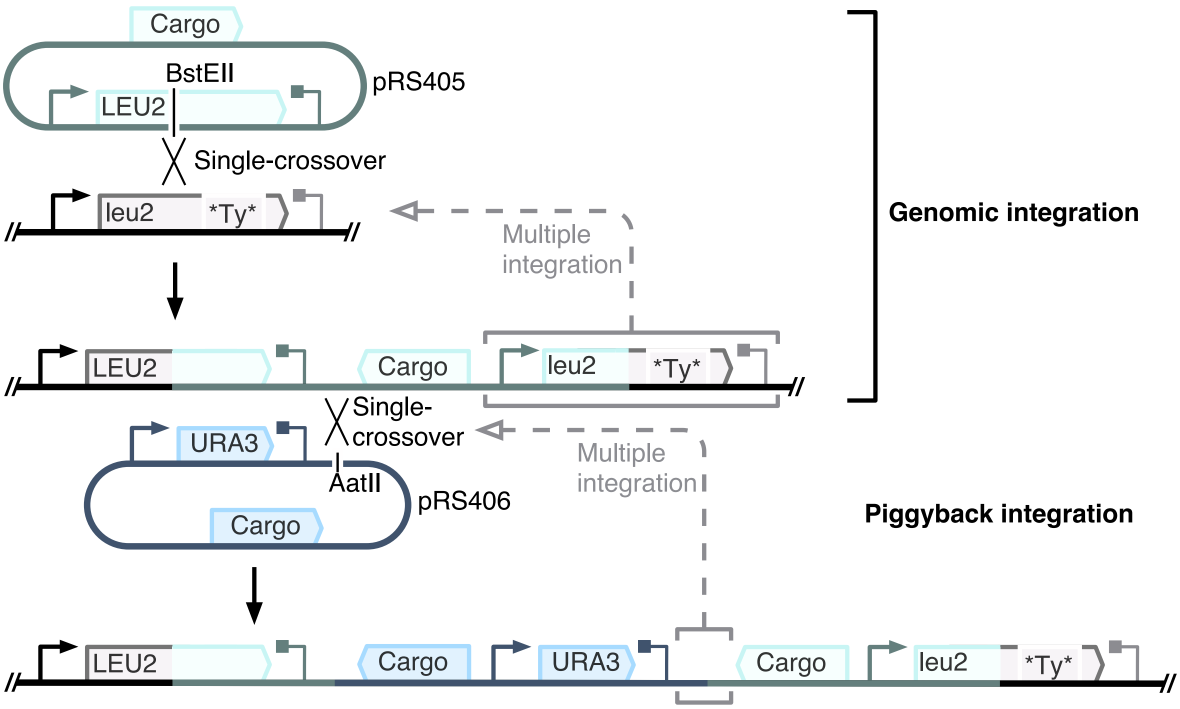


**Figure 2.2:** Integration of pRS based vectors into the yeast genome. Top section shows integration of pRS405 into the inactivated LEU2 locus on the genome. *Ty* denotes a transposon element that was inserted into the original gene to inactivate it. pRS plasmids are linearised using a unique restriction enzyme in order to facilitate a single-crossover homologous recombination event that leads to the incorporation of the plasmid. The bottom section shows how a second pRS-based plasmid can be incorporated at the same locus through homologous recombination between the identical backbone sequences. This is the so-called piggyback method.

increased incidence of multiple integrations, because the linearised plasmid will typically be present in the cell at multiple copies. After integration of the first copy, the reconstituted (hybrid) auxotrophic marker sequence can be targeted for homologous recombination by the second copy (and so on). This is in contrast to an analogous process called ’ends-out homologous recombination’, where the original integration site is not reconstituted. This type of homologous recombination is discussed in **Figure 2.8** on page 67.

Successful integration is heavily dependent on a matching homology between the insert and the genome. This makes the pRS particularly strain specific. A suitable host strain is auxotrophic for the HIS3, LEU2, TRP1 and URA3 markers, but retains sufficient homology to these genes in order for homologous recombination to be successful. This can cause complications when a significant part of the gene has been deleted to inactivate it. If the corresponding gene on the pRS plasmid is linearised in this region, there will be no homology on the genome to facilitate integration.

This is particularly relevant for the HIS3 and TRP1 markers. The trp1-$\Delta$63 present in the YPH500 strain that is used for pRS plasmid integration is a near complete deletion of the TRP1 gene. Only the SnaBI and BspEI restriction sites linearise the vector at a location where sufficient (but still limited) homology remains for succesfull integration. In the HIS3 marker the deletion is such that no suitable unique restriction sites remain in the pRS403 plasmid. For this reason we did not use HIS3 based plasmids for integration into the genome. In URA3 any restriction site is allowed since inactivation of the host gene was achieved by insertion of a Ty element only. We used the EcoRV and NdeI restriction sites most extensively. In LEU2 we used the BstEII restriction site, since BbsI and AflII fall within a deleted region.

From the above, it follows that the choice of linearisation sites can be quite constrained. This must be taken into account during the design of the cargo sequences. The intended linearisation site must be unique and cannot be contained in the cargo sequence. Non-unique linearisation sites were avoided by codon optimisation, cloning site choice or selection of a different auxotrophic marker.

### Strains and media

The parental strain used for yeast genomic integration was *S. cerevisiae* YPH500 (MAT$\alpha$ ura3-52 lys2-801$^{amber}$ ade2-101$^{ochre}$ trp1-$\Delta$63 his3-$\Delta$200 leu2-$\Delta$1). All selection was done using auxotrophic markers in defined dropout media, prepared according to instructions in the book 'Methods in Yeast genetics'[148]. To increase the shelf-life of solid media in petri dishes, these were supplemented with Kanamycin at a rate of 50 $\mu$g/ml. For non-selective outgrowth we used YEP media (1% w/v Yeast Extract, 2% w/v Peptone from soy in MQ).

For normal growth the rich and selective media contained 2% dextrose (glucose). For induction of the GAL promoter 2% galactose was used instead. All media were supplemented with 40 mg/l Adenine sulphate, to prevent deleterious effects of the ade2-101$^{ochre}$ genotype. Deleterious effects of this genotype were found to be especially severe when using unsupplemented YEP media prepared with peptone made from casein. Soy based peptone did not show such severe effects. For this reason we primarily used soy based peptone in addition to the supplementation with adenine.

### Transformation

Yeast transformations were done using competent cells prepared in advance and stored at -80 °C. This significantly reduces the time and planning usually needed for yeast transformations. The protocol is directly derived from the Lithium Acetate/PEG mediated transformation procedure and is documented elsewhere by Gietz and Schiestl[149]. This protocol typically gave between 250 and 3500 transformants per microgram of DNA. Between 200 and 1000 ng of linearised plasmid DNA was used per transformation.

### Knock-outs

Here we describe the method used to generate the eIF2A knock-out strain used in **subsection 4.3.3** on page 137. A visual representation of the process is shown in **Figure 2.3** on the next page. The first step in the process is the generation of 3 PCR products using the primers that are listed in **Table 2.2** on the following page. Two of these products are the upstream and downstream homology flanks. These are 1kb regions amplified from the genome of the target strain (YPH500) using colony PCR. They are necessary to facilitate efficient integration of the third fragment into the genome. The third fragment is a transcription unit for the dominant KanMX selection marker.

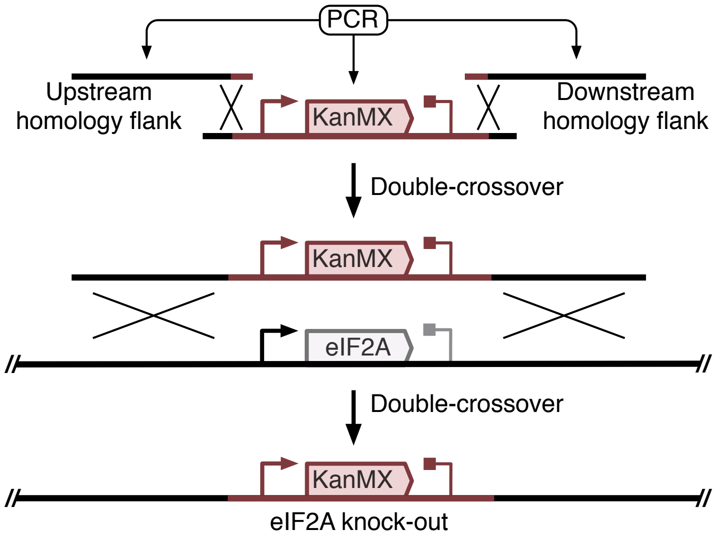


**Figure 2.3:** Process of generating the eIF2A knock-out in *S. cerevisiae*. 3 DNA fragments are generated using PCR. Each of the fragments contains a small homology region to one or both of the other fragments. After transformation of these fragments, they are assembled into a larger fragment through a double-crossover homologous recombination event. This large fragment, which now contains 1kb homology flanks to the targeted (eIF2A) region on the genome is then recombined into the genome in a second double-crossover homologous recombination event. The result is a strain containing the KanMX dominant marker at the location on the genome that was previously occupied by eIF2A.

**Table 2.2:** Primers for the generation of an eIF2A knock-out strain. Capital letters indicate the annealing part of the primer.

| Name | Sequence | Description | Direction |
|---|---|---|---|
| TW198 | AATGGTCTTCCGGTATGCA | Upstream flank | Fw |
| TW199 | gcaagctaaacagatctatattaccctGCGGTCGGGTAATAATATC | Upstream flank | Rev |
| TW200 | gaattcatcgatgatatcagatccaAACTAGAAGAAACTGATGTATC | Downstream flank | Fw |
| TW201 | CCCAATACACGACAAAATAC | Downstream flank | Rev |
| TW202 | tgtacgatattattacccgaccgcAGGGTAATATAGATCTGTTTAG | KanMX marker | Fw |
| TW203 | gtatggatacatcagtttcttctagttTGGATCTGATATCATCGATG | KanMX marker | Rev |

400 $\mu$g of each of the 3 purified fragments is transformed into yeast. In order for a successful knock-out to occur, these fragments must be homologously recombined into one larger fragment. The primers used for amplification each contain a 25bp tail that is homologous to the adjacent fragment. This combines into a 50bp overlap between each of the fragments, which is sufficient to efficiently facilitate homologous recombination into a larger fragment of approximately 3kb.

Recombination into the genome happens at a much lower rate than in short, linear fragments that are supplied in high concentration, making it necessary to add the 1kb flanks to the KanMX gene. With the flanks added, another double-crossover homologous recombination event takes place, replacing the target gene (eIF2A) with the dominant selection marker. The KanMX gene confers resistance to Gentamycin (G418), which is supplied at 300 $\mu$g/ml in YEPD media. Colonies are picked and screened for successful deletion of the eIF2A gene using colony PCR.

### 2.1.3 Flow Cytometry

Protein expression was primarily determined by flow cytometry. The transition between traditional restriction enzyme cloning and YTK-based construct assembly coincided with a change in flow cytometry equipment. Unless otherwise stated, constructs assembled with traditional methods were measured with previous generation equipment (described below) and YTK assemblies with next generation equipment described in the next section.

#### Culturing

After transformation, three colonies per transformation were picked into 2ml YEP-Dextrose medium for fluorescence determination. Cultures were grown to saturation, typically overnight. The majority of the measured constructs contained galactose inducible promoters. Induction was done in YEP-Galactose media for a minimum of 12 hours, typically overnight. 2-5ml cultures were inoculated with a 1000-fold dilution of the saturated YEP-D culture in 14ml round bottom snap-cap tubes (Cat No 734-0446, VWR) in a rotary shaking incubator (MaxQ 6000, Thermo Scientific) at 30 °C at 250 rpm. The day measurements were taken, the cultures were diluted and grown for a minimum of 4 hours to ensure logarithmic growth. Directly prior to data collection, cultures were diluted 10-100 fold in MQ, depending on culture density. Measurements were performed using 5ml round bottom snap cap tubes (VWR, cat no 734-0443).

When applicable, IPTG and ATc were added to the media at saturating concentrations of 10mM and 250ng/ml, respectively. IPTG was kept as a 1M (100x) stock solution in water. ATc was kept as a 100$\mu$g/ml (250x) stock solution in 50% ethanol.

#### Equipment

Flow cytometry assays for single-cell GFP and RFP expression were taken with a modified Becton Dickinson FACScan flow cytometer equipped with both a blue laser (488 nm) for yeast enhanced Green Fluorescent Protein (yeGFP) excitation and a green laser (561 nm) for mCherry excitation. Green fluorescence was detected with a 530 nm band pass filter (FL1) with gain 890. Red fluorescence was detected with a 610 nm filter (FL5) with gain 850. Data was collected using CellQuest Pro (v5.1.1; Becton Dickinson Co.)

#### Data analysis

The flow cytometer that was used for these experiments did not have a reliable 96-well plate reader. For this reason, we could not obtain fluorescence readings for our samples by averaging a large number of measurements as this would be prohibitively time consuming. Fluorescence readings therefore had to be obtained from single measurements consisting of 10,000 events. Fluorescence is not only correlated with the strength of expression in a cell, it is also correlated with cell size. The observed standard deviation of the fluorescence is therefore inflated, since the cell size in a logarithmically growing culture is a wide distribution. Using a workflow outlined in **Figure 2.4** on the next page, we attempted to correct for the variability introduced by cell size.

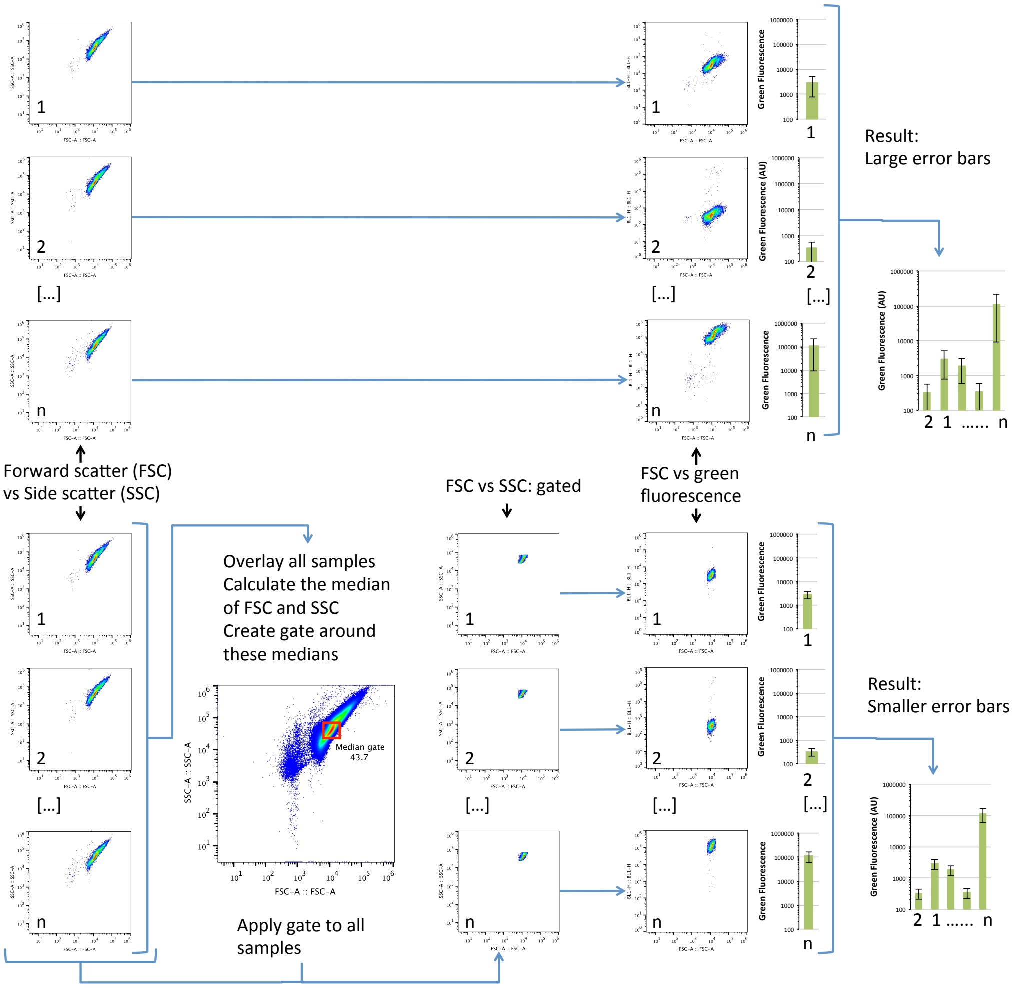


**Figure 2.4:** Workflow for flow cytometry data analysis. In our approach, the populations are gated around the median of forward and side scatter (FSC and SSC). This way, variation of cell size is eliminated as a factor in the obtained variance in observed fluorescence levels of each of the samples. This compares favourably to an approach where standard deviation is calculated directly on the full population, as variation in cell size contributes significantly to the observed variance.

Forward Scatter (FSC) and Side Scatter (SSC) are flow cytometry measurements that are collected as a standard procedure. These measurements are a proxy for cell size and complexity. By creating a tight gate around a subset of events with closely matched FSC and SSC, we selected a subpopulation of cells with similar cell size. By calculating the deviation of fluorescence in this subset, we get a more accurate measure for the variability in strength of expression in the cells, rather than the variability in cell size.

To create a fair comparison between different samples, the same FSC/SSC gate was applied to all samples within a particular experiment. To determine the optimal boundaries of this gate, we overlaid all samples within an experiment and calculated the median of the FSC of all events and the median of SSC of all events. These values were then used as the midpoints for the boundaries of the gate. This ensured that the largest proportion of closely matching cells in terms of cell size were selected for further analysis in all samples.

Of this gated population we determined the median fluorescence and Median Absolute Deviation (MAD). We did not use the more commonly used mean and standard deviation, because these measures are designed for linear datasets. Since flow cytometry data is logarithmic, a small number of outliers can disproportionally affect the mean and StDev. Outliers are common in flow cytometry experiments, because a small number of cells can carry over to the next measurement in a series of measurements. The median is not affected by small numbers of outliers, and is therefore a preferred measure of the global magnitude of the expression in a particular sample. The MAD is to the StDev what the median is to the mean: it also gives a more accurate measure of variance in a sample with logarithmic values and small but non-negligible numbers of expected outliers[150]. To calculate the MAD, all absolute differences are calculated between the value of each sample and the median of the population. Then the median of the generated list is taken to arrive at the MAD. We have developed a custom Matlab script that calculates all the above parameters on a set of measurements.

When outcomes were close, or particularly critical, we calculated the statistical significance of the observed differences between certain measurements. To mitigate the effect of outliers, we used a 20% trimmed t-test[151], implemented in the yuenv2 function within the Rallfun-v33 statistics package for R[152]. Because flow cytometry generates high volumes of data and suffers from systematic errors, statistical significance is easily reached, even for two identical samples. For this reason, we only reported significant differences when the explanatory measure of effect size was found to be 0.5 or higher. This corresponds to a high effect size, or a Cohen's d higher than 0.8 in the regular t-test[153,154]

As we mentioned previously, integrations using the pRS system are particularly prone to multiple integration events. Multiple integration is a serious problem for the characterisation of biological circuits, because the resulting measurements do not accurately reflect the state of the circuit, since the reporter or any other component of the circuit may be present in a double or higher dose. This leads to serious inaccuracies in the obtained characterisation data and must be eliminated. To do this, we tested three colonies for each transformation. We observed their expression levels and in each case where one of the transformants diverged significantly from the other two, we tested all colonies for multiple integration using colony PCR. In the vast majority of cases we then found the divergent colony to have acquired multiple copies of the construct and was subsequently dropped from the analysis. From the remaining (closely matching) samples, one was selected for presentation in this thesis.

### 2.1.4 TALE assembly

The highly repetitive DNA sequences for TAL-effectors were assembled using a popular kit designed especially for this purpose[18]. The assembly process is highly similar to that of the Yeast ToolKit, which is described in the following section. The assembly was carried out according to the instructions given in the paper and its accompanying materials. The required plasmid sequences were acquired from AddGene. The binding specificities of the two TAL-effectors used in this work are given in **Table 2.3** on the following page. Because of their repetitive nature, TALE sequences were always cloned into the constructs last. Similarly, in a yeast strain with multiple constructs, the plasmid harbouring the TALE was transformed last. We recoded the TAL21 to increase its evolutionary stability and the sequence of this construct is shown in **Figure 2.5** on the next page.

**Table 2.3:** Recognition and RVD sequences of the TAL7 and TAL21 TAL-effectors.

| | | Full RVD repeats | | | | | | | | | | | | | | | | | | | half-repeat RVD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| TAL21 | recognised base | C | C | A | T | T | C | T | A | A | A | C | A | C | A | A | C | A | T | A | A |
| | RVD identity | HD | HD | NI | NG | NG | HD | NG | NI | NI | NI | HD | NI | HD | NI | NI | HD | NI | NG | NI | NI |
| TAL7 | recognised base | A | T | C | T | C | T | T | C | A | A | C | A | C | A | A | C | A | T | A | A |
| | RVD identity | NI | NG | HD | NG | HD | NG | NG | HD | NI | NI | HD | NI | HD | NI | NI | HD | NI | NG | NI | NI |

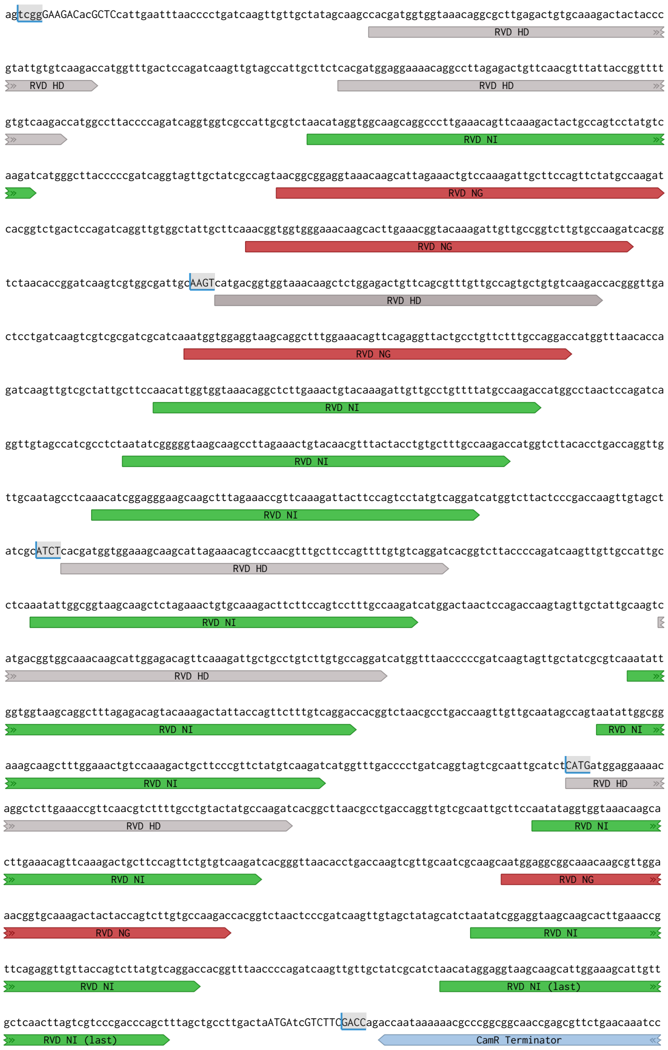


**Figure 2.5:** DNA sequence of the recoded TAL21.

## 2.2 DNA assembly with the Yeast ToolKit (YTK)

During the course of the project a new cloning system for yeast became available that dramatically increased the throughput and reduced the lead times in the assembly of DNA constructs. This method is referred to as the Yeast ToolKit and, like TALE assembly, relies on Golden Gate based methods of DNA assembly[8]. It relies on a large library of characterised parts that can be assembled in a hierarchical fashion.

**Figure 2.6** on the following page shows the hierarchy of assembly steps in the YTK. At the lowest level it contains part plasmids. Many part plasmids have been pre-defined and ship with the kit, but new parts can be created if required. Because the method relies on restriction enzymes for assembly, the method is not entirely sequence independent. The BsaI, BsmBI and NotI restriction sites are forbidden in part sequences and must be removed prior to incorporation of new parts into the a YTK part-level plasmid. New parts are first amplified by PCR to include both the BsaI and BsmBI sites at both extremities of the product. The enzymes are type IIs restriction enzymes, that cut outside their recognition site. In the first reaction, the BsmBI site is eliminated to yield a part-level plasmid with only BsaI restriction sites.

Parts are subsequently assembled into cassette-level plasmids. In this reaction, eight part-level plasmids donate their cargo DNA to yield a cassette which is typically a transcription unit. In this reaction the BsaI sites are eliminated. Two of the parts re-introduce a BsmBI site, flanking the transcription unit. These can subsequently be used in the next level of assembly. The parts are assembled in a pre-defined order, determined by the specific sequence contained in the scar-site next to the BsaI recognition sequence. Because of this standardised assembly method, the resulting cassettes can be fully defined in terms of the identities of the part-level plasmids used for construction. If the part-level plasmids are taken from the library, just a reference to the plasmid names suffices in defining the exact sequence of the cassette plasmid.

In the final level of assembly, the cassette-plasmids can be further assembled into multigene plasmids. This part of the assembly protocol is optional, since in many cases a single transcription unit will suffice. Up to 5 cassette plasmids can be incorporated into the multigene level in addition to one backbone plasmid which determines the location of integration in the yeast genome and the selection marker for transformation.

Interestingly, in this cloning system the selection marker for transformation and the location of integration are de-coupled, unlike in pRS-based systems. Location of integration is determined by two regions of homology that are included in the backbone of both the cassette-level and multigene-level plasmids. In between these homology regions are located the ORI and selection marker for propagation in *E. coli*. These are removed prior to integration in yeast through a restriction digest with NotI.

**Figure 2.7** on page 64 shows the library of characterised parts that are included in the kit. For flexibility, some parts in the 8-part cassette assembly are sub-divided into an **a** and a **b** part. This allows, for example, the ORF part (part number 3) to be split into a degradation tag domain (part 3a) and a coding sequence (part 3b). Correct assembly of all parts is dependent on the unique scar sequences adjacent to the restriction enzyme recognition site that is used in that particular round of assembly. This means that the method is not scar-less. However, certain scar sequences are intelligently chosen such that they minimise the impact on the final construct.

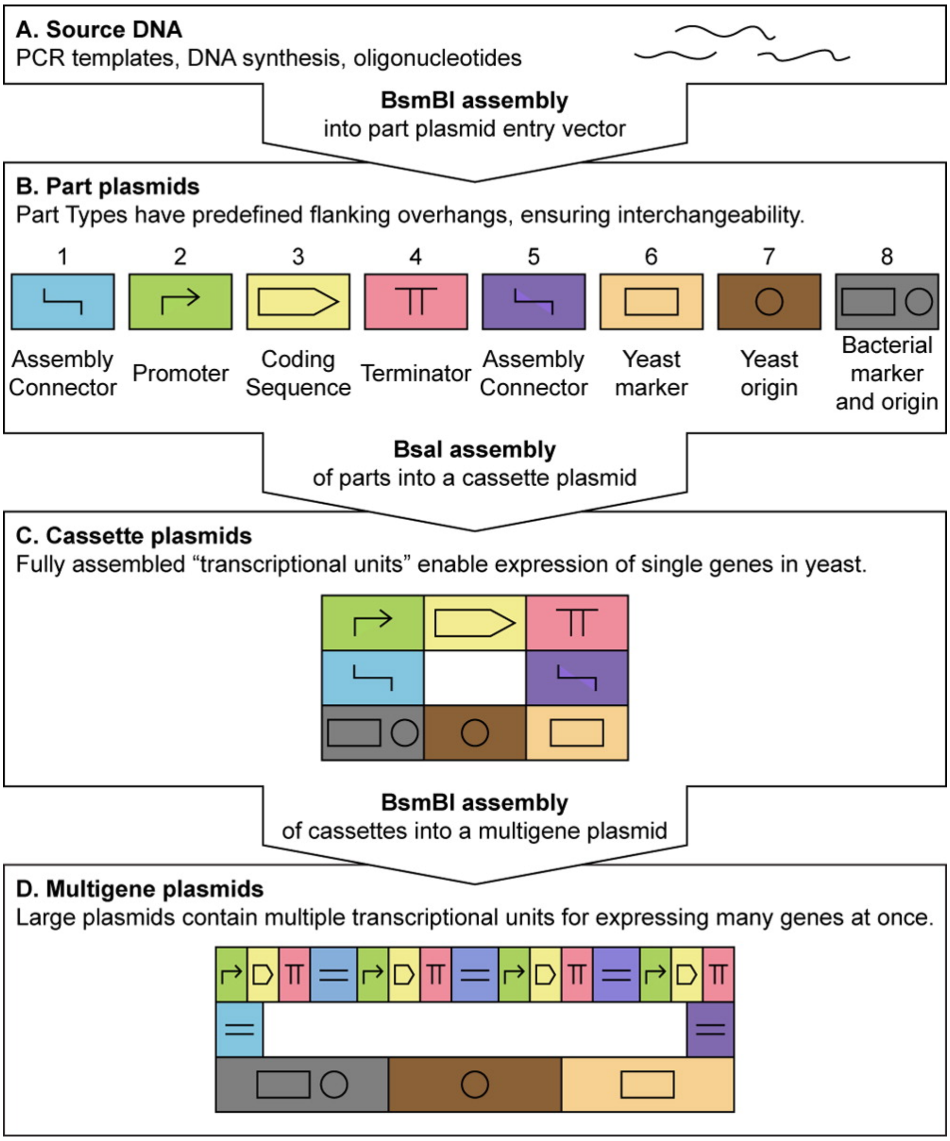

A. Source DNA
PCR templates, DNA synthesis, oligonucleotides
BsmBI assembly
into part plasmid entry vector
B. Part plasmids
Part Types have predefined flanking overhangs, ensuring interchangeability.
1
2
3
4
5
6
7
8
Assembly Connector
Promoter
Coding Sequence
Terminator
Assembly Connector
Yeast marker
Yeast origin
Bacterial marker and origin
BsaI assembly
of parts into a cassette plasmid
C. Cassette plasmids
Fully assembled "transcriptional units" enable expression of single genes in yeast.
BsmBI assembly
of cassettes into a multigene plasmid
D. Multigene plasmids
Large plasmids contain multiple transcriptional units for expressing many genes at once.

**Figure 2.6:** Hierarchy of assembly in the Yeast ToolKit assembly method. Figure from Lee *et al.* 2015[8].

For example, the scar that joins the promoter to the ORF includes the obligatory start codon sequence ATG. Likewise, the scar joining the ORF with the terminator includes a stop codon.

In **chapter 5** on page 158 we required a method for adding transcription factor binding sites upstream of promoters. To do this, we split the promoter part (part number 2) into two parts by introducing a unique scar site in the middle. The first fragment, part 2a, carries the TF binding sites, while the second fragment, part 2b, now supplies the promoter. The sequence of the 2a/2b scar is TTGA. Similarly, in **subsection 4.3.5** on page 147 we required a scar that included the 3' splice site YAG. This was done in order to split part 3 into a part containing the first exon plus the intron containing a gRNA and a part containing the second exon. These were called the 3alpha and 3beta parts and the junction between them was TTAG. In the same section, we required the first cassette in a multigene assembly to be incorporated in reversed orientation. This required incorporation of custom multigene assembly connector parts (part number 1 and 5).

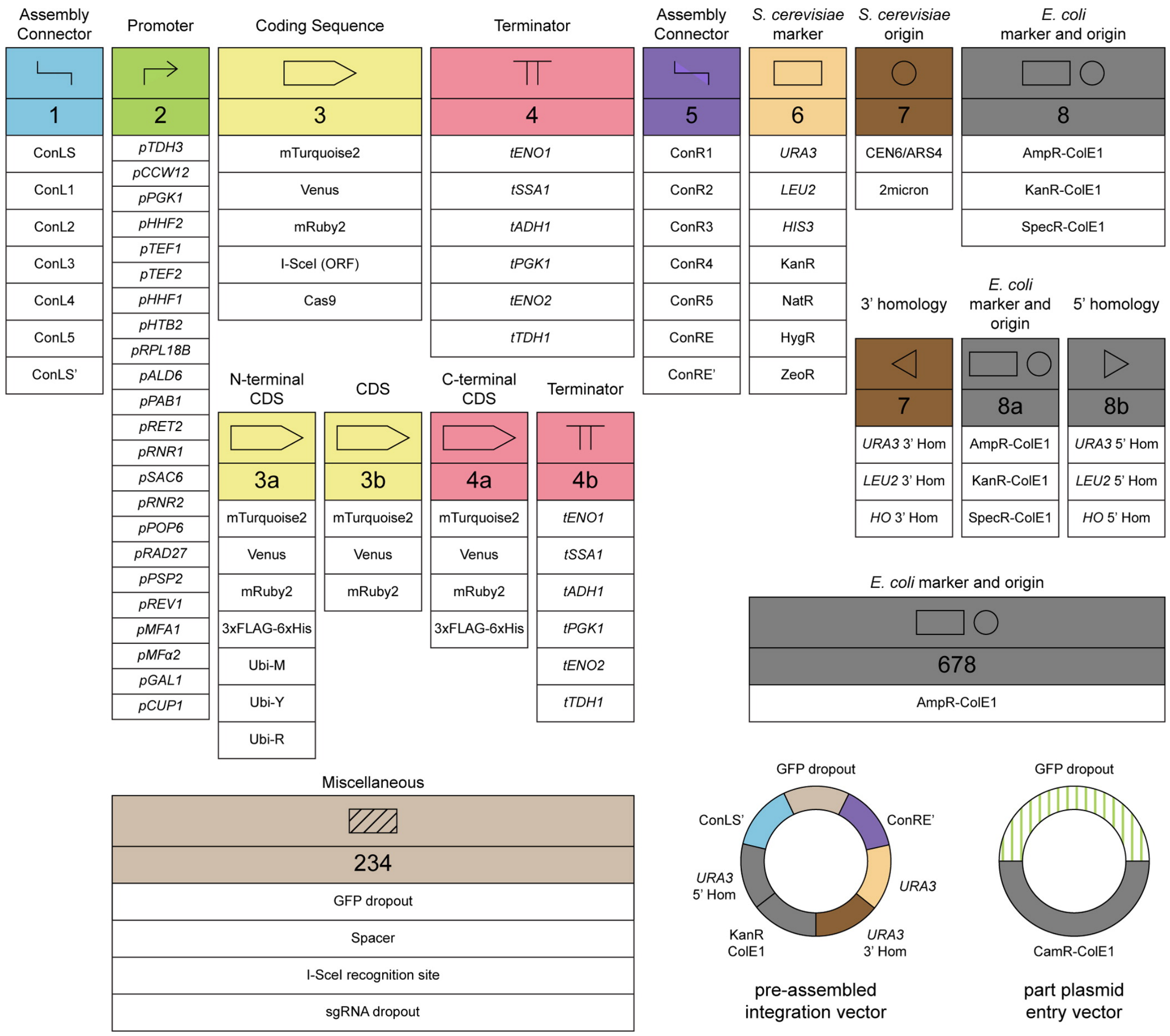


**Figure 2.7:** Library of characterised parts that forms the foundation of the Yeast ToolKit. Each position (1 through 8) fulfils a specific function in the final cassette plasmid. For example, position 3 supplies the ORF, while position 4 supplies the terminator. For flexibility, some positions can be sub-divided into 2 constituent parts. Position 3 can be fulfilled by supplying a 3a and a 3b part which carry a degradation tag and a coding sequence, for example. Figure from Lee *et al.* 2015[8].

## 2.2.1 Cloning in *E. coli*

The YTK was used in conjunction with a newly developed strain of *E. coli* in order to achieve assembly throughputs that had previously been impossible. *E. coli* was used for plasmid assembly. The constructs were subsequently integrated into the yeast genome. We describe each of these processes in more detail below.

### Strain, media, culturing and DNA manipulation

The *E. coli* strain that was used for all plasmid construction steps was the NEB Turbo strain (New England Biolabs C2984). This strain has markedly improved growth characteristics compared to traditional cloning strains such as DH5$\alpha$ and DH10$\beta$. Transformants growing on petri dishes could be picked after 8-9 hours and miniprep plasmid isolations could be performed as soon as 5 hours after inoculation, rather than o/n incubation with traditional strains. One possible disadvantage of the NEB Turbo strain is that it has retained its native copy of RecA, an enzyme involved in homologous recombination. This may increase evolutionary instability in constructs with homologous regions. In practice we have not observed a striking difference between NEB Turbo and DH10$\beta$.

Media types, incubation temperatures and antibiotic concentrations were identical to those described earlier in **subsection 2.1.1** on page 51. DNA manipulation was also performed as described above, with the exception that generic spin columns (NBS biologicals, Cat No NBS5005) and buffers were used for miniprep plasmid isolations, instead of the Qiagen kit. Miniprep buffers recipes are available on Open Wetware[155]. Golden Gate reactions were performed according to the methods described in the YTK paper[8]. Part-level constructs were always sequenced for determination of nucleotide level accuracy to the intended sequence. Subsequent steps are generally efficient and reliable to a degree that restriction enzyme digests suffice. In cases where circuit performance did not match expectations or experiments were particularly critical the higher level constructs were also sequenced as part of the QA process.

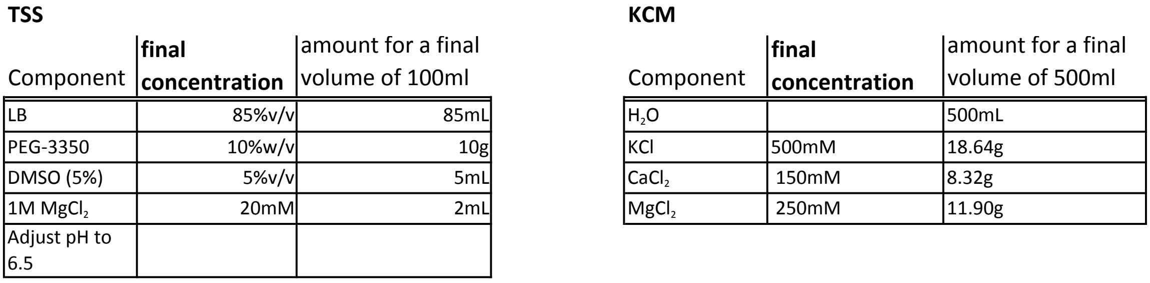
### Chemical transformation

For YTK transformations we utilised a chemical transformation protocol that allows the transformation steps to be performed on a thermocycler after mixing of the DNA and competent cells. This allowed further increase in throughput with lower costs and less hands-on time compared to electroporation. The protocol for preparation of chemically competent cells is as follows.

Required materials: Two sterile 2L flasks with 500mL of LB each. Two sterile 500ml centrifuge bottles or 10 50ml centrifuge tubes. A 10ml overnight culture of the appropriate strain of *E. coli*. Sterile TSS and KCM solutions, described in **Table 2.4** on the next page. The procedure is as follows:

1. Add 10 mL overnight culture to pre-warmed 1L LB (1:100 dilution).
2. Clean out and prechill the rotor and large centrifuge.
3. Pre-chill two large centrifuge bottles.
4. Grow culture to $OD_{600}$=0.5-1.0, 2-3 hours.
5. Shake on ice to stop the growth.
6. Spin cells at 3000rpm for 10min or 4700rpm for 5min.
7. Resuspend the pellets in the 50mL ice-cold TSS (25mL per 500mL culture batch).
8. Aliquot into PCR strips, 200$\mu$l per tube. approx 275 tubes total.
9. Freeze cells in liquid nitrogen bath, or place directly into the -80°C.

**Table 2.4:** Solutions required for the preparation of chemically competent cells.



**TSS**

| Component | **final concentration** | amount for a final volume of 100ml |
|---|---|---|
| LB | 85%v/v | 85mL |
| PEG-3350 | 10%w/v | 10g |
| DMSO (5%) | 5%v/v | 5mL |
| 1M $MgCl_2$ | 20mM | 2mL |
| Adjust pH to 6.5 | | |

**KCM**

| Component | **final concentration** | amount for a final volume of 500ml |
|---|---|---|
| $H_2O$ | | 500mL |
| KCl | 500mM | 18.64g |
| $CaCl_2$ | 150mM | 8.32g |
| $MgCl_2$ | 250mM | 11.90g |

Once competent cells have been prepared, the transformation can be carried out using the following steps:

1. Mix 50$\mu$l of the KCM into 200$\mu$l of competent cell prep, thawed on ice.
2. Add 50-100$\mu$l of comp cell-KCM cocktail to DNA (usually 90$\mu$l to 10$\mu$l golden gate reaction mix).
3. Incubate 10min on ice (or 4°C on a thermocycler).
4. Incubate 1min at 42°C.
5. Incubate 1min on ice (4°C on a thermocycler).
6. Incubate 37°C recovery for 15-60min (can skip for ampicillin).

### 2.2.2 Cloning in *S. cerevisiae*

Unlike the pRS system, the YTK is set up such that the selection marker and integration homology flanks are independent. This makes the YTK strain independent, since the selection markers no longer rely on inactivated essential genes (auxotrophies) that retain homology to the marker. The auxotrophies are allowed to be complete deletions, widening the selection of strains that are suitable for YTK integration. In addition, it allows dominant markers to be used instead of auxotrophic markers, which allows the system to be applied in virtually any strain of *S. cerevisiae*.

Another advantage of the integration mechanism in the YTK is the fact that it relies on ends-out rather than ends-in integration. This greatly reduces the probability for multiple integrations to occur. In ends-in integration, which relies on single-crossover events, the DNA sequence that is a product of an integration event can act as a substrate for a second integration event. We estimate that the frequency of multiple integrations in this system is 20-35%. While the products of ends-out integration, which relies on double-integration events, can act as substrate for subsequent integration events, the subsequent integration replaces the original integration on the genome. As such, there is no net increase in integrated constructs on the genome. We estimate that this leads to a multiple integration frequency of lower than 1%. This reduction brings about a meaningful decrease in the number of clones that need to be screened for multiple integration events. This further increases the throughput and cost effectiveness of the YTK method. A diagram of the ends-out integration process is shown in **Figure 2.8** on the following page.

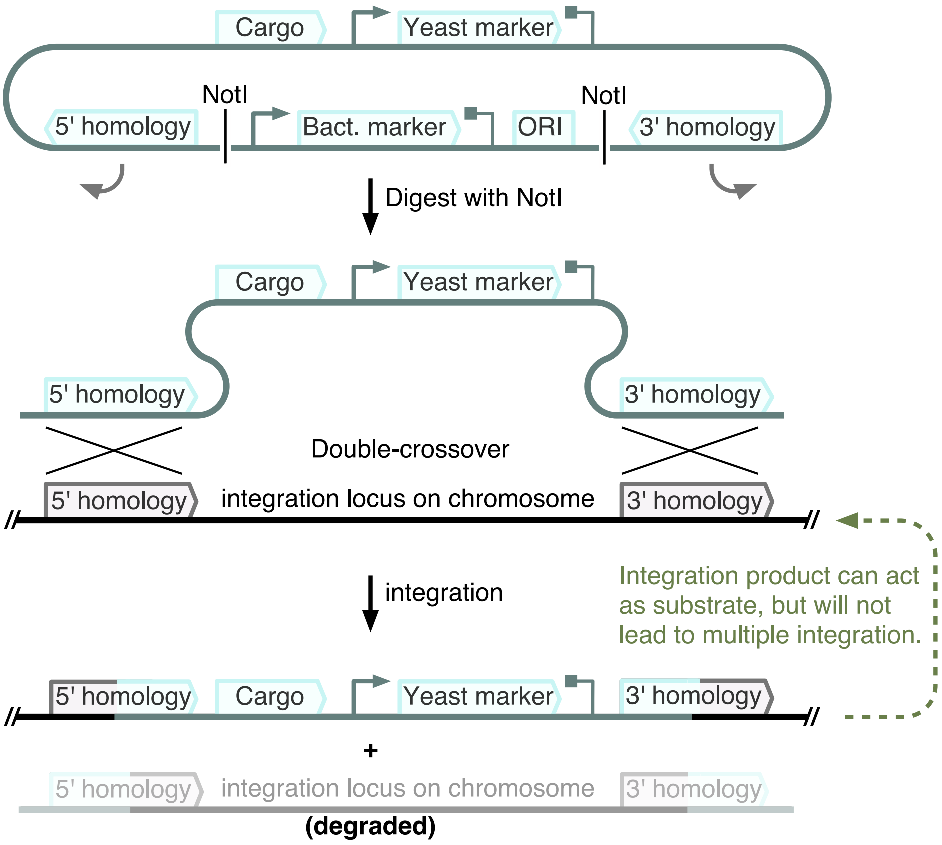


**Figure 2.8:** Mechanism of ends-out genomic integration used in the Yeast ToolKit. YTK construct is linearised with NotI to remove bacterial marker and ORI. Two homology flanks are exposed and pair with the corresponding homologous regions on the chromosome. Note how the yeast marker is independent of the integration locus. Technically the URA3 marker can be integrated into the LEU2 locus if this is desired. The construct is integrated in a double crossover event that replaces the pre-integration locus. For this reason, repeated integrations do not lead to multiple integrated copies on the genome.

#### Strain, media and transformation

The yeast strain that was used for constructs produced with the YTK was BY4741 (MATa his3$\Delta$1 leu2$\Delta$0 met15$\Delta$0 ura3$\Delta$0)[156]. This strain is lysine and adenine positive. Correspondingly, lysine cannot be used as an auxotrophic marker in this strain. Conveniently, media did not have to be supplemented with adenine for BY4741-based strains. Media, transformation protocols and strain handling were otherwise identical to what has been described for YPH500-based strains. Heat inactivation of NotI digested plasmids was found to not improve transformation efficiency and was therefore not performed for the majority of the work.

### 2.2.3 Flow Cytometry

Coinciding with the introduction of the YTK our lab also gained possession of next generation flow cytometry equipment. Unless otherwise stated, constructs assembled with the YTK were analysed on one of the two machines described below.

### Culturing

Along with higher accuracy and an increase in dynamic range, the two machines possessed reliable 96-well microtiter plate (MTP) autosamplers. This allowed for a significant increase in the number of transformants that could be screened per transformation while also decreasing the hands-on time required for the measurements. With the more accurate measurements it also became clear that culturing in minimal media reduced autofluorescence levels in wild-type cells. Lower autofluorescence leads to increased signal to noise ratios, which benefits the accuracy of the measurements.

With these new capabilities and insights we adapted the flow cytometry culturing protocol as follows: Per transformation 6-8 colonies were picked and grown to saturation in 700 $\mu$l YEP-Dextrose in 2ml 96-deepwell plates (VWR, cat no 732-0585). Cultures were grown overnight in a shaking incubator (Infors HT multitron MTP shaker) at 800rpm at 30°C, with breathe-easy film (sigma, cat no Z380059) covering the plate to prevent evaporation. This plate was then transferred to a new deepwell plate containing 700$\mu$l minimal media (compositions described earlier) with the relevant carbon source (glucose or dextrose). When required, colonies were inoculated into two plates, each with a different carbon source. Dilution at this step was 500 fold.

After overnight incubation of at least 12 hours, the cultures were backdiluted into a Costar 96 round-well flatbottom plate (VWR, cat no 3596). Dilution was 10-100 fold, depending on culture density. Total volume per well in these plates was 300$\mu$l of the same media that the particular sample had been incubated in overnight. Cultures were grown for a minimum of 4 hours before initiation of the measurements, to ensure logarithmic growth. These plates were then used directly for data acquisition without prior dilution in MQ, since the new machines allowed up to 30,000 events per second to be recorded.

### Equipment

Two flow cytometers were used for data collection on YTK-based constructs. Which machine was used is indicated in the caption of each relevant figure. The first machine was the Becton Dickinson LSR-Fortessa X-20 cell analyzer, with accompanying 96-well plate reader. Three lasers are installed in this machine: blue 488 nm, violet 405 nm and a 561 nm yellow/green laser. Green fluorescence (blue laser) was detected at a voltage of 450, red fluorescence (yellow laser) at a voltage of 604, while forward and side scatter were detected at voltages of 27 and 154, respectively.

The second machine was the Attune NxT Acoustic Focusing Cytometer (Invitrogen), with accompanying 96-well plate reader. Two lasers are installed in this machine: blue 488 nm and a 561 nm yellow laser. Green fluorescence (blue laser) was detected at a voltage of 450, red fluorescence (yellow laser) at a voltage of 480, while forward and side scatter were detected at voltages of 40 and 340, respectively.

### Data analysis

10,000 events were collected for each sample. Only events with forward and side scatter values greater than $10^3$ were counted. Populations were gated for sufficient mRuby2 expression in

**Table 2.5:** Typical touchdown PCR protocol.

| Temperature | Time | Cycles |
|---|---|---|
| 98 °C | 30 s | 1 |
| 98 °C | 10 s | |
| 61 °C decreasing 1 degree every cycle to 54 °C | 30 s | 8 |
| 72 °C | 30 s /1 kb | |
| 98 °C | 10 s | |
| 53 °C | 30 s | 27 |
| 72 °C | 30 s /1 kb | |
| 72 °C | 5 min | 1 |
| 4 °C | hold | |

strains that contained a constitutively expressed red fluorescent control. For fluorescence level determination, medians of the red and green channel were calculated using FlowJo 10.0.7r2. Averages of the median values of 6 to 8 transformants were reported. Error bars in these figures represent the standard deviation across the median values. We assume all 6-8 transformants are genetically identical. This is not necessarily the case. When obvious outliers were detected we performed a colony PCR to determine whether the transformant contained multiple integrations of the construct. If this was found do be the case, then the individual transformant was eliminated from the analysis.

When outcomes were close, or particularly critical, we calculated the statistical significance of the observed differences between certain measurements. Since these data were not reliant on single flow-cytometry experiments, we do not have to take into account a high occurrence of outliers. Significance could therefore be calculated with the commonly used two-sample, two-tailed t-test, as implemented in the *ttest2* matlab function. Results were reported as significant when the p-value was lower than 0.001. Effect sizes did not have to be calculated for this type of measurement, since systematic error in the flow cytometry measurements was mitigated through the averaging of the results for multiple (6-8) populations.

## 2.3 Molecular biology techniques

### Long-term storage of bacterial and yeast strains

Long-term storage of living biological samples such as strains of bacteria and yeast was done in 15% glycerol (final concentration). Glycerol stocks were held at -80 °C in an ultra-low temperature freezer (New Brunswick).

### PCR

Phusion DNA polymerase (New England Biolabs, Cat No m0530s) was used for all PCR reactions, according to the specifications in the Phusion manual[157]. In most cases a touchdown PCR protocol was used. A typical protocol is shown in **Table 2.5**. In all other cases, such as 2-step PCR, a protocol was derived directly from the Phusion manual[157].

**Table 2.6:** Typical colony PCR protocol.

| Temperature | Time | Cycles |
|---|---|---|
| 95 °C | 60 s | |
| 95 °C | 30 s | |
| 60 °C | 30 s | 32 |
| 72 °C | 60 s | |
| 72 °C | 7 min | |
| 4 °C | hold | |

### Colony PCR

Colony PCR of both Yeast and *E. coli* was performed by suspending a colony in 50 $\mu$l of 0.02 M NaOH with a sterile toothpick. After incubating this solution at 99 °C for 10 minutes, a 0.6 $\mu$l aliquot was used as template in the PCR reaction. Volume of the reaction was 12 $\mu$l. Since colony PCR was typically done with a large number of reactions (40+), Go Taq G2 Green Master Mix (Promega, Cat No M7823) was used for time and cost efficiency. In this protocol, addition of NaOH is essential, however, at high concentrations it will inhibit the PCR reaction. More than 5% of the NaOH solution in the PCR is not recommended and more than 10% has been shown to completely inhibit the reaction. Typically, targets for colony PCR were 1 kb or less. A typical protocol is shown in **Table 2.6**.

### Total RNA isolation

Total RNA was isolated from yeast using the YeaStar RNA Kit (Zymo Research, Cat No R1002), according to the manufacturer's instructions. Cultures were grown to saturation overnight and backdiluted 1:100 the next morning. Cultures were then grown to an O.D.$_{600}$ of approximately 2, to ensure logarithmic growth. At this point, 1.5 ml of the culture was used for RNA isolation. The volume that was used of each of the cultures was adjusted to ensure that the same amount of biomass was used for every culture.

### Reverse transcription

**Table 2.7:** Primers for specific reverse transcription.

| **Name** | Sequence | **Target** |
|---|---|---|
| TW165 | ACGACGTTAGTCCAGTCCTT | TPI1 |
| TW167 | CATTATCGTTGGGCTGGTCT | DOA1 |
| TW169 | ATGGGTAATACCAGCAGCAG | yeGFP |
| TW171 | TTCCACGATGGTGTAGTCCT | mCherry |
| TW173 | TAATGGCTCTCATTGCACCC | LacI |

400 ng of total RNA of each of the samples was subsequently used in the reverse transcription (RT) reaction to produce cDNA that could be used for qPCR. The RT reaction was performed in a total volume of 10 $\mu$l, using the Tetro cDNA synthesis kit (Bioline, Cat No BIO-65043) according to the manufacturer's instructions. Rather than random hexamers, we used specific primers which are listed in **Table 2.7**. For each reaction, a negative control lacking the reverse transcriptase was included.

### qPCR

cDNA obtained in the RT reaction was diluted 300x and used for qPCR. 4.6 $\mu$l of diluted cDNA was used in a total reaction volume of 10 $\mu$l. The Kapa universal qPCR 2x mastermix kit (KAPA biosystems, Cat No kk4601) was used according to the manufacturer's instructions. The primers (0.2 $\mu$l per primer per reaction) for each of the screened targets are listed in **Table 2.8**. Product lengths varied per target, causing differences in the measurements. This was not corrected for because the variations in length are relatively minor. Measurements were performed with the Eppendorf MasterCycler RealPlex qPCR termocycler and accompanying software. The default temperature cycling program was used, which is a 2-step protocol: denaturation for 10 min at 95°C followed by 50 cycles of 15 s at 95°C, 1 min annealing and extension at 60°C.

**Table 2.8:** Primer sequences and product lengths for targets in qPCR

| Name | Sequence | Product length | Direction | Target |
|---|---|---|---|---|
| TW164 | GTGTCGGTGTCATCTTGTGT | 124bp | Fw | TPI1 |
| TW165 | ACGACGTTAGTCCAGTCCTT | 124bp | Rev | TPI1 |
| TW168 | GGTGATGGTCCAGTCTTGTT | 129bp | Fw | yeGFP |
| TW169 | ATGGGTAATACCAGCAGCAG | 129bp | Rev | yeGFP |
| TW170 | TACGACGCTGAGGTCAAGA | 120bp | Fw | mCherry |
| TW171 | TTCCACGATGGTGTAGTCCT | 120bp | Rev | mCherry |
| TW178 | GTCAGCGACAACCCATATAC | 123bp | Fw | DOA1 |
| TW179 | CTGGTCTAGCGATATGCCATT | 123bp | Rev | DOA1 |
| TW180 | CAGACACCCATCAACAGTATTA | 128bp | Fw | LacI |
| TW181 | CTTGCTGAGACAGAACTGAG | 128bp | Rev | LacI |

Three technical replicates were performed for every biological sample. The data were analysed using the $2^{-\Delta\Delta C_T}$ (also 'dd-Ct') method[158]. The error was calculated as the standard deviation of the replicates. Because the C$_T$ value of the reference gene is subtracted from the C$_T$ value of the sample and because there also is uncertainty in the C$_T$ value of the reference gene, the error needed to be propagated. Since the data points are subtracted, error propagation can be calculated using the standard method for additive error propagation in uncorrelated variables. This calculation is given in the following formula:

$$Error(C_T1 + C_T2) = \sqrt{Error(C_T1)^2 + Error(C_T2)^2}$$

For each qRT-PCR experiment, two types of controls were included to monitor the level of DNA contamination of the cDNA and the used reagents. For every target (primer pair) we included a triplicate measurement of ddH$_2$O. This informed us if the used reagents had been contaminated with target DNA, since no signal should normally be found in these reactions. Secondly, for every target in every strain we included the -RT control samples produced during the cDNA synthesis. This informed us of the levels of genomic DNA isolated in the total RNA preparation. Typically, the signal for genomic DNA levels was at least an order of magnitude lower than the signal for the +RT sample.

# 3. Tuning expression levels with designed hairpins in the mRNA 5'UTR.

For biotechnology and synthetic biology there are various bioinformatics tools that can be used to aid in the design of genetic devices and circuits. These are most advanced in prokaryotic systems, especially *E. coli* where online tools can even be used to predict the expression level of a construct based on the DNA sequence of its parts. The most widely used of these part prediction tools is the RBS Calculator developed by Howard Salis and Chris Voigt as part of their research into predictable gene expression in *E. coli* synthetic biology[159]. Several similar ribosome binding site (RBS) prediction tools have since also been implemented by others and were recently reviewed by Reeve *et al.*[160]. All of these tools are based on a thermodynamic model that calculates the free binding energy between the ribosome binding site sequence on the mRNA and the ribosomal RNA within the 30S subunit of the ribosome. This calculated free energy can be used to predict how well the RBS sequence promotes the initiation of translation initiation from that site on the mRNA, which can in turn estimate the likely protein expression rate.

Owing to its popularity, the RBS Calculator has garnered hundreds of citations and is now used widely in synthetic biology and bacterial biotechnology. However, due to differences in the fundamental biology between prokaryotes and eukaryotes (see **section 1.5** on page 41), no equivalent tool exists for yeast or for eukaryotes in general. In this chapter, we aim to work towards the creation of such a tool, using the predictability of RNA base-pairing interactions which can give free energy values. By introducing a hairpin structure in the 5'UTR of yeast mRNA, we can modulate the translation initiation efficiency of the corresponding gene in a predictable manner. This gives us a basis for prediction and tuning of gene output, meeting an unmet need in yeast synthetic biology to fine-tune gene expression at the translation stage.

## 3.1 Introduction

The most basic representation of gene expression says that DNA makes RNA which makes protein. In prokaryotic synthetic biology we modify the promoter sequence of the gene to modulate the rate of the first step (transcription) and modify the RBS sequence to tune the second step (translation). In eukaryotic synthetic biology we typically only have the tools to modify gene expression output at the transcriptional level - regulated promoters and promoter libraries. In this chapter, we look at the opportunity to do something akin to that seen in the bacterial RBS Calculator system, where we exploit nucleic acid structures and modify expression. To help us understand the possible strategies for this work, we first examine how nucleic acid structures can be calculated and what factors affect their strength. From there we explore what is known about the effect of secondary structure on the translation of mRNAs, and finally we look at the current state of the art with regards to the use of RNA structure in tuning gene expression in yeast.

### 3.1.1 Computation of RNA folding strength and structure

A large body of theoretical and applied work exists on the modelling and prediction of base-pairing interactions in DNA and RNA. Since the concept of base-pairing is so fundamental to biology, applications span the entire breadth of the field, ranging from the calculation of primer annealing temperatures, to predicting expression strengths with the RBS Calculator, to predicting the structure of non-coding RNAs in order to elucidate their function.

The basic principle of nucleic acid pairing is well known: cytosine pairs with guanine and thymine pairs with adenine to form the iconic double helix. Because of a third hydrogen bond in the C-G pair, these bases form a stronger interaction than the A-T pair. In RNA, uracil (U) is found instead of thymine. In addition to its usual interaction with adenine, uracil can also pair with guanine, to form a weak interaction consisting of two hydrogen bonds.

The double helix is further stabilised by stacking interactions between the individual base-pairs. These are driven by Van der Waals forces originating from the favourable distance between the base-pair planes in the stack. Like in the base-pairing interactions, the strength of the stacking is dependent on the particular bases involved. These interactions are directional, meaning that the sequence 5'(G-C-C)3' with its complementary sequence 3'(C-G-G)5' will have a subtly different stability than the sequence 5'(C-C-G)3' with its 3'(G-G-C)5' complement, despite containing the same bases.

Because the formation of the double helix is based on discrete physical interactions, they can be quantified and used to compute the overall stability of a particular structure. Each of the stabilising interactions is associated with a thermodynamic contribution towards the stability. By summing the contributions of the interactions across the entire structure, a measure can be obtained for the stability, and thereby the likelihood, of a particular structure. By calculating the stability for all possible structures of a particular sequence, it is possible to calculate the most stable state.

The thermodynamic contributions of the discrete interactions in the structure are expressed in terms of the energy that is necessary to form them, per mole of substance. Since the formation of bonds is exothermic, energy is released in the process and the sign of the energy necessary to form the bond is negative. Another way to put this is that the system is always seeking to attain the state of the minimum free energy. For example, a hairpin of low stability can have an associated Minimum Free Energy (MFE) of -12 kcal/mol, while a highly stable hairpin could have an MFE of -41 kcal/mol. These values are also referred to as the $\Delta G$ and the lower the $\Delta G$ the stronger the structure.

Many software packages exist that calculate the MFE structure of a given DNA or RNA sequence. They follow the procedure outlined above, although each is implemented differently in order to make the problem computationally tractable[161]. New approaches to make algorithms more accurate and faster are continuously being developed[162–164]. Several popular packages are Mfold[165], UNAfold[166], NUPACK[167] and RNAstructure[168]. The accuracy for standard RNA folding problems generally overlaps between many of these tools. For our purposes in this work, we chose to use RNAfold from the ViennaRNA suite of tools, as it is one of the most widely used packages and allows the algorithms to be set up and run on a local computer with ease. It can adjust the thermodynamic parameters to temperatures other than 37°C and it has a long history of being in active development[169,170]

### 3.1.2 Tetraloops

So far we have discussed secondary structure (base-pairing interactions), but RNA can also fold into more complex (tertiary) structures, that have a stabilising effect on the total structure. Incorporating these structures into the structure prediction algorithms is non-trivial. The folding is often more complex and less well defined in terms of thermodynamic contributions, leading to inflated computational cost and more uncertainty in the final prediction.

Examples of these complex tertiary structures include the following:

- Pseudoknots: regions of hairpin loops that engage in base-pairing interactions elsewhere in the RNA molecule.
- Tetraloops: hairpin loops that are inherently more stable because of the specific bases present in the loop. Additionally, tetraloop-receptor motifs can form base-pairs with the tetraloop residues, stabilising the complex further.
- G-tetrads: configuration of 4 guanine residues in a plane, frequently with a metal ion bound in the centre. Multiple G-tetrads can stack together to form a G-quadruplex.
- A-minor motif: this motif consists of an adenine residue that is inserted into the minor groove of the double-helix, typically interacting with a G-C base-pair.

However, most of these tertiary structures are not relevant for our work because they typically only form in long RNA molecules that have evolved to include them. In the work within this chapter, only the tetraloop structure is relevant, as it can form in even some of the most basic hairpin structures. Indeed, tetraloops are so stable that they act as nucleation points for RNA folding[171–173]. The unusual stability of tetraloops has been known for over 25 years[174] and it is surprising that they are not commonly included in algorithms for RNA structure prediction. Recently, there have been some advances in the determination of the thermodynamic contributions and folding mechanism of tetraloops, which may pave the way for more widespread inclusion into future RNA structure prediction software tools[175–177].

Tetraloops were discovered because they are highly conserved features found in ribosomal RNA[183]. They owe their stability to non-canonical hydrogen-bond formation between the bases within the tetraloop. Many sequence variations exist, but in most tetraloop families the first and last bases are conserved, while the second base is the least constrained (see **Table 3.1** on the following page). The most common tetraloops are the 5'-UNCG-3' and 5'-GNRA-3' families[184].

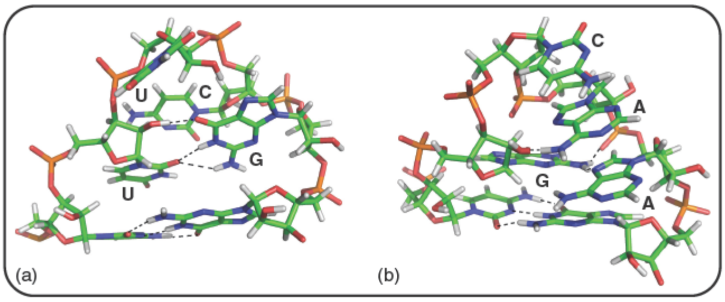


**Figure 3.1:** UUCG (a) and GCAA (b) tetraloop structures, members of the UNCG and GNRA tetraloop families, respectively. Hydrogen bonds are shown as dotted black lines. Figure reproduced from Cheong et al.[186].

Examples of these tetraloop families are also shown in **Figure 3.1**. Of these two families, the UNCG family is more thermodynamically stable than the GNRA family[185].

Not only do tetraloops initiate and stabilise secondary structures, but they can also mediate RNA-RNA contacts via the tetraloop-receptor motif, via kissing hairpin loops, A-minor interactions, and also by pseudoknots[187]. These kind of interactions stabilise the tertiary structure of the RNA and can be essential for any catalytic functions the RNA may have[188]. Of higher significance for the work in this chapter is the fact that tetraloops can also mediate RNA-protein interactions and have been shown to occur *in vivo* between RNAs and a multitude of endogenous and viral proteins[189–191]. These interactions can be used for interesting applications, such as fusing a tetraloop-binding protein to GFP in order to do real-time visualisation of an mRNA containing a tetraloop RNA structure[192]. In our plans to modulate translation by introducing RNA folding into mRNAs, RNA-binding proteins such as these also need to be considered, as these could modulate translation by steric hindrance or by being engineered to recruit translation initiation factors.

### 3.1.3 Evidence for translation inhibition by mRNA structure

It has been known for nearly three decades that secondary structures in the 5'UTR of an mRNA can inhibit protein expression in yeast[193,194]. Early experiments focused primarily on individual case studies, but more recently it has become clear that this is a general property of yeast biology. For example, three genome-wide studies have shown that a negative correlation exists between the efficiency with which an mRNA is translated and the structure over its translation start site[106,111,195]. This finding was further validated by the Segal lab, when they published the results of a study that randomised the 10 positions upstream of the start codon on the mRNA. Here, they found a significant association between thermodynamically stable secondary

**Table 3.1:** Common RNA tetraloop family sequences and references.

| sequence (IUPAC) | Sequence (expanded) | References |
|---|---|---|
| UNCG | U-A/U/C/G-C-G | Molinaro *et al.* 1995[178] |
| GNRA | G-A/U/C/G-A/G-A | Correll *et al.* 2003[179] |
| CUYG | C-U-C/U-G | Jucker *et al.* 1995[180] |
| UNAC | U-A/U/C/G-A-C | Zhao *et al.* 2012[181] |
| ANYA | A-A/U/C/G-C/U-A | Huang *et al.* 2005[182] |

structures (lower MFE) and reduced protein levels[114]. This indicates not just a correlation, but a causal relationship between mRNA structure and reduced expression.

The effect of mRNA structure has not just been found in yeast but also in higher eukaryotes such as mammalian cells[196,197] and plants[198]. In these systems, the amount of inhibition was the highest when the hairpin was positioned close to the 5' cap, whereas in yeast the strongest inhibition has been found to occur when the hairpin included the start codon and possibly the first dozen bases of the ORF[104,114,193–195].

The location that confers the highest inhibition offers some insights into the mechanism by which the hairpins repress protein expression. The current understanding is that the hairpin interferes specifically with the scanning of the 40S ribosomal subunit (or more specifically: the 43S translation initiation complex) along the 5'UTR of the mRNA. This is supported by the following findings:

- There is a genome-wide selective pressure against secondary structure around the start codon, but not in the ORF[106,111,195].
- In a mutant screen with an essential gene, revertants arose that had gained a new start-codon upstream of the hairpin[193].
- Targeting a hairpin binding protein to the mRNA increased repression of the hairpin, indicating that disruption of ribosomal scanning on the 5'-untranslated region, and not restriction of translational initiation *per se*, modulates the stability of nonaberrant mRNAs[131].
- Yeast lacks the DHX29 RNA helicase, which is involved in mRNA structure resolution. Higher eukaryotes have the DHX29 RNA helicase and correspondingly have longer '5 UTRs[199].
- mRNAs carrying 5'-secondary structures have been shown to have biphasic polysome distributions, indicating that the mRNA molecules are distributed between untranslated and well-translated subpopulations. This suggests that once 5'-secondary structures are unwound, they reform slowly relative to the rate of translation initiation in yeast[194].

### 3.1.4 Current applications of mRNA hairpins and aims

Hairpins in yeast mRNA have been used to regulate and tune expression in several instances. In most cases, the hairpins have been combined with other types of RNA-based translational regulation. In one example, RNA pseudoknots were used to induce ribosomal frameshifting to switch between expression states[200]. In two other cases, RNA binding proteins have been used to induce translational repression of the corresponding mRNA[201,202]. Finally, ATc-binding aptamers have been applied in one study to enable inducible knockdown of genes. These aptamer motifs were introduced into the 5'UTR of the targeted genes and in the presence of the ATc inducer, change their folding to inhibit translation[203].

None of these examples rely on the inhibition caused by stable hairpins alone, and therefore do not use the predictive power that is offered by the many software tools for RNA structure prediction. In one case, the authors utilise the repressive power of mRNA hairpins to tune bistability in a positive feedback loop, showing the potential of the approach but instead relying

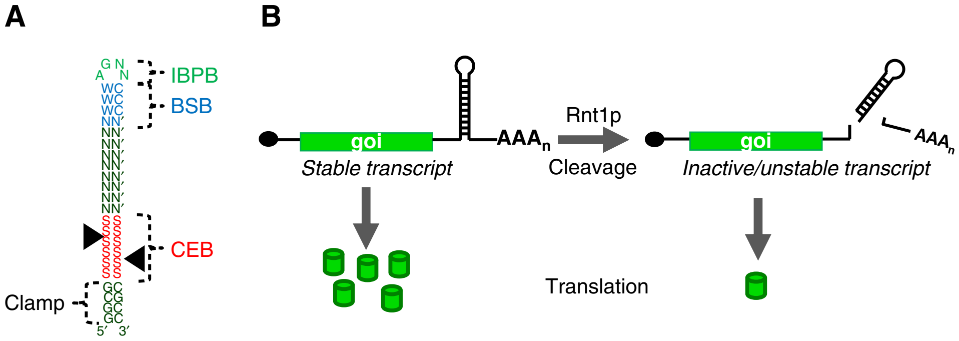


**Figure 3.2:** A system for tuning protein expression levels through modulating the cleaving efficiencies of Rnt1p on sequences placed in the 3'UTR of an mRNA. Figure adapted from Babiskin and Smolke, 2011[205].

on trial and error to find the correct expression levels[204]. In another case, the relationship between the minimum free energy of hairpin folding and the resultant expression levels is investigated, but this is not applied in a framework for prediction and implementation[199]. In yet another approach, also shown in **Figure 3.2**, mRNA hairpins acting as cleavage substrates are applied in library-creation framework, but the yield of library members is so low that enriching via cell-sorting techniques needs to be applied to accumulate suitable numbers for downstream applications[205].

**In this chapter,** we aim to unify elements from the examples above into one coherent framework for the design and implementation of RNA sequence libraries with defined ranges of gene expression. In this framework, we intend to incorporate the most advantageous features of existing approaches, in order to create a method that:

1. Introduces hairpins into the 5'UTR of mRNAs to modulate expression via modifying translation efficiency.
2. Utilises the power of RNA structure prediction to create libraries of parts that can be used to give predictable expression strengths.
3. Can be used for forward engineering in synthetic biology applications.
4. Is straightforward to implement at the DNA cloning level.
5. Does not require specialised equipment, such as fluorescence activated cell sorting (FACS).
6. Can be implemented with short turnaround times.

## 3.2 Experimental strategy

The workflow and software tool developed in this chapter were designed to generate libraries of 5'UTR sequences that yield a known range of expression outputs from a translated mRNA. By taking advantage of the ability to synthesise DNA oligomers that have undefined (N) or partially-defined (e.g. Y or W) bases at known positions, the premise was to allow a user to purchase a custom oligo that, through simple cloning steps, would yield yeast cells containing a library of 5'UTRs that give a diverse yet bounded range of expression outputs. The library could immediately be used in experiments where expression diversity among different cells is desirable such as for optimising expression levels during combinatorial cloning of a heterologous metabolic pathway. Alternatively, if sequences giving defined expression outputs are required, these can be selected from individual colonies within the characterised library.

The protocol for 5'UTR library generation was developed primarily for speed and accessibility and an overview is shown in **Figure 3.3** on the next page. A typical synthetic biology project consists of a design, build and test phase and this approach was taken here. Briefly, in the design phase a hairpin scaffold is selected and degenerate bases are rationally introduced into the design at key positions. The library design is simulated to give expected expression levels *in vivo* and when approved, a short series of cloning steps is performed using Golden Gate based cloning and primers incorporating the selected degeneracies. After obtaining yeast transformants, characterisation is performed using flow cytometry. Each of these steps is further explained below.

### 3.2.1 Design

The design phase starts with several computational steps to simulate and predict the behaviour of the library *in vivo*. Several bioinformatics tools exist that are instrumental to this process. In this procedure we use the ViennaRNA suite for RNA structure prediction and free energy calculation[169]. The first step is the selection of a scaffold hairpin, which by definition represents the strongest structure present in the library. A subset of nucleotides in this structure is then substituted for degenerate nucleotides. A set of design rules detailed in **section 3.3.1** on page 84 guides the choice of hairpin structure and the locations of degeneracy substitutions.

Based on the degeneracies, a list is compiled of all possible 5'UTR sequences that could arise. The predicted structure and free energy of folding are calculated using an offline version of RNAfold. The resulting $\Delta G$ values are converted to predicted average expression levels *in vivo*. The prediction is compared to the desired spread of expression levels in the library and if necessary the design can be refined by making adjustments to the nucleotide degeneracies.

### 3.2.2 Build

The build phase starts when the sequences have been finalised and primers containing the selected degeneracies have been ordered. The cloning strategy was designed to take 3 days or less for DNA construct preparation or up to 7 days from primer to fully characterised library. **Table 3.2** on page 80 shows an overview of a typical cloning schedule.

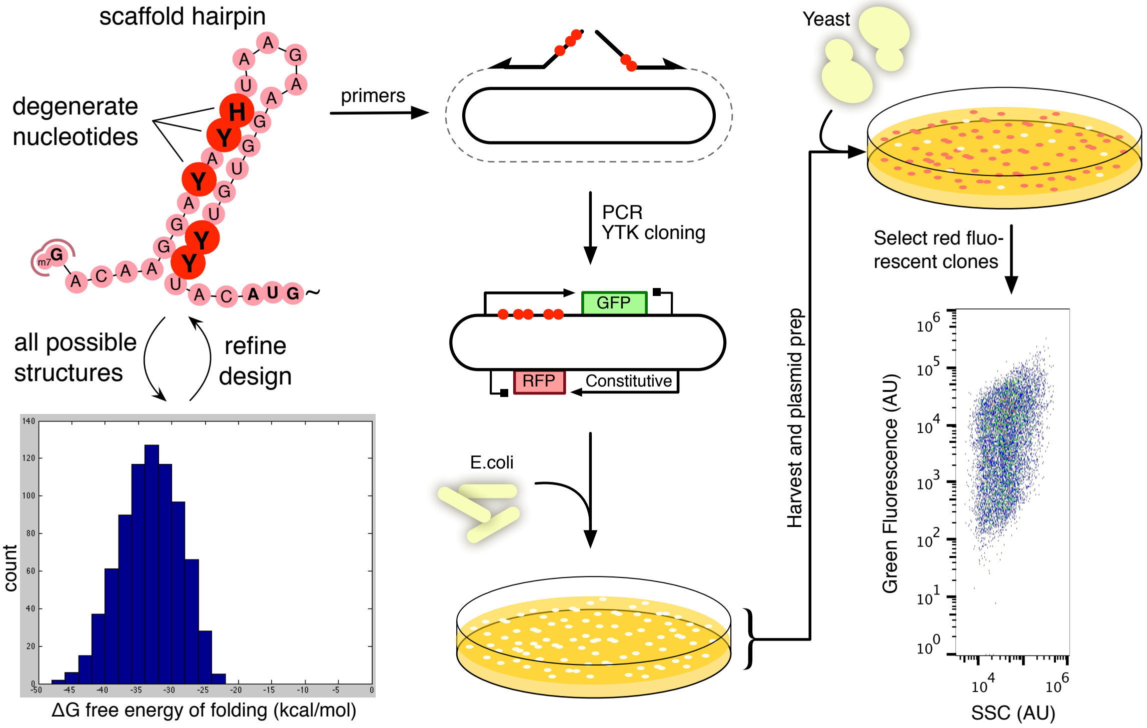


**Figure 3.3:** The process of 5'UTR hairpin library creation. Degenerate nucleotides are inserted at various positions into the design of a hairpin scaffold. The minimum free energy of all possible sequences is calculated and visualised in a histogram. If necessary the composition and location of the degenerate nucleotides is adjusted to produce the required distribution of free energies. Primers synthesised with degenerate bases are then ordered when the design meets the requirements. A library of plasmids for *E. coli* transformation is created using PCR and a subsequent Golden Gate based assembly step is implemented in Yeast ToolKit (YTK) format. The library of *E. coli* transformants is harvested and plasmids prepped for yeast transformation. Transformant colonies are pooled and analysed using flow cytometry. Clones that do not show constitutive red fluorescence are discarded in quality control for correct assembly. The diversity of the library of hairpins is reflected in the spread of green fluorescence over three orders of magnitude.

The proposed schedule can be shortened when time is critical. Day one and two can be combined to shorten lead time by a day. If the yeast promoter does not require induction, flow cytometry can be performed when yeast colonies are large enough to be picked or pooled, with only a few hours of culturing in liquid media. This can shorten the lead time by an additional day. Total lead time can thus be as short as 5 days, making this a particularly rapid method for expression library creation in contrast to synthetic promoter library production.

The high efficiency of this method is partly afforded by use of the Yeast ToolKit (YTK) system and by extension Golden Gate cloning[8]. The few accessory plasmids needed in the cloning process can be quickly generated using the collection of parts included in the YTK system. If required parts are missing, the kit allows for flexible incorporation of custom parts into the workflow.

**Figure 3.4** on the following page illustrates the process of plasmid construction in more detail. The degenerate hairpin sequence (indicated with red dots) is introduced into the 5'UTR

**Table 3.2:** Typical cloning schedule for the creation of a promoter library in yeast. When time is critical, cloning steps of day 1 and 2 can be combined.

| | |
|---|---|
| Day 1 | PCR to incorporate degeneracies<br>DpnI treatment to remove template<br>Golden Gate ligation with BsaI |
| Day 2 | Multigene Golden Gate assembly with BsmBI<br>High efficiency transformation of *E. coli* |
| Day 3 | Pooling of *E. coli* tranformants<br>Plasmid prep and NotI digest<br>High efficiency yeast transformation |
| Day 4-6 | Outgrowth on plates<br>Pooling and inoculation into (inducing) minimal media over night |
| Day 7 | Flow cytometry analysis |

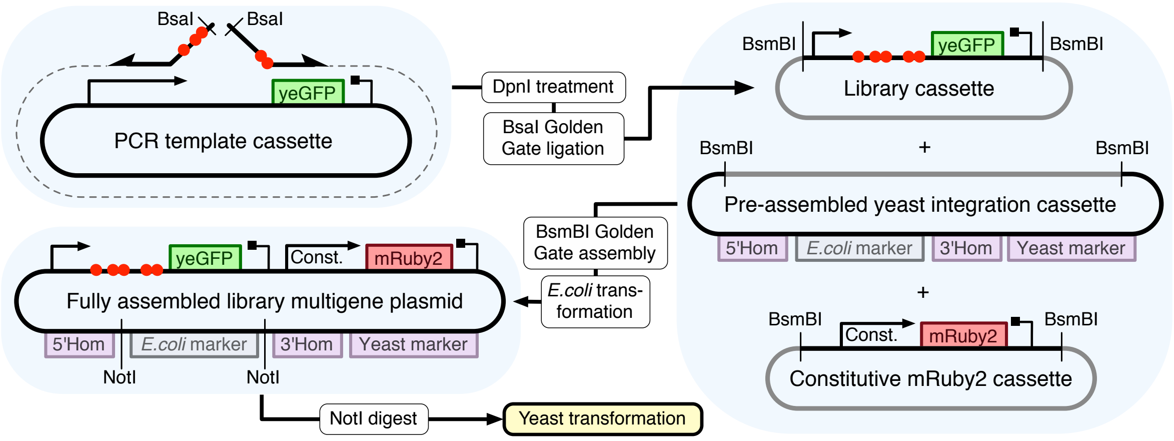


**Figure 3.4:** Cloning workflow for 5'UTR hairpin library construction using the Yeast ToolKit (YTK) system. **1)** Primers encoding the hairpin library of choice (indicated by the red dots) are used to amplify a template cassette carrying the selected promoter and a green fluorescent reporter gene (yeGFP). A DpnI digestion is performed to eliminate the template and the PCR product is subsequently self-ligated in a BsaI Golden Gate reaction. **2)** The resulting plasmid is purified and directly used in the multigene assembly, bypassing the typical intermediate transformation in order to preserve library diversity. The multigene BsmBI Golden Gate assembly is performed with two additional cassettes: a preassembled yeast integration cassette containing a selectable marker and homology regions for integration, and a constitutively-expressing red fluorescent reporter gene (mRuby2). **3)** NEB turbo competent *E. coli* is transformed with the multigene assembly. The thousands of resulting transformants are pooled and miniprepped. Finally the library of plasmids is digested with NotI before being used in a high efficiency yeast transformation resulting in hundreds to thousands of 5'UTR library candidates.

by PCR. A cloning step assembles the PCR product into the final integration vector. In principle, the PCR incorporating the library sequences could have been performed on the final integration vector directly. In practice, however, PCR on large plasmids can be problematic and may require time consuming optimizations. For the sake of robustness it was decided that a cloning step would be included to allow for PCR on small plasmids only.

To limit the number of cloning steps and the associated loss in library diversity, the PCR product is used directly in the multigene assembly step without passage through *E. coli*, as the YTK protocol suggests. To this end, the PCR product is DpnI treated to remove the template plasmid and subjected to a BsaI reaction to facilitate self-ligation. Equimolar quantities of this product are used with a pre-assembled yeast integration cassette and a cassette for constitutive mRuby2 expression in a multigene assembly step.

To further speed up the process of cloning and to maximise library diversity, the multigene assembly reaction is transformed into the fast growing NEB Turbo competent *E. coli* strain. To allow for parallelization, the transformation was performed with chemically competent cells. The turbo strain can be grown up for miniprep in as little as 5 hours, allowing considerable time savings.

To complete the library construction process, the pooled and miniprepped multigene assembly is digested with NotI in preparation for yeast transformation. The transformation itself was carried out using a high efficiency protocol[206] and a large amount of linearised DNA (3-5 $\mu$g) to typically yield thousands of library candidates.

### 3.2.3 Test

Performance of the designed and built library is finally assessed through flow cytometry analysis. A recurring challenge in cytometric data analysis is the use of Arbitrary Units (AU) in data collection. Depending on the settings used during data collection and the specific instrument used, identical specimens can give wildly different fluorescence values. This poses problems when data needs to be compared across collection dates, users, instruments and labs. Standardised ways of calibrated analysis have only recently been developed[207] and are not in widespread use.

In a crude attempt to overcome this issue, the acquired fluorescence values were normalised to the cellular autofluorescence. Dividing the fluorescence of measured events by the median fluorescence of a population of parental (untransformed) cells yields a unitless number. This number represents the fluorescence levels of a construct as a multiple of the autofluorescence of a cell. In order to maximise the signal-to-noise ratio, measurements must be performed in minimal media because rich growth media causes higher autofluorescence levels.

Golden Gate assembly is efficient, but not to 100%. The YTK system allows for visual screening of assembly success using a bacterial GFP dropout gene in *E. coli*. However, at a scale of thousands of library members, visual inspection and manual picking of hits is unworkable. A cassette for constitutive expression of mRuby2 is included to solve this problem. Any incorrect assemblies indicated by a lack of red fluorescence can simply be excluded from the analysis. This allows the entire *E. coli* transformation plate to be pooled and miniprepped and transformed into yeast without the need to exclude *E. coli* colonies with incorrect assemblies. Secondly, double integration events into yeast can be detected from the doubling of the red fluorescent signal and these cells can also be excluded from the analysis.

Taking the above into account, a typical workflow for flow cytometry data acquisition and analysis is as follows:

1. Pool yeast transformants and incubate overnight in minimal media under inducing conditions (a galactose carbon source was used in these experiments).
2. Dilute a sample from the overnight growth and take flow cytometry measurements.
3. Gate samples to only include events with typical SSC and FSC values for yeast.
4. Gate samples to only include events with typical red fluorescence values.
5. Determine median fluorescence in parental strain.
6. Divide fluorescence values of library samples by median autofluorescence.
7. Plot events on a density coloured scatter plot (fluorescence versus SSC).
8. When the sample is not a library, additional statistics are gathered, such as median and robust standard deviation.

## 3.3 Materials & Methods

### 3.3.1 Library design

As detailed below, library design consisted of selection of a scaffold hairpin followed by the selection of nucleotides that are to be replaced with degenerate nucleotides. Several computational steps then lead to the prediction of a distribution of expression levels resulting from the designed library. Here, we go into detail about the design considerations and computational packages used in each of these steps.

#### Location of the hairpin

Evidence exists that indicates that the position of the hairpin in the '5 UTR plays a role in the observed expression inhibition. In particular, hairpins incorporating the start codon have been shown to have a particularly large inhibitory effect[114]. This poses challenges for library design, as it is difficult to model and predict the impact of various types and strengths of pairing to the start codon and surrounding sequences. Including the start of the ORF in hairpin design also makes the system less modular, since changing the ORF will now impact base pairing in (part of) the stem.

Conversely, placing the hairpin far away from the start codon can lead to other problems. Not all yeast 5' UTRs are sufficiently long to introduce the hairpin at a distance of, for example, 50bp upstream of the start codon. Additionally, a varying distance to the start codon may have unintended effects. Taking these notes into consideration led to the conclusion that the optimal location for insertion of the hairpin is 5-15 bp upstream of the ATG. Care should be taken to avoid the start codon from being unintentionally incorporated into the hairpin.

#### Scaffold hairpin selection

The scaffold hairpin is defined as the hairpin structure that the hairpin library is based on. It generally represents the strongest member of the library, as any degeneracy introduced will typically lower the stability of the structure. Any sequence can in principle be used as a scaffold, as long as it conforms to the following design requirements:

- The sequence is entirely or largely palindromic (it is a strong hairpin).
- The sequence has a minimum free energy of folding that is as low as or lower than the lowest required member of the library.
- There is no premature start codon contained in this sequence.
- The loop of the hairpin does not contain an unusually stable sequence known as a tetraloop (see below).
- The two halves of the hairpin can be contained in the tails of primers used to amplify the selected promoter-ORF combination.
- The sequence does not contain BsaI, BsmBI or NotI restriction sites.

- The sequence does not contain other forbidden restriction sites or sequences that interfere with the function of the construct in the chosen application.

In our experiments, the UUCG tetraloop was shown to have a dramatic and unpredictable effect on expression. Inclusion of tetraloop sequences such as GNRA[179,183], UNCG[178,183] and CUYG[180,183] (where R is A or G and Y is C or T) is therefore advised against. Tetraloops can be used with caution when cloning limitations constrain the design to extremely short hairpins. Their utility is in the fact that they can lower the effective MFE into the desired range without increasing hairpin length.

Initially the hairpin scaffolds used in this project were based on the ideal Lac operator sequence. These were subsequently extended and modified to accommodate stronger libraries and shorter cloning primers. However newly designed scaffold can be based on the current 5' UTR - by adding an inverted repeat of this sequence - or on a de novo structure created through one of many inverse RNA structure prediction solutions available[169,208,209].

### Degeneracy selection

In principle, the degeneracies can be inserted at any base paired position in the hairpin. The requirements outlined above extend to all possible sequences that can arise from the inserted degenerate nucleotides. Any identified problems can be rectified manually by choosing different degeneracies. Substitution of the cytosine in a G-C base pair is preferred, as the opposing guanine is capable of forming the weaker non-canonical G-U pair in addition to the strong canonical pair. This way, variety can be introduced into the hairpin without disrupting its basic structure. In terms of degenerate nucleotide code this is the substitution of the cytosine in a G-C base pair by the Y (C or T) degenerate nucleotide.

More variety can be created still by also allowing a non-pairing base at this position. Both G and A nucleotides would accomplish this. However, introducing the G nucleotide carries the risk of introducing both strong and weak pairing with unintended bases elsewhere in the sequence. Therefore inclusion of guanines at locations where these were not originally present is generally avoided. This leaves adenine as the third option in addition to the original C and T possibilities, resulting in the degenerate base H.

Avoiding potential strong pairings for the reason above, W is the only degeneracy generally substituted at either side of the A-T pair. Additionally, degeneracies introduced at both sides of the pair are less effective, as the probability for a matching basepair decreases drastically as the number of possibilities at each side of the pair increases.

Exceptions can be made when these guidelines clash with the design rules stated in the previous section. However, the position causing a problem can also be left unchanged in favour of a substitution elsewhere. Typically, hairpins can be extended to such a length that not every base pair needs a substitution in order to create the desired library.

### Sequence list generation

For the next step in the process, a fasta file needs to be generated containing all possible sequences that can arise from the selected degeneracies. This was done using a basic python script that loops through all possibilities at every position in the mRNA from the transcription start site to the end of the start codon. There are indications that secondary structure involving the start codon has a strong impact on repression. We therefore chose to include it and any preceding sequence in the calculations. The cutoff directly after the ATG is still somewhat arbitrary and may be optimised in the future.

### Folding energy and expression level prediction

The generated list of sequences in multi-FASTA format serves as the input for the RNAfold script used for the folding energy estimation. This is part of the ViennaRNA package 2.0, which is a widely used suite of tools centred around the many facets of RNA structure analysis and prediction[169].

The key non-default parameter used in this script is `-T30`. This modifies the energy parameters used in the script to simulate 30°C, the typical temperature for yeast growth. Other parameters used for these calculations are `-d2 -noLP -noPS`. Following from this, the full command for script execution is:

```
RNAfold -T30 -d2 -noLP -noPS < Sequence_list_1.fa > Output_MFE_list_1.txt
```

The MFE value for each of the input sequences is then extracted from the output file. This list is used to create a histogram for visualisation of the distribution of folding energies and further to calculate the predicted distribution of expression levels. The logistic fit obtained in **Figure 3.4.1** on page 91 takes the MFE values as input and outputs the predicted normalised median fluorescence value for each of the members of the library. Finally, the obtained expression values are plotted as a histogram and compared to the intended expression profile. At this point adjustments can be made to the degeneracies in the hairpin scaffold.

## 3.3.2 Library cloning

The YTK cloning system was used for library plasmid construction[8]. NEB Turbo chemically competent *E. coli* (Catalog #C2984I) were used for transformation of library constructs, according to the manufacturer's protocol. **Table 3.3** on the following page shows a list of all libraries created for this study. The library plasmid is a multigene plasmid created from 3 cassette-level plasmids. One of these is an in vitro product created via PCR (see also **Figure 3.4** on page 80). The corresponding PCR reactions for each cassette part in these libraries are shown in the adjacent table. Primers are defined in **Table 3.7** on page 89.

Other cassettes used in the multigene assembly reactions and as template for the PCRs are listed in **Table 3.5** on page 87, along with part plasmids used in their assemblies. Any used parts that were not defined as a standard part in the YTK kit are listed in **Table 3.4** on the following page. BY4741 was used for all yeast transformations, using a high efficiency yeast transformation protocol[206].

**Table 3.3:** Yeast ToolKit multigene level library plasmids with corresponding PCR-generated cassette-level parts. PCR amplifications were subjected to DpnI treatment and a BsaI-Golden Gate digestion-ligation prior to incorporation into the multigene YTK assembly. Template plasmids are described in the table below.

| Library | Description | Cassette 1 | Cassette 2 | Entry vector |
|---|---|---|---|---|
| HL1 | pLX-yeGFP -8.0 kcal/mol library | PCR-HL1 | T701.1 | pYTK096 |
| HL2 | pLX-yeGFP -20.2kcal/mol library | PCR-HL2 | T701.1 | pYTK096 |
| HL3 | pLX-yeGFP -23.4 kcal/mol library | PCR-HL3 | T701.1 | pYTK096 |
| HL4 | pLX-yeGFP -25.8 kcal/mol library | PCR-HL4 | T701.1 | pYTK096 |
| HL5 | pLX-yeGFP -28.8 kcal/mol library | PCR-HL5 | T701.1 | pYTK096 |
| HL6 | pLX-yeGFP -32.2 kcal/mol library | PCR-HL6 | T701.1 | pYTK096 |
| HR1 | pLX-mRuby -32.2 kcal/mol library | PCR-HR1 | T700.1 | pYTK096 |
| HR2 | pLX-mRuby -28.8 kcal/mol library | PCR-HR2 | T700.1 | pYTK096 |
| HR3 | pLX-mRuby -23.4 kcal/mol library | PCR-HR3 | T700.1 | pYTK096 |
| HT1 | pLX-yeGFP tetraloop library | PCR-HT1 | T701.1 | pYTK096 |
| HT2 | pLX-yeGFP tetraloop library | PCR-HT2 | T701.1 | pYTK096 |
| HT3 | pLX-yeGFP tetraloop library | PCR-HT3 | T701.1 | pYTK096 |
| HT4 | pLX-yeGFP tetraloop library | PCR-HT4 | T701.1 | pYTK096 |
| HG1 | pLX-yeGFP -28.8 kcal/mol library | PCR-HG1 | T701.1 | pWS065 |
| HC1-TDH | pTDH3 -28.8 kcal/mol library | PCR-HC1 | T701.1 | pYTK096 |
| HC1-PGK | pPGK1 -28.8 kcal/mol library | PCR-HC2 | T701.1 | pYTK096 |
| HC1-TEF | pTEF2 -28.8 kcal/mol library | PCR-HC3 | T701.1 | pYTK096 |
| HC1-YRA | pYRA1s -28.8 kcal/mol library | PCR-HC4 | T701.1 | pYTK096 |
| HC1-POP | pPOP6 -28.8 kcal/mol library | PCR-HC5 | T701.1 | pYTK096 |

| PCR reaction | Fw | Rev | Template |
|---|---|---|---|
| PCR-HL1 | TW318 | TW323 | T827 |
| PCR-HL2 | TW317 | TW323 | T827 |
| PCR-HL3 | TW317 | TW394 | T827 |
| PCR-HL4 | TW317 | TW392 | T827 |
| PCR-HL5 | TW317 | TW393 | T827 |
| PCR-HL6 | TW317 | TW438 | T827 |
| PCR-HR1 | TW411 | TW394 | C6 |
| PCR-HR2 | TW411 | TW393 | C6 |
| PCR-HR3 | TW411 | TW438 | C6 |
| PCR-HT1 | TW317 | TW329 | T827 |
| PCR-HT2 | TW319 | TW330 | T827 |
| PCR-HT3 | TW319 | TW329 | T827 |
| PCR-HT4 | TW318 | TW330 | T827 |
| PCR-HG1 | TW439 | TW440 | T827 |
| PCR-HC1 | TW431 | TW432 | T897 |
| PCR-HC2 | TW431 | TW434 | T898 |
| PCR-HC3 | TW431 | TW433 | C7 |
| PCR-HC4 | TW431 | TW435 | T899 |
| PCR-HC5 | TW431 | TW437 | T901 |

**Table 3.4:** Custom part-level plasmids for the Yeast ToolKit used for the creation of 5'UTR hairpin libraries.

| Name | Description | Type | Sequence including BsmBI sites (standard backbone not shown) |
|---|---|---|---|
| T655 | pLX Gal1 derived LacI repressible promoter | 2 | GGTCTCAAACGGAAGTACGGATTAGAAGCCGCCGAGCGGGTGACAGCCCTCCGAAGGAAGACTCTCCTCCGTGCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACTAGCTTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATGAACGAATCAAATTAACAACCCTAGGATGATAATGCGATTACTTTTTTAGCCTTATTTCTGGGGTACTGCAGCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGCAAAAACTGTTGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAATATTACTTCTTATTCAAATGTAATAAAAGTATCAACAAAAAATTGTTAATATACCTCTATACTTTAACGTCAAGGAGAAAAACCCCAAATATGTGAGACC |
| T675 | yeGFP | 3 | GGTCTCATATGGTTTCTAAAGGTGAAGAATTATTCACTGGTGTTGTCCCAATTTTGGTTGAATTAGATGGTGATGTTAATGGTCACAAATTTTCTGTCTCCGGTGAAGGTGAAGGTGATGCTACTTACGGTAAATTGACCTTAAAATTTATTTGTACTACTGGTAAATTGCCAGTTCCATGGCCAACCTTAGTCACTACTTTCGGTTATGGTGTTCAATGTTTTGCTAGATACCCAGATCATATGAAACAACATGACTTTTTCAAGTCTGCCATGCCAGAAGGTTATGTTCAAGAAAGAACTATTTTTTTCAAAGATGACGGTAACTACAAGACCAGAGCTGAAGTCAAGTTTGAAGGTGATACCTTAGTTAATAGAATCGAATTAAAAGGTATTGATTTTAAAGAAGATGGTAACATTTTAGGTCACAAATTGGAATACAACTATAACTCTCACAATGTTTACATCATGGCTGACAAACAAAAGAATGGTATCAAAGTTAACTTCAAAATTAGACACAACATTGAAGATGGTTCTGTTCAATTAGCTGACCATTATCAACAAAATACTCCAATTGGTGATGGTCCAGTCTTGTTACCAGACAACCATTACTTATCCACTCAATCTGCCTTATCCAAAGATCCAAACGAAAAGAGAGATCACATGGTCTTGTTAGAATTTGTTACTGCTGCTGGTATTACCCATGGTATGGATGAATTGTACAAAGGATCCTGAGACC |
| pTMP030 | YRA1s promoter | 2 | GGTCTCAAACGAAACTTGTGGGCGCAATTATAAAACACTGCTACCAATTGTTCGTTTTCTGTTCATTAACACATAAAAAACCCTTATGTAACTATATTTACAAAGTAAATACGTATATTAAAGCTATTTTACCACTACCACAGAGTTCTTTGTCCAGTTGCTAGTATTTTTTTTTCGCGACGAGGCAGGGGCGGGTAGACGTGTTGTTTTTCCACGGCTTTCGGCTCACCACTTGAAGAACTATAAAAGGCCGCCAAATTTATCCTTTTTCACTTCTTCCGTTCGCTTTTTTCTGTCATTCCTATCGTGTGTTTAGTAGTAGGTTTTTTTGTTAGAAGAAGTTTTATCCGAAAACTATCGAtGACAAATAGATAAAAAAATCTCCCTCGTTCTATTTGAAACTTTAAGAAATCCATATTAAGAAAATACCTACATCTGCTAAAGATCTATGTGAGACC |

**Table 3.5:** Yeast ToolKit cassette level plasmids used in the construction process.

| Name | Description | Parts used in assembly | | | | | | |
|---|---|---|---|---|---|---|---|---|
| T827 | pGAL1-yeGFP cassette | pYTK002 | T655 | T675 | pTMP065 | pYTK068 | pYTK095 | |
| C6 | pGAL1-mRuby2 cassette | pYTK002 | T655 | pYKT034 | pTMP065 | pYTK068 | pYTK095 | |
| T897 | pTDH3-yeGFP cassette | pYTK002 | pYTK009 | T675 | pTMP065 | pYTK068 | pYTK095 | |
| C7 | pTEF2-yeGFP cassette | pYTK002 | pYTK014 | T675 | pTMP065 | pYTK068 | pYTK095 | |
| T898 | pPGK1-yeGFP cassette | pYTK002 | pTMP030 | T675 | pTMP065 | pYTK068 | pYTK095 | |
| T899 | pYRA1s-yeGFP cassette | pYTK002 | pYTK018 | T675 | pTMP065 | pYTK068 | pYTK095 | |
| T901 | pPOP6-yeGFP cassette | pYTK002 | pYTK024 | T675 | pTMP065 | pYTK068 | pYTK095 | |
| T700 | pTEF-yeGFP cassette | pYTK004 | pYTK013 | T675 | pYTK054 | pYTK072 | pYTK095 | |
| T701 | pTEF-mRuby2 cassette | pYTK004 | pYTK013 | pYTK034 | pYTK054 | pYTK072 | pYTK095 | |
| pWS065 | pre-assembled HIS integration cassette | pYTK008 | pYTK047 | pYTK094 | pYTK073 | pYTK076 | pYTK088 | pYTK090 |

**Table 3.6:** Sequences of the various hairpin libraries used in this study. Hairpin stems are annotated in blue, tetraloops in red and the start codon in green.

| Library name | Library sequence | MFE in kcal/mol |
|---|---|---|
| HL1 | AGTATCAACAAAAAgaattgtgagcgctWaHaattAttWWYgtYRagDAGAAAAACCCCAAATATG | -8.0 |
| HL2 | AGTATCAACAAAAAgaattgtgagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATG | -20.2 |
| HL3 | AGTATCAACAAAAAgaattgtgagABWtAatcagagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATG | -23.4 |
| HL4 | AGTATCAACAAAAAgaattgtgagWSWtWatcagagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATG | -25.8 |
| HL5 | AGTATCAACAAAAAgaattgtgagMSMtMatcagagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATG | -28.8 |
| HL6 | AGTATCAACAAAAAgaattgtgagYSYtYatcagagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATG | -32.2 |
| HR1 | AGTATCAACAAAAAgaattgtgagABWtAatcagagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATG | -32.2 |
| HR2 | AGTATCAACAAAAAgaattgtgagMSMtMatcagagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATG | -28.2 |
| HR3 | AGTATCAACAAAAAgaattgtgagYSYtYatcagagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATG | -23.4 |
| HT1 | AGTATCAACAAAAAgaattgtgagYSYtYttcggagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATG | -32.5 |
| HT2 | AGTATCAACAAAAAgaattgtgagCGCtCttcggagcgctYaYaattHttWWYgtYRagSAGAAAAACCCCAAATATG | -33.2 |
| HT3 | AGTATCAACAAAAAgaattgtgagYSYtYttcggagcgctYaYaattHttWWYgtYRagSAGAAAAACCCCAAATATG | -26.5 |
| HT4 | AGTATCAACAAAAAgaattgtgagCGCtCttcggagcgctWaHaattAttWWYgtYRagDAGAAAAACCCCAAATATG | -24.8 |
| HG1 | AGTATCAACAAAAAYttaaYaYtgWSWtWatcagAGCGcagtgttaagttTTYgtYGagYAGAAAAACCCCAAATATG | -28.8 |
| HC1 | tctttaagaattgtgagYgYtYatcagagHgYtYaYaattYttaaRgaAGATCTATG | -28.9 |

**Table 3.7:** Primers used in this study. Capital letters indicate annealing sequence and locations of nucleotide degeneracies. W = A or T; H = A, C or T; Y = C or T; R = A or G; D = A, G or T, V = A, C or G; S = G or C, K = G or T

| Name | Sequence | Direction | Used for library |
|---|---|---|---|
| TW318 | `ttttggtctcaagcgctWaHaattAttWWYgtYRagDAGAAAAACCCCAAATATGGTTTC` | Fw | HL1 |
| TW317 | `ttttggtctcaagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATGGTTTC` | Fw | HL2, HL3, HL4, HL5, HL6 |
| TW411 | `ttttggtctcaagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATGGT` | Fw | HR1, HR2, HR3 |
| TW323 | `ttttggtctcacgctcacaattcTTTTTGTTGATACTTTTATTACATTTG` | Rev | HL1, HL2 |
| TW394 | `ttttggtctcacgctctgatTaWVTctcacaattcTTTTTGTTGATACTTTTATTACATT` | Rev | HL3, HR3 |
| TW392 | `ttttggtctcacgctctgatWaWSWctcacaattcTTTTTGTTGATACTTTTATTACATT` | Rev | HL4 |
| TW393 | `ttttggtctcacgctctgatKaKSKctcacaattcTTTTTGTTGATACTTTTATTACATT` | Rev | HL5, HR2 |
| TW438 | `ttttggtctcacgctctgatRaRSRctcacaattcTTTTTGTTGATACTTTTATTACATT` | Rev | HL6, HR1 |
| TW317 | `ttttggtctcaagcgctYaYaattYttTTYgtYGagYAGAAAAACCCCAAATATGGTTTC` | Fw | HT1 |
| TW318 | `ttttggtctcaagcgctWaHaattAttWWYgtYRagDAGAAAAACCCCAAATATGGTTTC` | Fw | HT4 |
| TW319 | `ttttggtctcaagcgctYaYaattHttWWYgtYRagSAGAAAAACCCCAAATATGGTTTC` | Fw | HT2, HT3 |
| TW329 | `ttttggtctcacgctccgaaRaRSRctcacaattcTTTTTGTTGATACTTTTATTACATT` | Rev | HT1, HT3 |
| TW330 | `ttttggtctcacgctccgaaGaGCGctcacaattcTTTTTGTTGATACTTTTATTACATT` | Rev | HT2, HT4 |
| TW439 | `ttttggtctcaagcgcagtgttaagttTTYgtYGagYAGAAAAACCCCAAATATGGT` | Fw | HG1 Gal repression lib. |
| TW440 | `ttttggtctcaCGCTctgatWaWSWcaRtRttaaRTTTTTGTTGATACTTTTATTACATT` | Rev | HG1 Gal repression lib. |
| TW431 | `ttttggtctcaatcagagHgYtYaYaattYttaaRgaAGATCTATGGTTTCTAAAGGTGA` | Fw | All constitutive libraries |
| TW432 | `ttttggtctcatgatRaRcRctcacaattcttaaagaTTTGTTTGTTTATGTGTGTTTAT` | Rev | HC1-TDH3 |
| TW434 | `ttttggtctcatgatRaRcRctcacaattcttaaagaTGTTTTATATTTGTTGTAAAAAG` | Rev | HC1-PGK1 |
| TW433 | `ttttggtctcatgatRaRcRctcacaattcttaaagaGTTTAGTTAATTATAGTTCGTTG` | Rev | HC1-TEF2 |
| TW435 | `ttttggtctcatgatRaRcRctcacaattcttaaagaTTAGCAGATGTAGGTATTTTCTT` | Rev | HC1-YRA1s |
| TW437 | `ttttggtctcatgatRaRcRctcacaattcttaaagaTTTGATTTGCTTTTATCTTTTTT` | Rev | HC1-POP6 |

### 3.3.3 Library testing and analysis

Libraries were tested using flow cytometry, as described in Materials and Methods **subsection 2.2.3** on page 67. yeGFP and mRuby2 were selected as the two strongest fluorescent proteins currently available[210]. *S. cerevisiae* BY4741 was the strain used for the implementation of the libraries. Cloning was done using the Yeast ToolKit, as described in **section 2.2** on page 62.

## 3.4 Results

### 3.4.1 Matching RNA structure to fluorescence levels

One of the central concepts in synthetic biology is forward engineering, which constitutes a design process where a system is built from the ground up using a high level understanding (model) of the relevant factors in a system. In contrast, much of the traditional work in metabolic engineering and cell biology has been conducted using a reverse engineering approach. In this approach, genetic circuits that are the result of the process of evolution are adapted, combined and deconstructed to learn about their design and to create desired functionality. Forward engineering ultimately provides a more powerful framework for the design of desired functionality, since the creator is not limited to the designs created in evolution or, more specifically, the known and understood subset of designs created in evolution. In the case of biology, some degree of reverse engineering tends to be unavoidable, as our current level of understanding is typically not sufficient for a forward design approach of the entire system.

For the proposed system developed in this chapter, the biggest challenge for forward engineering was the question of how the strength of the modelled RNA structure relates to the reduction in translation and gene expression. We set out to first determine this transfer function by designing a test hairpin library, characterising it and then selecting colonies that had a wide range of expression levels. From the selected colonies we sequenced the 5'UTR regions of each construct and these sequences were then analysed to determine their predicted folding strength using RNAfold.

When plotting the free energy of folding ($\Delta G$), which is a quantitative measure for the strength of the hairpin structure, against the normalised green fluorescence, a clear pattern arises. As hypothesised, there is a steep decline in expression levels as the $\Delta G$ approaches more negative values. This is shown in **Figure 3.5** on the next page.

However, this decline does not start until the $\Delta G$ reaches approximately -22 kcal/mol. This indicates that the structures with energies less than -22 kcal/mol are not strong enough to interfere with the biological processes that constitute translation. A $\Delta G$ of -22 kcal/mol is strong enough to form stable structures at 30°C, so a more sophisticated explanation is required. It is likely that the RNA helicase activity of constituents of the translation initiation complex play an important role in this.

The relation between folding energy and expression level can be modelled in at least two ways. The most basic function is a generalised exponential shown below:

$$Fl_n(\Delta G) = c \cdot e^{k\Delta G}$$

Where $Fl_n$ is the normalised fluorescence, $\Delta G$ is the free energy contribution of RNA folding (hairpin strength), $c$ is a constant and $k$ is the slope of the curve. The equation for the curve fitted to the GAL promoter is as follows:

$$Fl_n(\Delta G) = 47874 \cdot e^{0.25\Delta G}$$

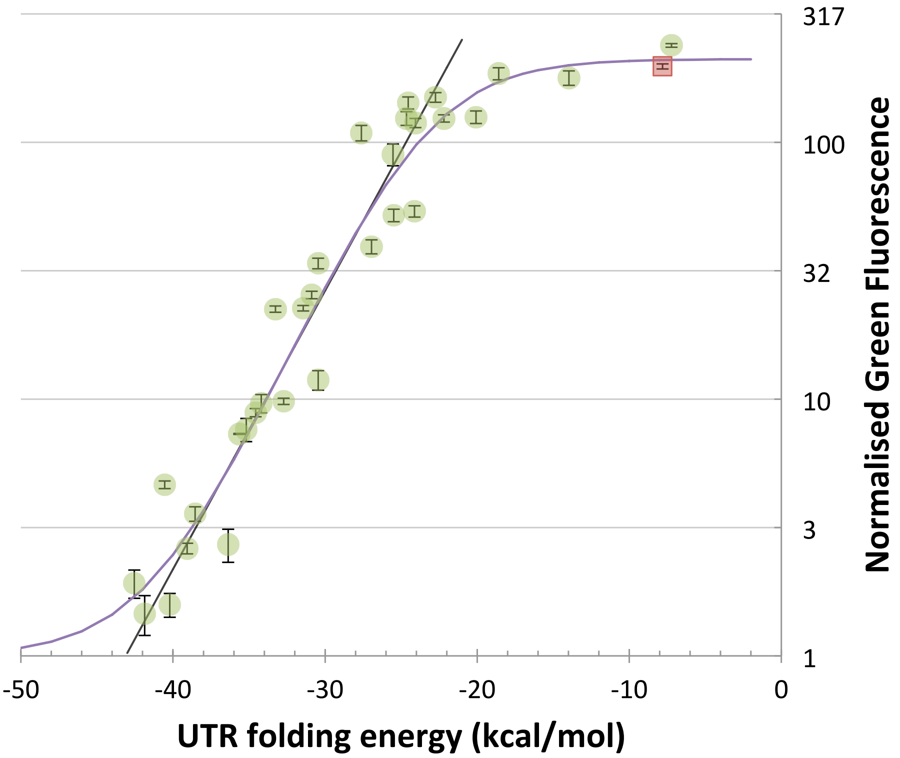


**Figure 3.5:** Relationship between the 5'UTR structure (folding energy) and the normalised green fluorescence from yeast cells. Green fluorescence was measured for 31 isolated library members. The members were sequenced to obtain the 5'UTR sequences and ultimately the corresponding folding energy. Straight black line: exponential fit ($Fl_n(\Delta G) = 47874 \cdot e^{0.25\Delta G}$), curved purple line: logistic growth fit ($Fl_n(\Delta G) = 1 + \frac{210}{1+e^{-0.3(\Delta G-23.5)}}$). The red box shows the properties of the original construct that the library was derived from. Fluorescence is normalised against autofluorescence of the parental strain. A value of 1 indicates autofluorescence levels.

This curve intersects the maximum fluorescence (indicated by the red square in **Figure 3.5**) at a $\Delta G$ of -22.0 kcal/mol and the lower boundary of 1 (signifying cellular autofluorescence) at -43.1 kcal/mol. It is however somewhat crude to only consider this equation valid between these points. A more elegant solution is to fit the equation for logistic growth. The general equation for logistic growth is as follows:

$$Fl_n(\Delta G) = 1 + \frac{P_{max}}{1 + e^{-k(\Delta G - x_0)}}$$

Where $Fl_n$ is the normalised fluorescence, $\Delta G$ is the free energy contribution of RNA folding (hairpin strength), $P_{max}$ is the maximum promoter strength, $k$ is the slope of the exponential part of the curve and $x_0$ is the location of the midpoint of the curve. The equation for the curve fitted to the GAL promoter, the equation is as follows:

$$Fl_n(\Delta G) = 1 + \frac{210}{1 + e^{-0.3(\Delta G - 23.5)}}$$

The advantage of this expression is that the fit can easily be adapted to different promoters by changing the term for the maximum value $P_{max}$. The obtained fit could then be used to predict the expression of specific hairpin containing constructs and whole libraries of such constructs.

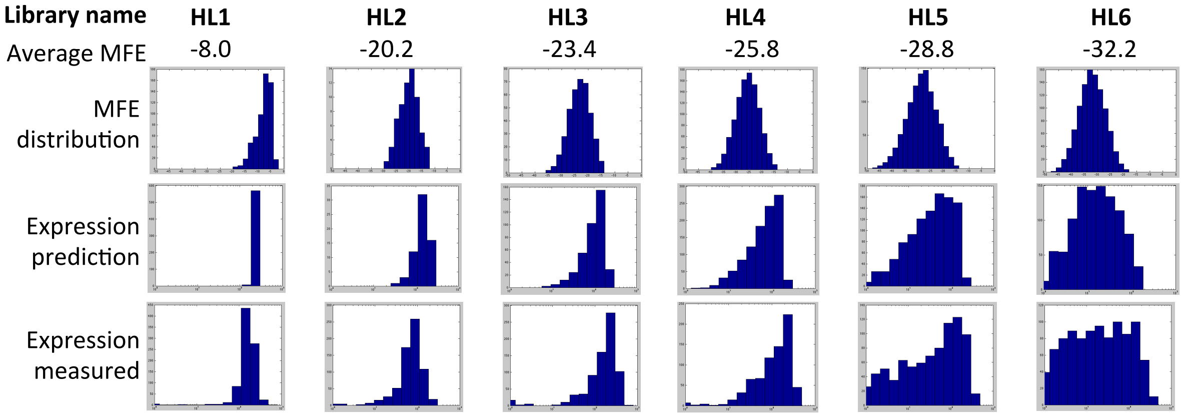


**Figure 3.6:** Correlation between the predicted strength of 5’UTR RNA structures in libraries and the measured gene expression distributions of these libraries. A total of 6 libraries are shown, whose average minimum free energy (MFE) of folding is given in kcal/mol. In the top row of panels, the histogram of the distribution of the MFE in the 5’UTRs of the different libraries is shown. The horizontal axes for these panels ranges on a linear scale from -50 kcal/mol on the left to 0 kcal/mol on the right. The middle row converts these into a histogram of predicted expression levels using logistic fit. The third row shows the distribution of fluorescent reporter gene expression levels in yeast cells as measured by flow cytometry. In the lower two rows, the horizontal axis corresponds to normalised green fluorescence ranging on a logarithmic scale from 1 on the left to 1000 on the right.

### 3.4.2 Library design

Library design can be challenging when the behaviour of individual library members is not known. The logistic correlation between the 5’UTR structure free energy and the gene expression strength can be employed to inform library design. To test whether the logistic equation indeed has predictive power, we designed, built and tested libraries with distributions of increasing average structure strengths (i.e decreasing minimal free energy). These libraries were named HL1 for the weakest distribution at an average MFE of -8.0 kcal/mol, through to HL6 for the strongest distribution at an average MFE of -32.2 kcal/mol. The sequence used to generate these libraries is given in **Table 3.6** on page 88.

For each library the minimal folding energies (MFE) of all possible members were calculated. These individual energies were converted into predicted gene expression levels using the logistic fit to the above equation. The predicted distributions were then compared to the experimentally obtained distributions for green fluorescent protein gene expression in yeast cells. An overview of these comparisons is shown in **Figure 3.6**.

Qualitatively there is a good agreement between the prediction and the experimental results. Experimental results show some peak broadening compared to the predicted results. This becomes apparent when the predicted expression is particularly uniform, such as in the low structure library. For example, compare the predicted versus measured expression for the left-most library in **Figure 3.6**. Here, we expect that the experimental distribution is widened because small and large cells deviate from the predicted average, despite having the same relative expression. This follows from the inherent stochasticity in gene expression.

Quantitatively the largest deviations arise from a shift of the maximum. In the weak structure libraries the predicted peak lies at a higher fluorescence than the experimental values. This may be caused by the altered sequences having a slight negative impact on the promoter, i.e. causing a slightly lower transcription rate. In contrast, stronger structure libraries show a predicted peak at a lower fluorescence than the peak in the experimental values. The reason for this may be a sub-optimal fit in the logistic curve. **Figure 3.5** on page 92 shows that the number of data points between -20 and 0 kcal/mol is relatively low compared to the number above -20 kcal/mol. The fit in this region may therefore not be as accurate as the remaining region. This may have caused the predicted expression peak to occur at slightly lower MFE values than the measured data.

For the goal of forward engineering in synthetic biology, these libraries and predictions positively serve their purpose. Use cases are more likely to be qualitative than quantitative, as it is challenging to account for every member of a large library in a computational model. Rather, the predicted distribution will likely be used to qualitatively match a target distribution which will be a uniform distribution in many cases. If the optimum promoter strength is unknown, a uniform distribution will give the most efficient way of sampling the expression level space.

Frequently this uniform distribution can be matched most accurately by a broad Gaussian curve as shown in the right most column in **Figure 3.6** on the previous page. Although different combinations of strong and weak libraries can be used to create a truly uniform predicted library, the results show that a broad Gaussian is flattened in the experimental results to provide a uniform sampling of expression levels. This simplifies library design considerably and makes this an attractive method for promoter strength screening.

### 3.4.3 Tetraloops

As described in the chapter introduction, tetraloops are 4 base pair sequences that form at the U-turn of the hairpin structure. Their special properties stabilise the hairpin structure allowing the structure to fold more easily. The thermodynamic contribution of tetraloops has widely been reported to be between -1 and -4 kcal/mol depending on the type of tetraloop and its surrounding sequence[175,176,185]. Addition of tetraloop sequences to the libraries offers potential for additional regulation through the binding of proteins to the mRNA. However, tetraloops will need to be characterised carefully to allow their effects to be incorporated into the predictive model.

**Figure 3.7** on the following page shows how the addition of the tetraloop affects the expression distribution. Library HL6 (columns 1) and library HT1 (column 2) show an identical library with and without the UUCG tetraloop sequence. Despite virtually no difference in the predicted folding energies and predicted expression distributions, there is a marked difference in the measured expression profile. The expression profile for the library with tetraloop has shifted mostly to the low end of the spectrum - only slightly higher than the autofluorescence levels of yeast cells.

Having established the presence of a strong effect of inclusion of the tetraloop sequence, we look at the thermodynamic contribution of the loops in order to incorporate the effect into the predictive model. Comparing the experimental expression profiles in the HL6 (1st column) and HT3 (5th column) libraries, it can be argued that the library with a mean minimum free energy

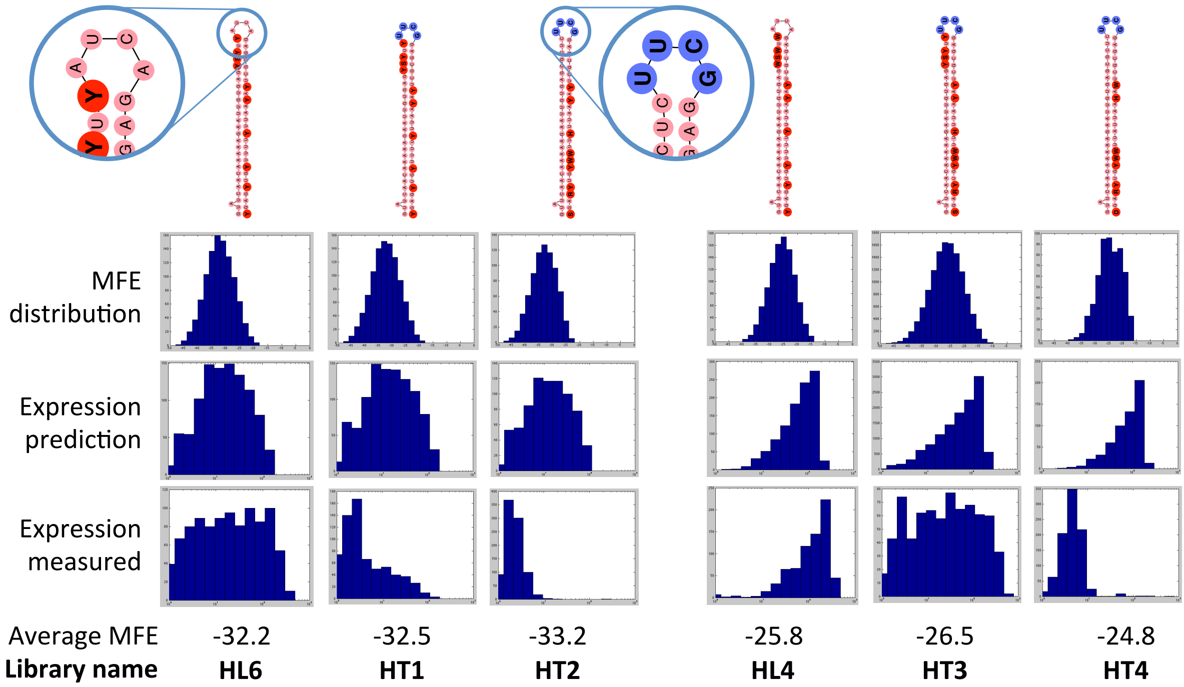


**Figure 3.7:** Two sets of libraries characterising the effect of the highly stable tetraloop on fluorescence levels. Left set (first three columns) each have MFE averages of -32 to -33 kcal/mol. The right set has MFE averages of -24.5 to -26.5 kcal/mol, as indicated below the histograms. The illustrations at the top show the scaffold hairpin and introduced degeneracies for each library. The blue bases in the RNA structures indicate the presence of the strong UUCG tetraloop. The top row of histograms shows the MFE distributions for the 5'UTRs of each of the 6 libraries. The horizontal axes for these panels ranges on a linear scale from -50 kcal/mol on the left to 0 kcal/mol on the right. The middle row converts these into a distribution of predicted expression levels. Expression profiles are predicted by mapping the structure MFE to an associated expression level using the function derived from **Figure 3.4.1** on page 91. The third row shows the distribution of fluorescence levels in yeast cells as measured by flow cytometry. In the lower two rows, the horizontal axis corresponds to normalised green fluorescence ranging on a logarithmic scale from 1 on the left to 1000 on the right. Note that measured expression levels deviate substantially from the predictions for libraries that contain a tetraloop.

of -32 kcal/mol without the tetraloop sequence is experimentally equivalent to the library with a mean MFE of -26 kcal/mol with the tetraloop. This would indicate a net average thermodynamic contribution of -6 kcal/mol for the UUCG tetraloop.

However when the HT3 library (column 5) is compared to HT4 (column 6), it becomes apparent that the net contribution can not be expressed as a simple average. The two libraries which both have the UUCG tetraloop have very similar mean MFEs, yet distinctly different expression profiles. Looking at the locations of the nucleotide degeneracies reveals that the context of the tetraloop is important for its effect on expression. HT4 has no diversity in the strong 7 bp stem that is connected to the tetraloop. As a result, the expression from the members of this library is reduced to levels nearing autofluorescence. This indicates an apparent thermodynamic contribution between 0 and -23 kcal/mol, assuming repression is complete at -44 kcal/mol.

We conclude that repression is almost complete when the UUCG tetraloop is combined with a strong stem. This would also explain the bimodal distribution of expression results in the second library that is shown. In the context of library creation this is problematic, as it does not facilitate the creation of gradual libraries with many intermediate expression strengths. Instead, the response becomes all-or-nothing, which defeats the purpose of generating a library.

In summary, the apparent thermodynamic contribution in our system ranges from 0 to -20 kcal/mol or more, with a likely minimum of -6 kcal/mol. The discrepancy between these observations and the reported measurements of -1 to -4 kcal/mol is implausibly large, even when taking into account that past reported measurements have typically been carried out *in vitro* at 37°C, versus the *in vivo* at 30°C conditions used here. It is more likely that the tetraloop-stabilised hairpin poses a distinct challenge to the RNA helicase machinery in the process of scanning ribosome assembly. This would cause a more severe impact on translation efficiency than the calculated thermodynamics would suggest.

### 3.4.4 Effect of 5'UTR structure on the transcription process

Instinctively, it would be expected that the repressive effect of a hairpin in the 5'UTR region of an mRNA is caused by the formed structure interfering with translation, presumably by reducing the efficiency of translation initiation as the ribosome scans the 5'UTR region for an AUG initiation codon. However, it is also a possibility that the repressive effect is partly or wholly exerted at the stage of transcription. DNA is naturally a duplex and G-T (G-U) pairs cannot exist in DNA, so formation of the designed hairpins in DNA is unlikely. There can still be an effect on the transcription machinery, as the mRNA exits the RNA-polymerase and starts to form a hairpin, possibly interfering with the process, for example by leading to premature transcription termination. In addition, mRNA degradation through the no-go decay pathway (see **subsection 1.5.4** on page 45) or nuclear export of the mRNA (see **subsection 1.4.2** on page 38) could also be affected.

To rule out with direct evidence that the mRNA is not degraded, we conducted a set of control experiments using quantitative RT-PCR (qPCR). Measurements were normalised to the TPI1 reference gene, which has been shown to be more robust than the commonly used ACT1 reference, particularly in conditions of growth on alternative carbon sources such as galactose, which we used extensively[211]. As an additional reference, we included the weakly expressed DOA1 gene, which is expressed at an average of 2.6 copies per cell[212]. For more detailed information, including the primer sequences that were used for this experiment, we refer to Materials and Methods **section 2.3** on page 71.

Using this method, mRNA levels per cell were observed for a set of constructs with strong and weak 5'UTR hairpins. The strongest was a hairpin loop with a folding strength of -32 kcal/mol that included a tetraloop for added stability. Next was a strong hairpin with an MFE of -28.4 kcal/mol. Then came an intentionally weakened hairpin with a folding strength of -4.16 kcal/mol and finally the original, unmodified promoter.

If hairpins were to interfere with transcription, mRNA levels should drop for constructs with strong hairpins. However, **Figure 3.8** on the following page shows this is not the case. The mRNA levels per cell for the different constructs remain similar and do not decrease. If anything

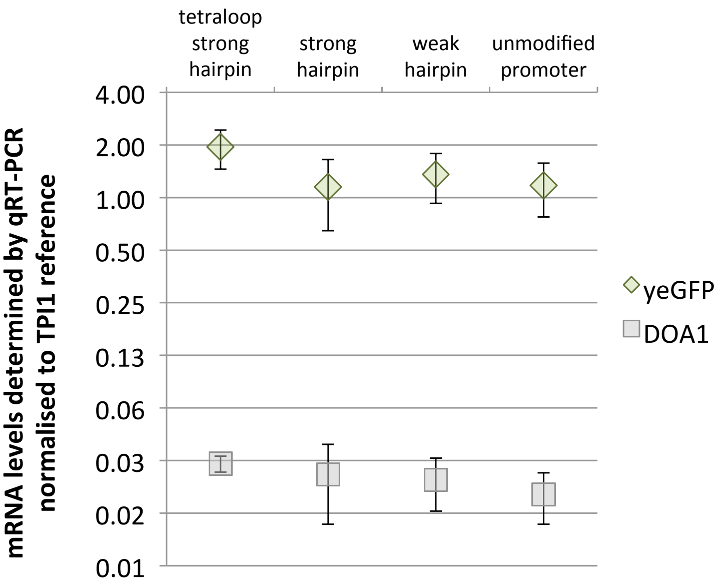


**Figure 3.8:** Low impact of 5'UTR structure on transcript levels per cell. Two-step qPCR was used to determine transcript levels in strains with hairpins of various strengths. The strong tetraloop hairpin has a folding strength of -32 kcal/mol, the strong hairpin has an MFE of -28.4 kcal/mol and the weak hairpin an MFE of -4.16 kcal/mol. Hairpins are present in the 5'UTR of yeGFP expressed from a pLX GAL1-derived promoter. Transcript levels are normalised against the TPI1 mRNA, using the dd-Ct method[158]. TPI1 is a strongly expressed gene (~200 mRNAs per cell), while DOA1 is a weakly expressed reference gene (2.6 mRNAs per cell on average). Error bars indicate standard deviation of technical triplicates.

there are slightly higher transcript levels for the mRNA construct with the strongest hairpin (the tetraloop strong hairpin). Thus, any effect imparted by the designed hairpins is not leading to fewer mRNAs. The effect on expression from the 5'UTR hairpins must therefore occur after transcription, and all gathered evidence indicates that it is at the stage of translation.

### 3.4.5 Modularity with different promoters

Having demonstrated that non-tetraloop structures designed into the 5'UTR of a yeGFP-encoding mRNA can predictably down-tune expression, we next sought to demonstrate that this approach is modular, i.e. that the technique is generally applicable to be used in constructs with other modular DNA parts, such as differet promoters. To demonstrate this we introduced a 5'UTR haipin library into 5 constructs containing different constitutive promoters of different strengths, using these in place of the previously-used pGAL1 promoter. The same hairpin library (named HC1) was used in all cases and was chosen to have an average minimum free energy of -28.9 kcal/mol. This is expected to produce a uniform distribution between autofluorescence levels and the maximum expression level of the various promoters. We expect this to be the most common use case for the technique.

The experimental data shown in Figure 3.9 on the next page demonstrates that the library does indeed produce a fluorescence distribution in all selected promoters that is consistent with expectations from the predicted MFE. While this is only a small sample of possible promoters, this result implies that the 5'UTR library approach is modular with respect to upstream promoters

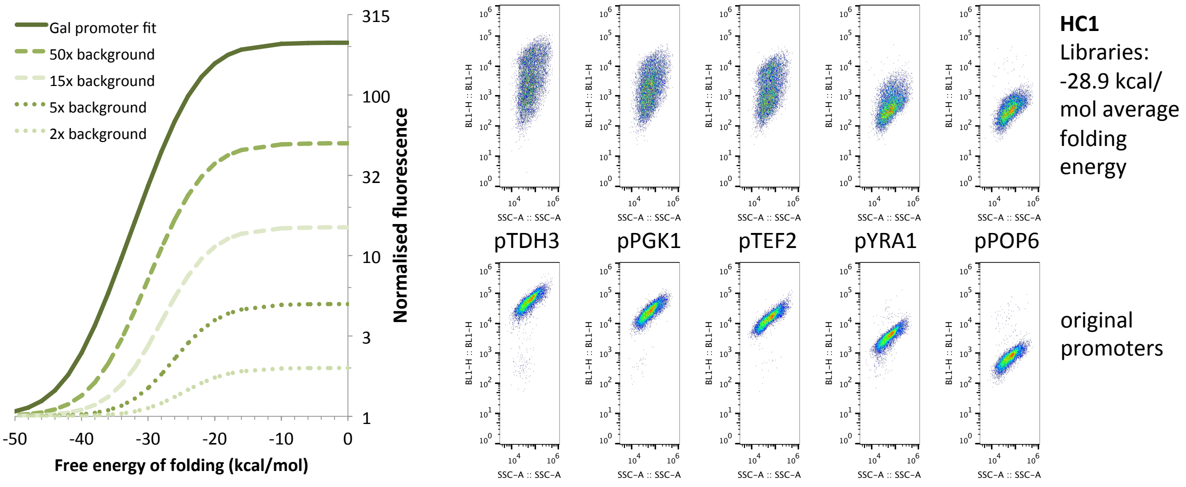


**Figure 3.9:** Library performance at various promoter strengths. **Left panel:** predicted strength distribution for promoters of lower strength than the GAL1 promoter, based on the maximum expression of a promoter (50, 15, 5 and 2 fold increase over background fluorescence). **Right set of panels:** The HC1 library with -28.9 kcal/mol average MFE hairpins in the 5'UTR for 5 different constitutive promoters of decreasing strength, shown in the top row. For comparison, the promoters with a control 5'UTR with no structure are shown in the bottom row. Panels are pseudocolour density plots showing 10,000 events each. Axes are sidescatter (SSC-A) versus green fluorescence (BL1-H) in arbitrary units.

(i.e. there is no context-dependency). This means that our approach could likely be applied to alter the expression from most, if not all Pol II promoters in yeast. However, when the results are inspected with more detail, it becomes apparent that the expression level distribution becomes progressively less uniform as weaker promoters are used. For these promoters, the distribution skews slightly towards the lower end of the spectrum. This phenomenon can be clarified using the logistic fit derived earlier.

In the first panel of Figure 3.9, the curves show the predicted fits for promoters with a different maximum expression value for each promoter used to create the fit. When comparing these, it becomes evident that the range over which they are exponential (manifesting as the straight part of the curve when plotted logarithmically) becomes narrower for weaker promoters. For example, if the maximum promoter strength is only 5 times as high as the autofluorescence of the cell, then virtually the full range of expression levels is captured between -34 and -20 kcal/mol. This contrasts with the original curve, where a range of -44 to -24 kcal/mol MFE is needed to capture the full range of expression levels.

Thus, when a library spanning the range of -44 to -24 kcal/mol is used for a weaker promoter as was the case in this experiment, a larger proportion of the members will obtain a sequence that is fully-repressing which in turn skews the expression level distribution towards the lower end of the spectrum. By measuring the normalised fluorescence of a promoter prior to library creation, the acquired information can be used to tailor the library MFE spread to the promoter strength, and hence create with more precision the required distribution of expression levels.

### 3.4.6 Modularity with different open reading frames

In addition to demonstrating modularity with respect to the upstream promoter, we also set out to demonstrate that this technique is also robust to changes in the downstream sequence, i.e. to changes in the open reading frame (ORF) sequence of the mRNA. To investigate this, a selection of hairpin libraries previously assessed with yeGFP were reconstructed with the red fluorescent protein mRuby2 ORF substituting for the yeGFP ORF. The green libraries that we selected were HL6, HL5 and HL3, and the corresponding red libraries are HR1, HR2 and HR3, respectively. The sequences used to generate these libraries are given in **Table 3.6** on page 88. Sequence identity between the two ORFs encoding these fluorescent proteins is as little as 11.5%, as identified by a BLAST alignment optimised for somewhat similar sequences (blastn)[213].

The synthetic constructs were tested in yeast and flow cytometry was used to obtain red fluorescent expression values. A direct comparison of these results, and the previous results with yeGFP is shown in **Figure 3.10** on the following page and reveals that the results obtained for each hairpin library are matched closely for both the green and red fluorescent protein coding genes. This indicates that the method is modular and that it can be used with different ORFs that produce different proteins. In other words, it is sufficiently robust towards changes in the sequence downstream from the hairpin.

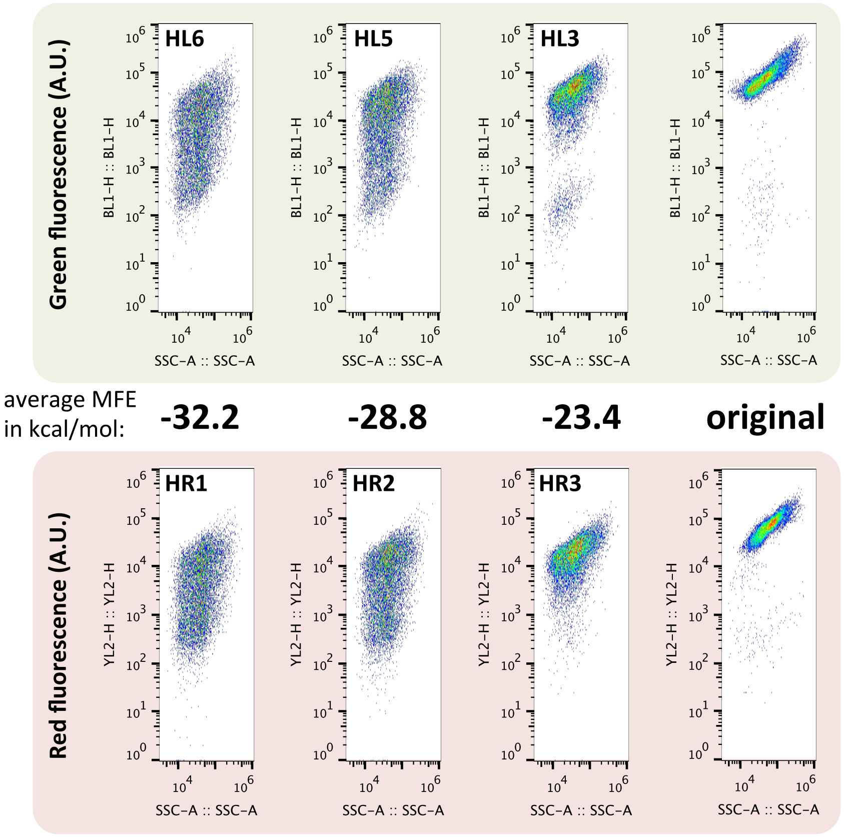


**Figure 3.10:** Comparison of three 5'UTR libraries of decreasing hairpin strength plus a non-structured control, all expressed from the GAL1 promoter, and cloned with either a green (yeGFP, top panel) or red (mRuby2, bottom panel) reporter gene. Panels are pseudocolour density plots showing 10,000 events each. Axes are sidescatter (SSC-A) versus green (BL1-H) or red (YL2-H) fluorescence in arbitrary units (A.U.). For each of the three libraries the average minimum free energy of folding (MFE) of all potential library members is shown in kcal/mol. A lower average MFE indicates stronger folding of the hairpins and consequently stronger repression.

### 3.4.7 Applications of gene expression tuning with 5'UTR hairpins

Fine-tuning gene expression is important in many aspects of synthetic biology, and the approach described here of using hairpin sequences in the mRNA 5'UTRs offers a new tool towards this in yeast and potentially in other eukaryotes. However, many genetic parts already exist that allow gene expression levels to be modulated. In particular, promoter libraries enable users to alter the amount of transcription and thus expression of any gene. Normally this is utilised with sets of different constitutive promoters which can be exchanged to change gene expression outputs[8].

For regulated promoters (e.g. inducible promoters that have on/off responses to external inputs), changing the output gene expression requires libraries of those regulated promoters to be made. This has previously been achieved by mutation of the regulated promoter at non-essential sequences in the promoter core region[43]. These mutations lead to versions of the promoters with different outputs when induced. However, they unfortunately also lead to promoters with different leakiness (i.e. output when not induced) due to mutation of bases close to transcription factor binding sites altering the efficiency with which they control regulation.

Importantly, leakiness of expression plays a major role in the implementation of genetic circuits. Typically, leakiness needs to be as low as possible for the successful implementation of designed circuits with predicted behaviours. Leakiness of the identified library members can vary considerably in the aforementioned approach of traditional random or targeted mutagenesis of promoters and core promoters. This hampers the ability of engineers designing and constructing genetic circuits to fine-tune expression levels within a system while also guaranteeing precise regulation.

The 5'UTR library approach developed here is promising solution to this problem, because the fine tuning of gene expression output is achieved by mutation away from those sequences that encode regulation. It therefore potentially offers a new way to modify protein levels from synthetic constructs without affecting the characteristics of their regulation at the promoter level.

To show that this approach indeed imparts inherently better leakiness characteristics compared to traditional methods, we directly compared GAL1-based promoter libraries created using our method versus current-generation mutagenesis-based library methods. The method we compare against, which I have described in detail in Methods in Molecular Biology[214], relies on targeted mutagenesis of the core promoter region of a high-strength promoter. For the mutagenesis library, promoters previously made by Ellis et al.[43] were used where the GAL1 promoter has an integrated Lac operator site (LacO) which can be repressed by the binding of the Lac Inhibitor protein (LacI) which is expressed elsewhere in the yeast genome. The promoters within the library vary in output strength due to mutations in their core region. In galactose media, they are turned on and with the addition of IPTG (**section 1.3.2** on page 31). In our 5'UTR method, we paired the strongest member of the Lac-repressed GAL1 promoter library (LX) with a library of 5'UTR hairpins with -28.8 kcal/mol average folding strength (named HG1). A total of 48 transformants were picked and three were discarded for aberrant expression of red fluorescent protein. The remaining 45 clones were compared to the library constructed with targeted mutagenesis. In both cases all constructs contained the yeGFP ORF and were assessed by the green fluorescence of yeast cells grown in galactose media with and without IPTG.

The experimental results comparing these two different approaches are shown in **Figure 3.11** on the following page. The top panel shows the design and the results of the targeted promoter mutagenesis library approach. The induced states of the library members give a graded range of yeGFP expression outputs useful for construction of synthetic gene circuits and pathways. However, the repressed state of many of the members shows considerable variation, with L3, L4, L5, L6, L9 and L15 showing a significant increase in leakiness compared to the original LX promoter. The results for constructs built exploiting the 5'UTR library method also show an excellent graded range of output expressions, but in contrast to the promoter library method, these also now show very little variation in the expression in the repressed state, all now showing around the same low leakiness as the parent promoter (LX).

Thus this new method for expression library creation outperforms the current approach of promoter library creation as all members exhibit essentially the same level of regulation, with no unexpected leakiness. The increased predictability of this approach can also aid *a priori* modelling efforts, because there is significantly lower risk that the promoter with the required maximum expression also exhibits increased leakiness or has unanticipated impaired regulation.

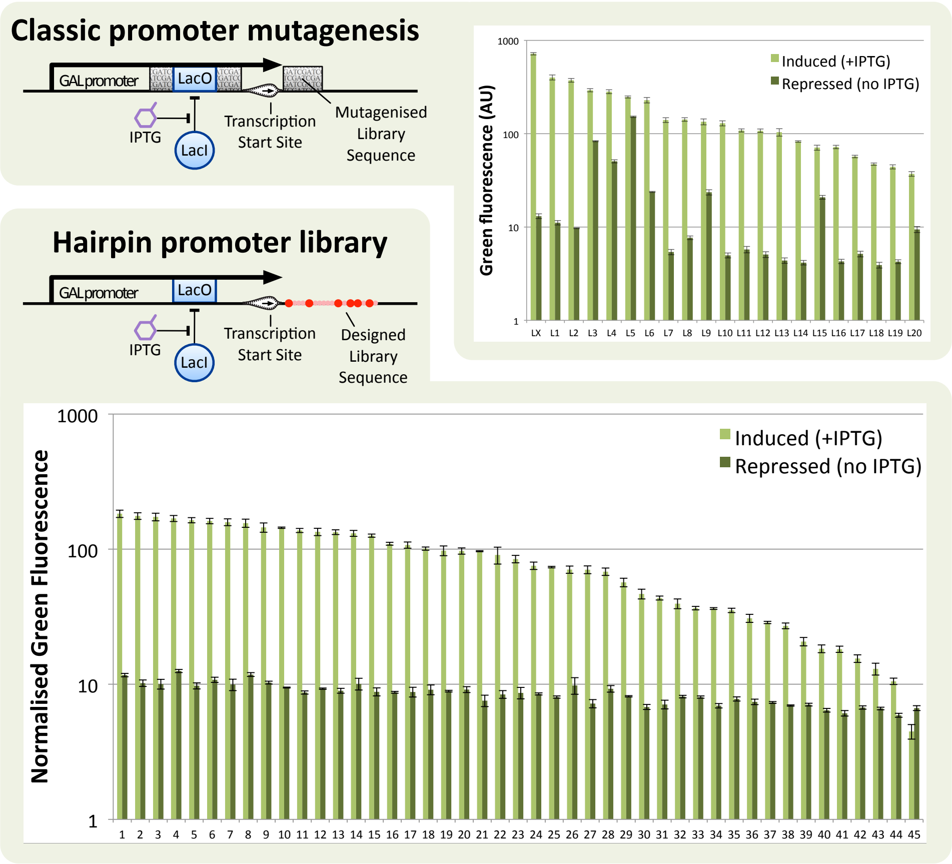


**Figure 3.11:** Comparison of methods for regulated expression library creation. Top: results and method employing random mutagenesis of selected regions in the core promoter region. In this approach, selected regions in the core promoter (grey blocks) are completely randomised and tested. From a pool of 350 candidates, the 20 best performing hits (L1-L20) and the non-mutated version (LX) are selected. Bottom: results and method employing 5'UTR hairpins as developed in this work. In this method, a hairpin library sequence (HG1) is placed directly following the transcription start site and transformed into yeast. From the resulting clones, 45 are directly picked and characterised. Promoters contain a Lac operator site which can be bound by the Lac Repressor (LacI) in order to repress the promoter. IPTG can subsequently be added to release LacI from the DNA, reversing its repressive effect and inducing yEGFP expression and generating green fluorescence.

Another, possibly more significant advantage of this technique, is the fact that far fewer colonies had to be screened to find this 45 member graded output library. In the promoter mutagenesis method, approximately 350 colonies had to be screened in order to isolate the 20-member graded range library. For the 5'UTR hairpin method, only 48 colonies from the designed library were screened and sequenced to create the 45 member library shown here (3 were discarded due to DNA assembly errors). In addition, no modified nucleotides or special polymerases are needed to generate the library, which can be constructed and screened in less than a week following delivery of the required oligonucleotides.

## 3.5 Discussion

In this chapter we set out to advance gene expression control tools for synthetic biology in yeast, by developing a method akin to that seen with RBS sequence design in bacteria. In *E. coli*, synthetic biologists routinely predict and adjust gene expression strength using tools such as the RBS Calculator. However, in *S. cerevisiae* such tools do not currently exist. Here, we successfully implemented a system that allows predictable tuning of expression strength by introducing a strong hairpin into the 5'UTR of an arbitrary gene. Not only is this new method more predictable than promoter-based techniques to vary gene expression output, it is also considerably faster and simpler to perform. The approach developed here will contribute to making yeast gene engineering more predictable, and in turn facilitate improved forward engineering of genetic circuits and advance the field of synthetic biology.

### 3.5.1 Connecting mRNA structure with expression strength

In order to allow prediction of expression we first needed to determine the relationship between the strength of the hairpin in the 5'UTR and its effect on expression. We showed in **Figure 3.5** on page 92 that a logistic curve offers a good fit between the strength of the hairpin measured in kcal/mol and expression normalised to autofluorescence levels.

The shape of this curve, which plateaus between -24 and 0 kcal/mol, is interesting and suggests a mechanistic basis for the reduction of translation that is not purely the result of structure in the mRNA, but an interplay between the structure and components of the translation initiation machinery.

As described in the introduction (see **section 1.5** on page 41), eIF4A is the primary translation initiation factor conferring RNA helicase activity within the eIF4F translation initiation complex. From the work of this chapter, we hypothesise that eIF4A facilitates the resolution of RNA structures with $\Delta G$s between 0 and -24 kcal/mol. eIF4A has been classed as non-processive helicases and can therefore not catalyze consecutive reactions without releasing its substrate. This can explain why it is unable to completely resolve stronger RNA structures in the way that DNA helicases can completely resolve the double helix in DNA replication.

The hypothesis that RNA helicases play a role in shaping the correlation between RNA structure and expression strength is also supported by the slope of the curve between -24 and -44 kcal/mol. If RNA structures of this strength could not be resolved at all, one would expect a 10-fold decrease in expression strength for every -1.36 kcal/mol added to the strength of the hairpin, since every -1.36 kcal/mol step increases the concentration of the folded fraction by 10-fold[215]. We find that in reality a difference of 10 kcal/mol is necessary to impart a 10-fold difference in expression. This lends further credibility to the hypothesis that RNA helicase activity (partially) counteracts the effects of hairpin structure in mRNA.

In this chapter, the established logistic fit between the mRNA structure and expression output was then used to create a series of libraries with different average hairpin strengths and predicted expression profiles. When we measured the expression levels of these designed libraries, the results matched predictions well, as shown in **Figure 3.6** on page 93. We found that a library with -32.2 kcal/mol average folding strength produced a library with a near uniform distribution of expression strengths, which is ideal for applications where a wide range of expression strengths needs to be tested.

### 3.5.2 The impact of tetraloops

Tetraloop sequences are evolutionary conserved in ribosomal RNA and other ribozymes, indicating their importance in maintaining secondary structure in RNA. They can also be bound by RNA-binding proteins, offering potential ways for further regulation of the mRNAs. However, for this to be of use, the inclusion of tetraloops must have a predictable impact on expression. We therefore set out to determine how expression levels were affected in libraries with tetraloops.

Indeed, the inclusion of a 4 base-pair tetraloop sequence in the loop of the hairpin dramatically reduced observed expression levels. Unfortunately, the $\Delta G$ contribution of tetraloops is not accounted for in the RNAfold software that we used for RNA structure prediction. This led to marked discrepancies between the predicted and observed expression profiles for libraries including tetraloops, as is shown in **Figure 3.7** on page 95. However, even if RNA structure prediction software included the nominal contributions of tetraloops (reported between -1 and -4 kcal/mol) the predictions would still underestimate the effect on expression, as the apparent $\Delta G$ contribution of tetraloops in our system is at least -6 kcal/mol. We hypothesise that this is an effect of the interaction between the tetraloop and translation initiation factors, the RNA helicase eIF4A in particular. The tetraloop structure resolution by eIF4A could be so problematic that it confers a disproportionate effect on translation compared to its nominal $\Delta G$ contribution.

Overall it seems that the application of the tetraloop in the context of library creation is challenging. It is non-trivial to predict when the tetraloop stabilised hairpin will form using current RNA folding algorithms and when it does it has a dramatic impact on expression. This makes it difficult to create a library of promoters with gradually decreasing expression strength. The potential benefit of additional regulation with RNA binding proteins allowed by tetraloops does not offset the decrease in predictability of expression strength. However, given their dramatic impact, tetraloops should remain a serious possibility when hairpin size is tightly constrained due to limitations in the methods for DNA assembly, for example.

### 3.5.3 The effect of hairpins on mRNA stability

In order to verify that the effect of an RNA hairpin in the 5'UTR of a coding sequence is indeed primarily at the translation level, we measured mRNA levels of constructs with hairpins of various strengths. These were compared to mRNA levels of constructs with no hairpin added.

In **Figure 3.8** on page 97, we show that mRNA levels are not affected by the inclusion of a hairpin. This confirms that the reduction of expression is not caused by a reduction of transcription, which is in line with expectations.

As described in **subsection 1.5.4** on page 45, several mRNA surveillance mechanisms exist in yeast, and one of these, the No-Go decay pathway, targets mRNA transcripts with stalled ribosomes. One may expect that transcripts with 5'UTR hairpins will form substrates for the No-Go decay pathway, since the stalled ribosomes at the hairpin would attract the No-Go decay machinery to cleave the mRNA. However, the results show that mRNA levels were not affected by hairpins, which is surprising, as the No-Go decay pathway was studied using very similar hairpin structures to those that were used here[216]. It is possible that the hairpin location in the 5'UTR plays a role. It is conceivable that the No-Go decay pathway only targets RNA bound to complete ribosomes actively involved in translation, and does not act when the small ribosome subunit is scanning the 5'UTR with the eIF4F complex.

Another possibility is that the RNA is being cleaved at the location of the hairpin, leaving the rest of the mRNA in the cytosol, which then somehow persists long enough to still be measured by qPCR. The qPCR assay we used was only capable of detecting the 3' end of the transcript and so in further experiments it would be interesting to design a construct containing a large enough 5' region upstream of the hairpin to allow separate detection of that section also by qPCR. If the transcript is indeed cleaved *in vivo*, the 5'probe will likely yield dramatically lower transcript levels, since it would not be amplified by a poly(T)-primed reverse-transcription reaction.

### 3.5.4 Robustness of the method

As detialed in the introduction (see **subsection 1.1.4** on page 16), robustness is of major importance in synthetic biology. The behaviour of biological parts can change dramatically depending on their genetic context. In a series of experiments, we showed here that the method we have developed is generally applicable and is not especially sensitive to upstream or downstream genetic context. This will facilitate adoption of the developed method by other researchers in the field as it can be considered a modular genetic part.

In the first set of robustness experiments, we looked at the behaviour of a single hairpin library (average folding energy of -28.9 kcal/mol) with a series of constitutive promoters of different strengths (see **Figure 3.9** on page 98). Consistent with the logistic model that we initially fitted, the expression distributions for libraries based on weaker promoters were somewhat biased towards lower expression. This means that promoters will ideally need to be characterised to determine their maximum expression before they can be partnered with 5'UTR hairpin libraries. However, as we expect that our technique will most commonly be partnered with strong promoters, we expect libraries with an average energy of folding between -32 and -29 kcal/mol to be suitable in most cases.

In the second set of robustness experiments, we looked at the context dependence with respect to the downstream sequence. To do this, libraries with different average folding energies that were initially constructed and tested with yeGFP were recreated but this time with mRuby2 as the fluorescent reporter. From the results shown in **Figure 3.10** on page 100 it became clear that the results were very reproducible even though the two reporters share negligible sequence homology. This gives initial verification that the method is not adversely affected by changes in downstream sequence, but ideally this would be validated by testing with further different downstream ORFs.

The biggest discrepancy between the two sets of libraries in this experiment is in the library with an average folding energy of -23.4 kcal/mol. In the yeGFP library, a small subpopulation is present with expression levels comparable to autofluorescence. This population is not present in the mRuby2 library. It is questionable whether this is a significant difference, since the subpopulation is a particularly small fraction of the total population. Additionally, a similar subpopulation is present in the populations of the control constructs with the same promoters but no hairpin library. This indicates that there may be a fraction of the total population that has not been induced by the galactose present in the medium. Another possibility is that a fraction of the library sequences contain DNA synthesis errors that prevent expression of these constructs, for example by introduction of an out-of-frame start codon. These errors could be propagated differently in the assembly process between the two libraries.

Overall, the characterisation with different parts showed that the system is robust towards upstream sequence variation (different promoters of various strengths) and towards downstream sequence variation as well (two reporters with different encoding sequences). This is preferable to RBS parts widely-used in bacterial synthetic biology, as the performance of these has shown to vary considerably based on changes to the upstream promoter sequence and downstream ORF sequences. The excellent modularity of our system will lower the barrier for third parties to implement it for fast construction of expression libraries in yeast.

### 3.5.5 Performance in a test case

Finally, in order to assess the performance of the method in a synthetic biology context, we compared it to an expression library created using the standard method of promoter mutation. The selected test case was based on the GAL1 promoter with an engineered LacI repression site and particular attention was paid to the performance of each library in the repressed state. In most applications of an inducible system, the repressed state needs to be as low as possible, since leaky expression generally interferes with device, circuit or pathway function. As shown in **Figure 3.11** on page 102, our method showed a decidedly more constant level of strong repression across the library as compared to a targeted mutagenesis method.

Unexpectedly, the level of leaky expression did not decrease with promoter strength. It was anticipated that the hairpin in the 5'UTR would equally diminish expression in both the induced and repressed state. This would cause the measured expression in the repressed state to decrease proportionally with a decrease in maximum expression seen when induced. However, the repressed expression outputs only show a very modest decrease as maximum induced output strengths decrease.

One possible explanation for this observation is that there could be a second, very weak transcription start site, situated downstream of where the hairpin is placed in the construct. However, it is challenging to support this hypothesis with readily available experimental evidence, since all high throughput TSS determination experiments that are available have been carried out in dextrose. Since the GAL1 promoter is tightly repressed in these conditions, we could not verify whether transcription initiation in the GAL1 promoter is focused or diffuse. This hypothesis does allow for the interesting possibility that if a second TSS were to be identified and deactivated, a library could be created with particularly low leakiness across all members of the library. Similarly, an internal ribosome entry site may be present downstream of the hairpin, allowing cap-independent translation to occur and bypassing the inhibitory effect of the hairpin. This also allows for the possibility for this sequence to be deactivated and a new library to be created from the modified version exhibiting low leakiness.

## 3.6 Conclusion

In this chapter, a new approach for controlling the expression levels of a gene was developed for yeast synthetic biology. By computationally designing short hairpin sequences to be inserted into 5'UTR regions, the amount of protein produced from mRNAs in yeast can be predictability tuned. The approach is quick and cheap to implement and can be considered modular as it works independently of upstream and downstream sequences. It brings a whole new genetic part to the yeast synthetic biology toolbox that outperforms previous methods to tune gene expression and enables prediction-based design. It is especially suited for the generation of libraries where proteins are expressed at a wide range of different levels within different cells within a population. We especially anticipate downstream use of this approach for the combinatorial engineering of metabolic pathways that benefit from varied levels of expression of the different enzymes, and for precise tuning of regulator levels within genetic circuits that require digital-like behaviours.

# 4. Transcriptional Interference

In this chapter we investigate the utility of transcriptional interference (TI) in synthetic circuits. Transcriptional interference is the negative (or potentially positive) effect that transcription from one promoter has on the transcription from a second promoter that is on the same region of DNA. Transcriptional interference could enable more efficient biological computation and more robust function by eliminating or reducing the need for the expression of transcription factors and instead allowing some of the regulation to take place *in cis* at the transcriptional level.

Here we choose to implement a bistable switch as a model circuit simplified by using TI. By using the specific architecture of two head-to-head facing promoters, we prove that transcriptional interference is present in our prototype system. While our design is a success at the transcriptional level we are unable to obtain protein expression from the mRNA transcripts of this system. We therefore explore a variety of (co-)transcriptional processes in an attempt to restore translation to enable use of TI. Finally, we seek to apply the generated RNA directly in trans-regulation by using CRISPR/dCas9, avoiding the need for transcription factor production altogether and theoretically further reducing the burden of the system on the host.

## 4.1 Introduction

In this introductory section we look at the challenges faced when attempting to perform logic computations in a biological context. We look at the conflicting approaches taken to achieve binary computation in synthetic gene circuits and the complexity that frequently arises as a result. We next look at how transcriptional interference can help solve this problem, what is meant by TI and in what forms it has been shown to exist in previous work. Finally, the use of CRISPR/dCas9 as a tool for gene regulation is introduced.

### 4.1.1 Binary computation in biological circuits

A frequently encountered failure mode in the construction of synthetic gene circuits is the graded response. In circuits where computation based on logic functions and decision-making based on the integration of a set of inputs takes place, signals typically need to be binary: i.e. 1 or 0, all or nothing. Since gene expression is typically a gradual process, these circuits must be engineered specifically to realise this behaviour. A classic way of achieving this is to use transcription factors with a strong non-linear response of gene output to transcription factor (TF) concentration. In these TF-based systems, the transition from low induction to high induction is condensed to a very tight range of TF concentrations. Any concentration below this range leads to low expression and concentrations above this range lead to high expression, which leads to tight switching behaviour in gene regulatory circuits that utilise them.

The problem with these transcription factors is that only a limited number are available and their binding specificity cannot be reprogrammed in the way that TAL-effectors and Cas9 can be. Conversely, the desired non-linear dose-response curves produced by the cooperative binding effects of almost all TFs, are seen with TALEs and Cas9, making them unsuitable for digital-style logic.

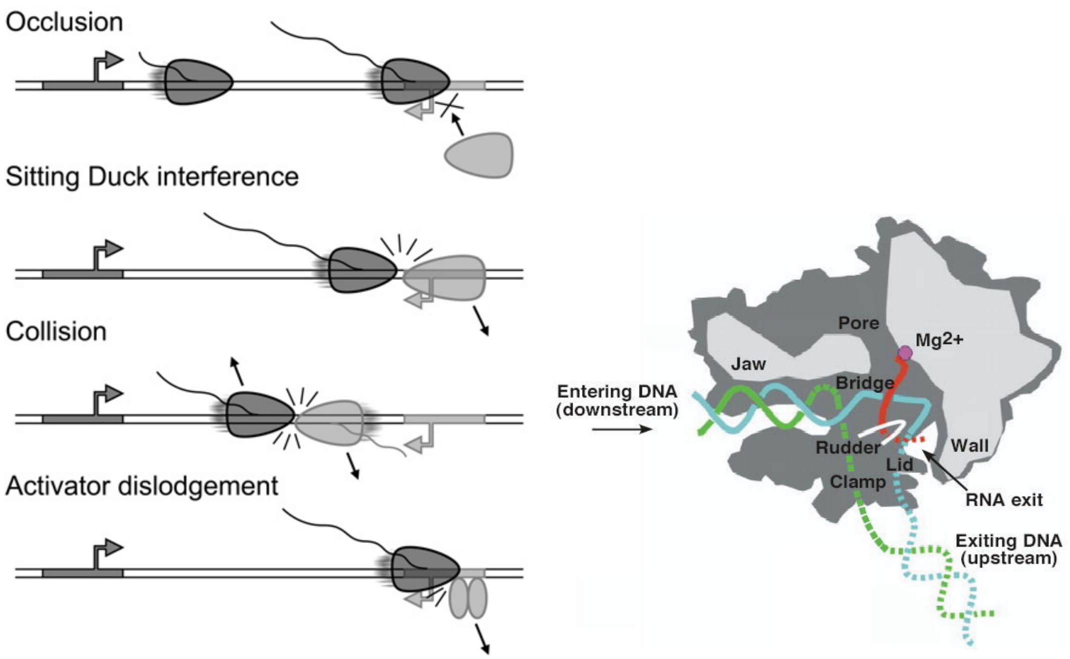
While there is a strong desire to move towards genetic circuits built from programmable TFs like TALES, if the lack of non-linear responses is not dealt with, the circuit will not be able to perform calculations based on logic functions and the required computation will fail. Rather than binary signals travelling through the system, a circuit built from TFs with non-cooperative binding will suffer from a graded response: input signals will partly induce the circuit genes and all downstream genes will be expressed in a semi-induced equilibrium, that balances activating and repressing signals without ultimately resolving the computation.

It is possible to avoid a graded response in circuits that cannot use TFs with a non-linear response or where the response curve is already insufficiently non-linear. This is done by adding additional feedback loops to the original circuit that effectively act to make the dose-response curves non-linear[217]. Typically, this approach requires additional transcription factors to be introduced into the system to carry out this added layer of feedback. This comes at the price of additional complexity and also adds more burden of protein expression onto the host organism. In the thesis introduction, we argued that complexity is one of the major and most fundamental challenges holding back synthetic biology (see **section 1.6** on page 47). In this chapter, we therefore seek to reduce the complexity of circuits and the burden on the host by implementing transcriptional interference as a novel way to add feedback loops into simple genetic circuits, without the requirement for additional TFs. In addition, our implementation of TI acts *in cis* at the RNA level, and so decreases the response times of the feedback signals and thereby improves the performance of the circuit by making the feedback quicker.

### 4.1.2 Introduction to Transcriptional Interference

Transcriptional interference can be seen as a complementary form of regulation. We define TI as the suppressive influence of one transcriptional process, directly and *in cis* on a second transcriptional process within a DNA region. This definition excludes other types of interference, such as RNAi, inactivation of RNA polymerase by RNA regulators and the effects of DNA replication on transcription.

Several distinct mechanisms simultaneously play a role in TI, and these are shown in panel **a** of **Figure 4.1** on the following page. In **occlusion**, an RNA polymerase cannot bind to a promoter because it is occupied by a traversing RNA polymerase originating from another promoter. The RNA Pol II elongation rate in yeast has been reported to be between 15 and 40 bases per second[212,218] and so from this it follows that transit time over a 100 bp core promoter region is typically on the order of 2.5 to 7 seconds. This type of TI therefore only plays a role when the second promoter is highly active. This model alone does not appear to be able to account for the full strength of TI observed in experiments[219]. However, when the RNA



**(a)** Four propsed mechanisms for transcriptional interference. See text for a detailed description of each mechanism. Image from Palmer *et al.* 2009[219].

**(b)** Position of DNA in the yeast RNA polymerase II holoenzyme. Coding strand shown in green, template strand in blue, RNA in red and the active site in magenta. Image from Hahn, 2004[220].

**Figure 4.1:** Illustrative diagrams showing proposed mechanisms behind the observed effects of transcriptional interference and the close interaction between RNA polymerase and DNA in yeast.

polymerase traversing the core promoter is made to pause in this region the effect of TI is more dramatic, but unfortunately this scenario is difficult to engineer[219].

As shown in panel **b** of **Figure 4.1**, RNA polymerase II closely interacts with both the template and the coding strand as it passes along the DNA. The DNA is partially melted, forming the transcription bubble from where the RNA is synthesised. This interaction with DNA and the melting of the DNA is incompatible with DNA binding by other proteins. As the polymerase passes over the DNA, other proteins that are already bound to the DNA are displaced. This gives rise to a variety of additional TI mechanisms, which we discuss below.

A mechanism that is thought to be of higher significance than the occlusion mechanism discussed above, is the so-called **sitting duck interference**. As discussed in the general introduction (see **subsection 1.3.2** on page 26), the pre-initiation complex assembles at the core promoter and promotes the ’firing’ of rounds of transcription. As it assembles at the core promoter, it is susceptible to being displaced from the DNA by a traversing RNA polymerase, in resemblance of a sitting duck that is an easy target in hunting. The sitting duck model was found to be of significant importance when it was discovered that TI was significantly reduced if the interfering polymerase was removed from the DNA before reaching the susceptible promoter[221]. Interference is greatest when the rate of formation of the initiation complex at the sensitive promoter is equal to its rate of firing. If the ratio of these rates is different from 1, then sitting duck interference is reduced. If it is less than 1, then transcription initiation from the PIC is fast and the PIC is less likely to be hit. If it is larger than 1, then any displaced PIC is rapidly replaced by another. Sitting duck interference is the dominant mechanism when the two promoters are in close proximity, because under these circumstances collisions between elongating polymerases are substantially less likely.

**Collision** occurs when two RNA polymerases transcribe a region of DNA in opposing directions. This leads to premature termination of the transcriptional process in at least one of the polymerases[222]. It is not exactly known how the two stalled RNA polymerases are rescued from this situation. One could simply be released from the DNA, while the other continues the transcriptional process. Alternatively, both could be released from the DNA, potentially through the action of host factors. It is also not known to what extent trailing polymerases can influence the outcome of a collision event. Intuitively, it should be the case that collision events should happen more frequently when the overlapping transcribed region increases. This notion is also backed by modelling efforts, along with the expectation that increasing promoter strength increases interference[223].

Finally, in a process similar to sitting duck interference, **activator dislodgement** can cause a reduction in transcription initiation from the sensitive promoter. In this type of TI, activators that play a role in assembly of the PIC are displaced by traversing RNA polymerases[224]. Depending on the affinity of the activator to the DNA, this can have a negative impact on the activation of the promoter. Conversely, the same principle applies for repressors, and thus this type of TI can also theoretically lead to increased activation of a promoter, depending on the DNA binding rate of the repressor.

### 4.1.3 Possible implementations of Transcriptional Interference

Transcriptional interference is found in numerous situations in nature in a variety of implementations. These can be divided into three main categories, as shown in **Figure 4.2** on the next page. The relative importance of each of the previously described TI mechanisms is subtly different in these three categories. Here we describe the three categories and give examples of instances of where they evolved in nature.

In the first implementation of TI, two convergently oriented promoters are situated head-to-head, directly facing each other without separation by other elements. In this type of TI, collision events play a minor role, since the overlap between the transcribed regions is small. Examples of this implementation can be found in coliphage 186, where it is involved in the transcriptional switch between lysis and lysogeny[225]. It is also found in a bistable switch involved in conjugation in *Enterococcus faecalis*[226]. Recently this implementation was successfully used in the engineering of a bistable switch circuit with TetR and LacI TFs in a mutually repressive configuration in *E. coli*[227].

In the second implementation, the promoters are also oriented convergently. In this type, however, the ORFs are located between the promoters, such that transcription of one ORF continues into the end of the opposing ORF. This type relies heavily on collisions and not so much on the sitting duck mechanism, as the polymerases have to transcribe through both ORFs in order to reach the other promoter. This increases the chance that the polymerase gets displaced from the DNA strand in a collision event before reaching the other side. We call this type of implementation the duelling configuration, in analogy to the involved parties first increasing the distance between each other before turning around for the confrontation.

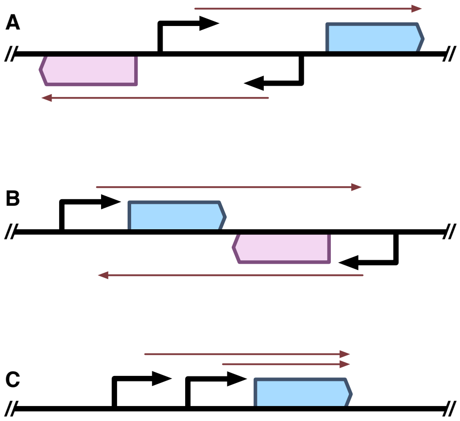


**Figure 4.2:** Implementations of transcriptional interference. **A**: head-to-head configuration. **B**: duelling configuration. **C**: chase configuration. Thick black arrows: promoters, blue and purple blocks: open reading frames, thin red arrows: areas transcribed by RNA polymerase.

The duelling configuration has been shown in several instances to occur in nature. In yeast, a significant but small effect was found in the convergent POT1 and YIL161w genes[228]. On a genome-wide level, it was also shown that expression levels of convergent pairs are significantly negatively correlated across a broad range of environments[229]. This indicates that TI plays a genome-wide role in *S. cerevisiae*. The duelling configuration has also been shown to be applicable as a tool for synthetic biology in recent work in *E. coli*[230,231], indicating the utility of TI in this field.

Finally, TI also occurs when two promoters are placed in tandem, which we refer to as the chase configuration. In this configuration the second promoter is placed directly upstream of the first and represses when active[232]. Transcription takes place in the same direction from both promoters, so collisions of RNA polymerase do not play a role in this type. Instead, this type relies mostly on sitting duck and activator displacement interference. In yeast, this type of implementation is found in the actin gene, where the actin intron contains a cryptic promoter that can be switched on by repressing the upstream promoter[233].

### 4.1.4 Transcriptional Interference at the GAL locus

In some cases, different types of transcriptional interference co-occur in nature. This is the case at the GAL1-GAL10 locus in yeast. At this locus, a combination of the duelling and chase configuration is present, which is especially relevant since the GAL1 promoter is used extensively

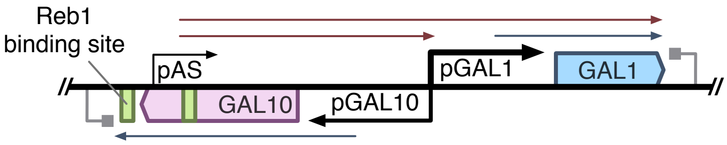


**Figure 4.3:** Transcriptional interference at the GAL1/GAL10 locus is *S. cerevisiae*. Non-coding mRNA is transcribed from a cryptic antisense promoter (pAS) at the 3' end of the GAL10 ORF. Thick black arrows: promoters, blue and purple blocks: open reading frames, thin arrows: areas transcribed by RNA polymerase (blue: coding, red: non-coding).

in this work. As shown in **Figure 4.3** on the preceding page, the GAL10 ORF contains a cryptic promoter that gives rise to non-coding antisense transcripts. In work by the Mellor and Vogelauer labs, it was shown that this promoter is flanked by Reb1 binding sites, one within and one outside of the ORF, and that it is dependent on these sites for its function[234,235]. From this promoter, two non-coding transcripts arise. One covers the GAL10 transcription unit, whereas the second covers the entire GAL1/GAL10 locus, ending at the GAL1 terminator.

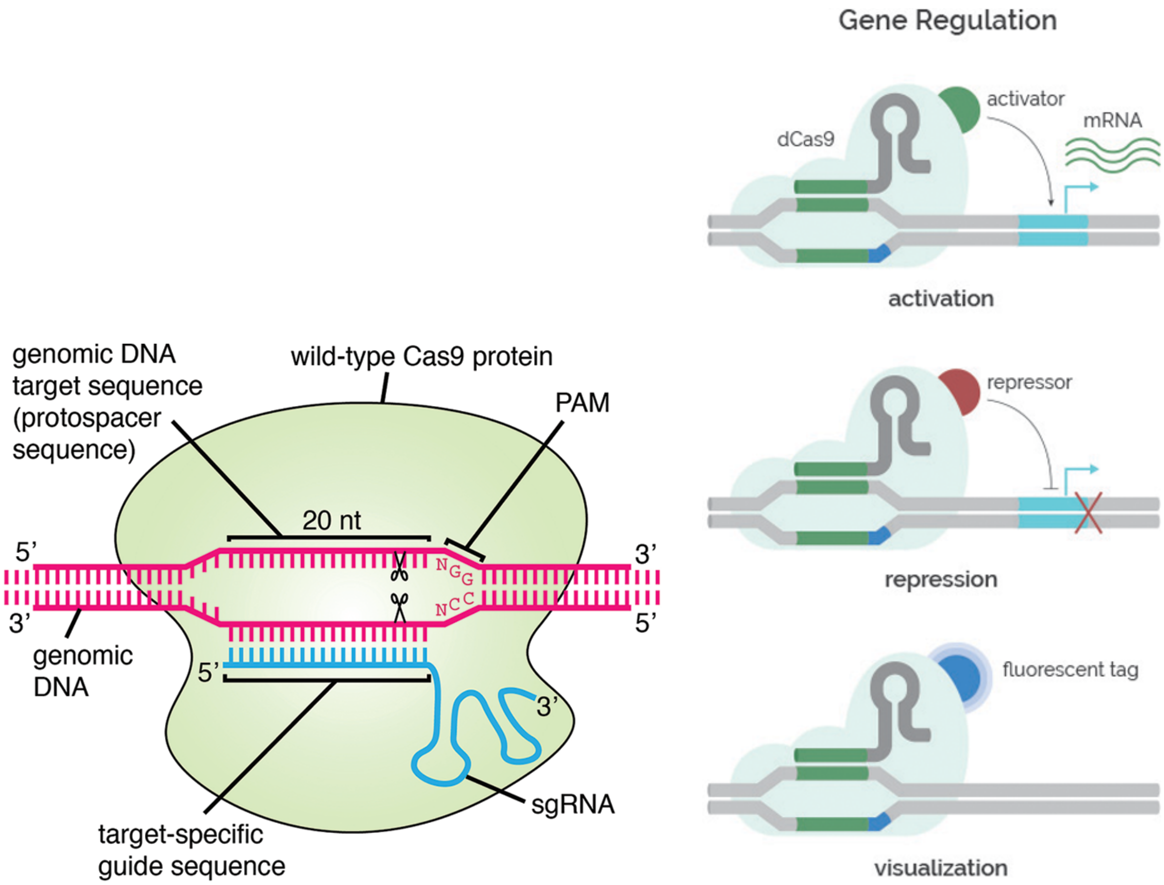
Transcription from the pAS promoter is not always detectable. It is especially detectable when the bidirectional GAL1/GAL10 promoter is repressed, as is the case when glucose is present in the growth media. This is consistent with the hypothesis that transcription in one direction inhibits transcription in the opposite direction. Strikingly, transcription from the pAS promoter also results in histone methylation, which in turn represses the GAL1 promoter[236]. This shows that the interference can work both ways and that it is consistent with the required function.

This evidence, together with results from other loci[237,238], support the notion that transcriptional interference can be used for implementation of a bistable switch.

### 4.1.5 CRISPR/dCas9 for gene regulation

Towards the end of this chapter we explore the use of dCas9 to implement regulation in our system and so here, we give a brief introduction of this protein with respect to its origin and function. dCas9 is the catalytically inactive (dCas9 stands for dead-Cas9) version of the Cas9 protein which is part of a bacterial adaptive immune system[239]. In this immune system, Cas9 is capable of selectively targeting foreign DNA and inducing a double stranded break to neutralise it. The targeting is achieved by a short RNA molecule that contains a 20 bp sequence which hybridises with the target DNA. In the natural system the Cas9 protein requires several short segments of RNA for targeting. Researchers have been able to link these to form a single RNA molecule that can target Cas9 to the correct location on the DNA, which is called the guide RNA (gRNA). In the natural system, the gRNA equivalent is encoded on the host cell genome. To prevent Cas9 from targeting the genome, it also requires that a specific sequence be present adjacent to the target sequence. This sequence is called the Protospacer Adjacent Motif (PAM). Different versions of Cas9 require different PAMs and the version that is most used in the context of synthetic biology is the one isolated from *S. pyogenes*, which requires a PAM DNA sequence of NGG. These features are shown in panel **a** of **Figure 4.4** on the next page.

The ease with which Cas9 can be targeted to specific regions of DNA has revolutionised many areas of life sciences and biological engineering. By creating a mutant without the ability to cleave the targeted DNA, the binding-only version called dCas9, now also offers a whole further variety of programmable functions, especially with regards to gene regulation[242]. For example, in yeast fusion of a GAL4 activation domain to dCas9 turns this protein into an RNA-guided transcription factor that induces expression from promoters when their sequence matches that of the co-expressed gRNAs. Fusion of the MxiI repression domain to dCas9 reverses the function, turning the protein into a repressor protein. In other studies, the location of DNA can be mapped with dCas9. By fusing a fluorescent tag to the protein, it can make real-time localisation of DNA targets within cells now possible. For a visual representation of just a few possibilities with



**(a)** Recognition and Cas9 induced cleavage of genomic DNA. PAM: protospacer adjacent motif, the short NGG sequence is required for binding and cleavage by the *S. pyogenes* Cas9. sgRNA: synthetic guide RNA, necessary for hybridisation with the DNA binding site. Figure adapted from Agrotis and Ketteler, 2015[240].

**(b)** dCas9-based regulation and visualisation. Protein fusions with activator domains, repressor domains and fluorescent tags can be used to implement non-native functions of the dCas9 protein. Figure from ERS, 2016[241].

**Figure 4.4:** Illustrations of the CRISPR/Cas9 enzyme complexed to DNA and the use of the CRISPR/dCas9 system to program gene regulation.

dCas9 see panel **b** of **Figure 4.4**. Evidently, this tool offers great potential for applications not just in synthetic biology but for any field in the area of molecular biology.

## 4.2 Aims and strategy

In this chapter, we aim to establish transcriptional interference as a novel type of regulation that can be applied generally in synthetic circuits. Already, recent work by others in *E. coli* has established the potential benefits of TI, and we believe that this new form of engineered regulation offers a reduction of circuit complexity, increases the predictability of the system (by allowing more accurate modelling) and decreases the response time of the feedback by eliminating intermediate steps required for gene expression.

As detailed above, several implementations of TI are possible, and from these options we have chosen here to implement the head-to-head configuration. Initial experiments (data not shown) found that TI in the chase configuration was not strong enough to have the required effect.

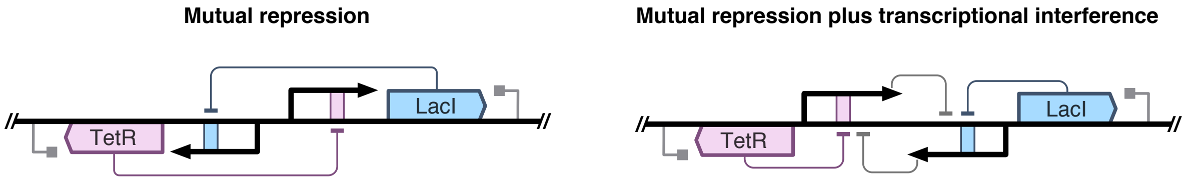


**Figure 4.5:** Comparison of bistable switch designs without (left) and with transcriptional interference (right). Repression originating from TI is indicated with grey lines.

The duelling configuration was also considered, however, for this implementation terminators must be removed to allow transcription to continue into the opposing ORF. However, as we described in thesis introduction **subsection 1.5.4** on page 45, the nonsense mediated decay (NMD) pathway degrades mRNA with exceptionally long 3'UTRs. Since this configuration leads to a particularly long 3'UTR we expected these mRNAs to be substrates for the NMD surveillance mechanism. In addition, because transcription is not ended by a terminator sequence, we expected these transcripts to lack a poly(A) signal at the 3' end which could lead to potential problems with nuclear export and translation efficiency.

For these reasons, we concluded that the head-to-head configuration was the most promising for our work in engineered TI. In this configuration transcription can be ended normally by use of regular terminator sequences. It has also been applied successfully in *E. coli*, where it was used to construct a bistable switch[227]. Given the fundamental importance of bistable switches in synthetic biology and the fact that TI has also been found to play a role in natural switches, we also chose here to implement transcriptional interference within a bistable switch in yeast. When this work was started and for several years after, no synthetic long-lasting bistable switch had ever been reported in yeast, despite several attempts[43,243] and also despite the successful implementation in other organisms over a decade earlier[1].

Previously, a bistable switch in yeast was implemented in a mutual-repression configuration using the LacI and TetR repressors usually found in prokaryotes. As shown in **Figure 1.16** on page 32, these repressors can be regulated through the addition of small molecules to the growth media and they are also known to exhibit cooperative binding, making them ideal for robustness of the bistable switch. While these repressors have been used successfully for bistable switches in *E. coli* when implemented in yeast, they do not form a fully bistable switch. Instead, the genetic switch would always revert to one of the two states after some time and could not be built to retain memory for sufficiently long enough times[43].

In this chapter, we therefore propose to stabilise this implementation of the bistable switch with the addition of transcriptional interference in the head-to-head configuration. **Figure 4.5** shows the difference between the original implementation and the one proposed here. Theoretical work has shown that cooperativity is not strictly necessary to achieve bistability and that instead it can be achieved by the construction of additional feedback loops[244]. We therefore hypothesise that the addition of regulation in the form of transcriptional interference to a bistable design where the cooperative effects of the repressors are not sufficient to achieve full bistability, can stabilise the circuit to realise full bistability.

## 4.3 Results

In this chapter we implement transcriptional interference as a tool for augmenting regulatory circuits, in particular the model system of a bistable switch. After establishing the functionality of TI at the mRNA level, we observe that the chosen implementation does not lead to protein expression. We then examine a number of solutions, including cap-independent translation initiation and intron sequences. Finally, we implement a dCas9-based system for functionalisation of the produced mRNA.

### 4.3.1 Transcriptional interference in promoters going head-to-head

The core of the bistable switch design to be built and tested in this chapter relies on a convergent promoter configuration. This configuration introduces a unqiue constraint to the promoter sequences that is not normally a consideration needed for typical genetic circuits. In this design, the promoter situated opposite the forward facing promoter is always incorporated into the 5'UTR of the first promoter and this adds the extra requirement that the sequence does not interfere with downstream processes, such as translation of the mRNA. Primarily, this means that the promoter sequence must not encode a premature start codon into the 5'UTR of the other gene on the reverse strand. If the promoter sequence does contain one CAT site (the reverse complement of ATG), the best case scenario is that the intended protein gets fused to a random peptide sequence. In the worst case, the mRNA is degraded through the nonsense mediated decay pathway and the protein never gets made[245].

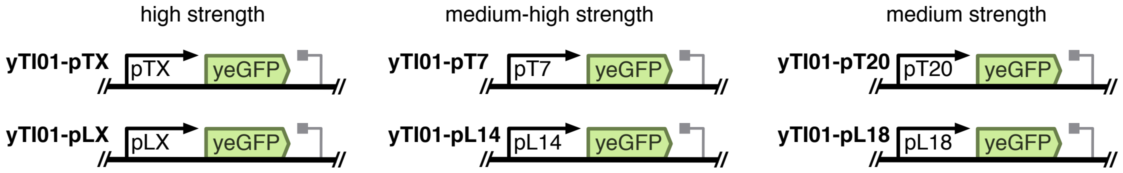
It should be clear that even the best case scenario can lead to unpredictable results. In order to avoid this, promoters had to be selected in this chapter that did not encode in any start codons into the reverse complement sequence. Unfortunately, it is to be expected that the short 3 base-pair CAT sequence occurs relatively frequently in promoter. Any particular three nucleotide sequence can be expected to occur once every ($4^3$=) 64 bases on average in randomly generated sequences. Promoters are not random sequences, however, this number is useful in assessing whether this could potentially pose a problem.

Promoter size is not straightforward to compute as the upstream boundary is frequently poorly defined. An upper limit for promoter size can be found by defining the promoter as the sequence in between two ORFs. This has been done for yeast and the median length was found to be 455 bp[246]. Indeed, many frequently used promoters including the GAL1 promoter are cloned as parts of 450 bp or more[8]. This is typically either because they have annotated sequence features up to this length or because of uncertainty on the location of the upstream boundary. At this size, the occurrence of premature start codons is likely to play a major role because these promoters are several times larger than 64 bp.

From the above considerations it follows that any promoter selected for the project would likely need to be modified to remove CAT motifs, because there are very few yeast promoters known to be short enough to not be likely to encode these premature start codons. Promoter selection was therefore made based on other criteria. We selected the synthetic promoter libraries developed by Ellis et al.[43] which offer promoter pairs repressible by two orthogonal protein repressors. This regulation is essential for the correct operation of the proposed bistable circuit. In addition, these libraries offer different strengths of promoters, which could aid troubleshooting and characterisation efforts. Finally, the two protein repressors, LacI and TetR, can be inactivated through the addition of IPTG and ATc, respectively. This further expands the toolbox for troubleshooting and characterisation of the proposed system.

The promoters of these two libraries were scanned for the occurrence of the CAT sequence which will cause premature translation initiation. Unexpectedly, only a single CAT was found in the sequence region that was shared between all promoters in the libraries and so for each of the selected promoters this sequence was changed to CCT. Doing so also created a new unique restriction site that facilitated cloning and diagnostic screening.

Not all of the available library members were devoid of CAT sequences in the diversified parts of their promoters and so we only selected library members that were CAT-free in this region. This decision was made because adding a mutation into the core promoter sequence could cause a significant change in promoter strength. We chose to select a high, medium-high and medium strength promoter from both the TetR repressible and LacI repressible libraries and worked with these for the project.

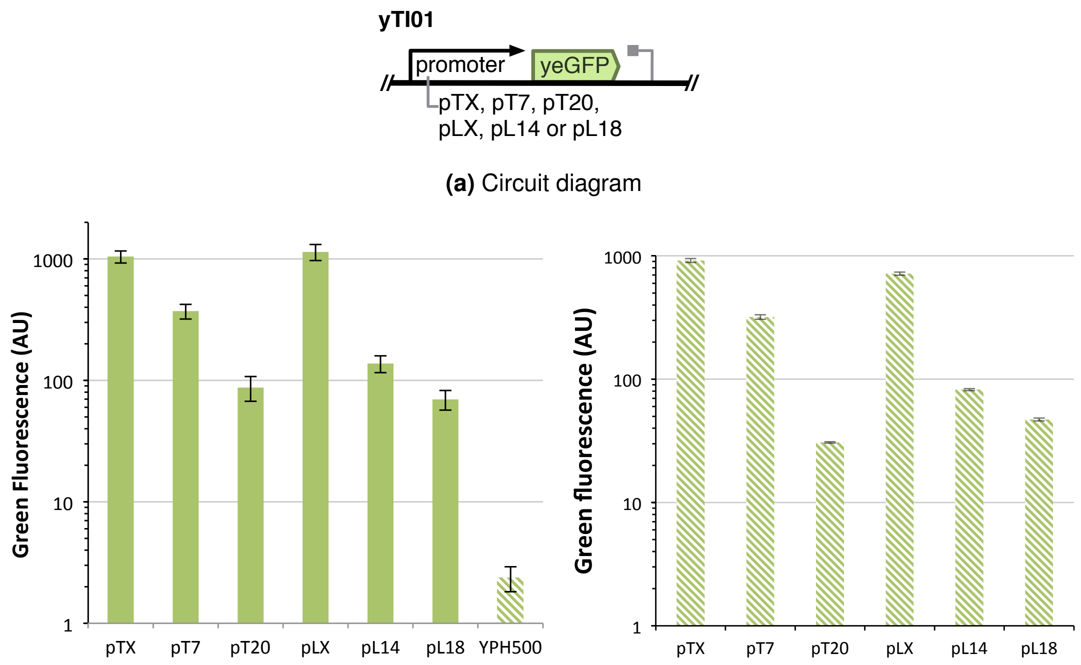


To test if the promoters behaved as expected with the CAT to CCT modification, they were each cloned upstream of yeGFP in a pRS406-based plasmid. This plasmid was integrated into the URA3 locus of YPH500 and used for characterisation. The selected TetR repressible promoters were pTX, pT7 and pT20 and the corresponding yeast strains incorporating these promoters were named yTI01-pTX, yTI01-pT7 and yTI01-pT20. The selected LacI repressible promoters were pLX, pL14 and pL18 and the corresponding yeast strains incorporating these promoters were named yTI01-pLX, yTI01-pL14 and yTI01-pL18.

These six strains were tested for yeGFP expression using flow cytometry after overnight induction with 2% galactose in YEP media (for method see Materials and Methods **subsection 2.1.3** on page 58). The parental strain YPH500 was included as a negative control. The measured data were analysed using Matlab and are presented in **Figure 4.6** on the next page.

The results were compared to the published expression strengths of the unmodified promoters. These had been characterised in a similar manner[43] but on a different yet related flow cytometer (a BD FACScalibur machine). A comparison of the results shows that the obtained expression strengths match the published dataset closely. A quantitative comparison is not possible because the published data set does not contain information on the level of autofluorescence of the parental strain.

The promoter that shows the most significant difference with the published data set is the T20 promoter. However, its expression is 0.23 times the level measured for pT7, which is a considerable reduction and suitable for our needs. The pL14 promoter was found to be more of a medium strength promoter than a medium-high strength one. Yet its expression level is twice as high as pL18, and this is a significant enough difference to be of use in further experiments.



**(a)** Circuit diagram

**(b)** Modified promoters

**(c)** Original promoters

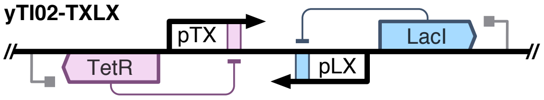
**Figure 4.6:** Characterisation of a set of modified promoters and comparison to the matching originals. Modified promoters contain a single A to C base change compared to the originals and, like the originals, are TetR repressible (pTX, pT7 and pT20) or LacI repressible (pLX, pL14 and pL18). Strains were cultured for 20 hours in inducing (2% Gal) YEP media. Bars represent median green fluorescence in arbitrary units (AU) of a single clone. Data collected on a BD FACScan flow cytometer. Error bars represent median absolute deviation of a gated population around the median of forward and side scatter. Controls (hatched): 'YPH500': non fluorescent wild type/parental strain. The reference values (hatched, subfigure **c**) are reproduced from Ellis et al.[43]. In this panel, error bars represent s.e.m. of three repeated measurements.

We conclude that the removal of the reverse premature start codon had no meaningful impact on the expression strengths from the promoters and that these matched expectations. This thus provided us with the tools for further construction and characterisation of the proposed circuit.

### Construction and testing of a toggle switch circuit.

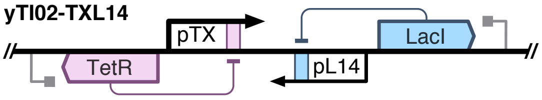
The previous experiments showed that the six repressible promoters were behaving as expected and as required. This allowed us to perfom the implementation of a bistable circuit to the design outlined in the aims and strategy section. In this design, mutual repression by the protein repressors is complemented by transcriptional interference in two promoters situated in a head-to-head configuration, also referred to as a convergent orientation.

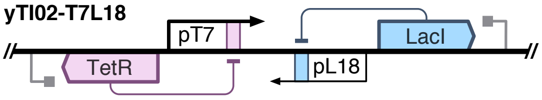
We hypothesised that a bistable switch would function most robustly when the two promoters driving the circuit would be of equal strength. Additionally, we expected the effect of transcriptional interference to be the strongest when the promoters have a high expression strength. For this reason we selected the high strength TX and LX promoters, which were shown to be both high strength and of very similar strength (within 10%) in the previous experiment.



This 'toggle' construct was assembled using conventional restriction enzyme cloning, based on the pRS405 vector. It was integrated into the genome of the YPC1 strain, which is detailed below. The resulting strain was called yTI02-TXLX and a diagram is shown above.

For troubleshooting and characterisation purposes, we also included toggle circuits that were artificially tipped to one state by using pairs of promoters that were not balanced in expression strength. This would increase the effect of transcriptional interference on the weaker promoter and magnify its effect, potentially giving more insight into the mechanism.

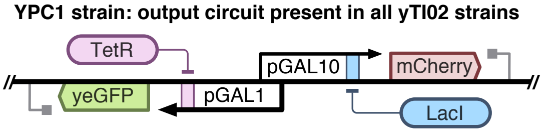




For the implementation we chose to put the strong TX promoter in a head-to-head configuration against the medium-strong L14 promoter creating a 7.6 fold difference in expression strengths according to the results in **Figure 4.6** on the preceding page. In a second construct we placed pT7 and pL18 in a convergent orientation for a difference in expression strength of 5.3 fold. The constructs were built in similar fashion to the TXLX construct. The strains containing these assemblies were named yTI02-TXL14 and yTI02-T7L18 respectively. A diagram of the constructs is shown above.

**Output circuit (YPC1)** In order to gain information about the state and behaviour of the system, the levels of the two repressors needed to be measured. Direct measurement of the protein levels was not a viable option. Consequently we opted for an indirect measurement by placing fluorescent proteins under control of two promoters repressible by either TetR or LacI. The strong Gal1 promoter repressible by TetR was placed to control yeGFP and the medium strength Gal10 promoter repressible by LacI was used to control mCherry red fluorescent protein expression. As TetR concentrations rise, yeGFP levels would go down, while LacI presence would repress mCherry expression.

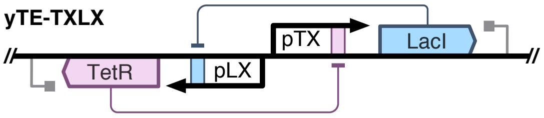
This is not a perfect system, as non-linearity in the response of the promoters could lead to low levels of repressor that go undetected. However, this would also hold true for the effect of the repressor on the promoters in the toggle circuit. Therefore, levels undetectable by the output circuit are very likely to be functionally insignificant. Hence we believe that this is an adequate method for determining the state of the toggle switch circuit.



A diagram for the output circuit is shown on the right, consisting of the two reporters driven by their respective repressible promoters. The circuit was constructed using conventional restriction enzyme cloning in a pRS404 based plasmid. The construct was added to all strains prior to integration of the toggle circuit and was integrated into the URA3 locus of YPH500, using the TRP1 marker for selection to create the strain called YPC1. The toggle circuits to be assessed were subsequently integrated next to the bacterial AmpR cassette of the previously integration into YPC1, this time using the LEU2 marker for selection.

The use of the TetR and LacI repressors offered another opportunity to gain more insight into the performance of the bistable switch. By introducing ATc and/or IPTG into the culturing medium, the repressive effect of TetR and LacI (respectively) could be suppressed completely. This allowed for another method to influence the balance of the toggle circuit and obtain information about its behaviour and performance.

A consequence of the indirect measurement of repressor levels is that when a repressor is inactivated through addition of a small chemical, its output cannot be read by the output circuit. Because of this, we could not obtain results for the state of the toggle when neither repressor was active. Instead we collected results for when only the TetR repressor was active (green fluorescence only), when only the LacI repressor was active (red fluorescence only) and when both repressors were active (green and red fluorescence).



In order to assess the behaviour of the head-to-head promoter constructs, a control for the repressed state of the switch was necessary. For this we used the previously described TXLX timer circuit[43]. This circuit functions like the yTI02-TXLX toggle circuit without the transcriptional interference component. In this circuit, the promoters are not placed in a convergent orientation (see diagram). Instead, the TX promoter directly drives LacI expression and pLX directly drives TetR expression. Due to incomplete repression the circuit behaves as an unstable bistable switch that is reset to the dominant state after a set amount of time.

Similarly to the yTI02 strains, the TXLX timer circuit was integrated into the YPC1 strain containing the output circuit. The resulting strain was called yTE-TXLX. This strain served as a number of crucial controls in characterisation of the yTI02 constructs. Firstly, it allowed us to check whether the small molecules (ATc and IPTG) used to set the activity of the repressors were active and being taken up by the cells.

Secondly, it allowed for the determination of expression levels when the repressors were fully expressed. This represents the maximum attainable level of repression that can be detected by the output circuit. This was done by monitoring the output circuit in conditions where only TetR was active (+10mM IPTG) or conditions where only IPTG was active (+250ng/ml ATc).

Finally, by monitoring the output of the circuit in conditions where both repressors were active, we gained insight into the performance the newly created circuits needed to match and exceed in order to improve functionality over the existing implementation.

As a final set of controls we included the parental strain YPH500 and the output circuit strain, YPC1, excluding any additional circuits. The latter was necessary to obtain the unrepressed expression levels which were used to compare all measurement against to see if any significant levels of repressor were being expressed. The YPH500 strain informed us about autofluorescence levels and indicated how strong the tightest repression in the yTE-TXLX control strain was.

These six strains (three constructs and three controls) were assessed for yeGFP and mCherry expression by flow cytometry after overnight induction with 2% galactose in YEP media (for method see Materials and Methods **subsection 2.1.3** on page 58). Each of the strains was tested in each of three conditions: +10mM IPTG (TetR active), +250ng/ml ATc (LacI active) and plain media (both TetR and LacI active) The measurements were analysed using Matlab and are presented in **Figure 4.7** on the following page.

The results show that all control strains showed expected behaviour. From a comparison between YPC1 and yTE-TXLX in 'TetR active' conditions, it follows that IPTG is functioning and leads to a 145-fold reduction in expression of yeGFP when TetR is expressed in the cell. Similarly 'LacI active' conditions show that ATc is functioning and that LacI can repress mCherry when expressed in the cell. At 8.8-fold, the repression is not as strong as for TetR, which may be explained by the fact that mCherry is a weaker fluorophore than yeGFP and that the GAL10 promoter is a weaker promoter than GAL1. These factors combined may lead to lower maximum expression and consequently lower reduction by repression from LacI.

The results for the experimental toggle circuits did not match expectations, however. For conditions where TetR was active, we expected repression to correlate with the strength of the Lac-repressible promoter i.e. strong repression for the LX promoter and weak repression for pL18. The measurements showed no reduction in expression in any of the yTI02 constructs compared to the positive YPC1 control. This means no biologically relevant amounts of TetR are being produced in the tested strains.

For conditions where LacI was active, we expected repression to correlate with the strength of the TetR repressible promoter i.e. strong repression for the TX promoter and weak repression for pT20. The measurements showed no repression in the TXLX and T7L18 strains. However, yTI02-TXL14 did show 2.9 fold repression versus YPC1. This repression is weaker than the 8.8 fold repression achieved by the yTE-TXLX controls strain. However, it does indicate that biologically relevant concentrations of repressor are produced in this particular strain. We can infer that this is in fact an effect of repressor concentration and not the effect of a mutation in the reporter circuit, from the fact that red fluorescence is comparable to YPC1 levels in conditions where LacI is inactive (data not shown).

The fact that yTI02-TXLX did not show repression, indicated that the strength of the promoter driving LacI was not the only determinant in the output of the circuit. The experiment proved that the facing promoter also plays a role. In turn, this supports the hypothesis that transcriptional interference can play a role in regulation of genetic circuits. Of the three tested strains, the difference in strength between TXL14 is the greatest and this may be the reason that this is the only strain that showed biologically relevant levels of LacI repressor.

In conditions where both repressors were active, we initially expected the circuits to display bistable properties. In yTI02-TXL14 and T7L18 we expected the LacI expression state to dominate because the promoter strengths were biased towards that state. However, with the unexpected results in the 'TetR active' and 'LacI active' conditions, we realised that achieving bistability in these strains was unlikely.

We note that even in the 'TetR active' state, where LacI is completely inactive, the LacI repressible promoters are incapable of producing biologically relevant levels of TetR. Given this fact, it follows that 'TetR and LacI active' conditions are essentially reduced to 'LacI active' conditions, as no TetR can be produced in this system. As expected, the results for green fluorescence show that no biologically active levels of TetR are produced. The results for red fluorescence reflect this, mimicking the levels for the 'LacI active' condition.

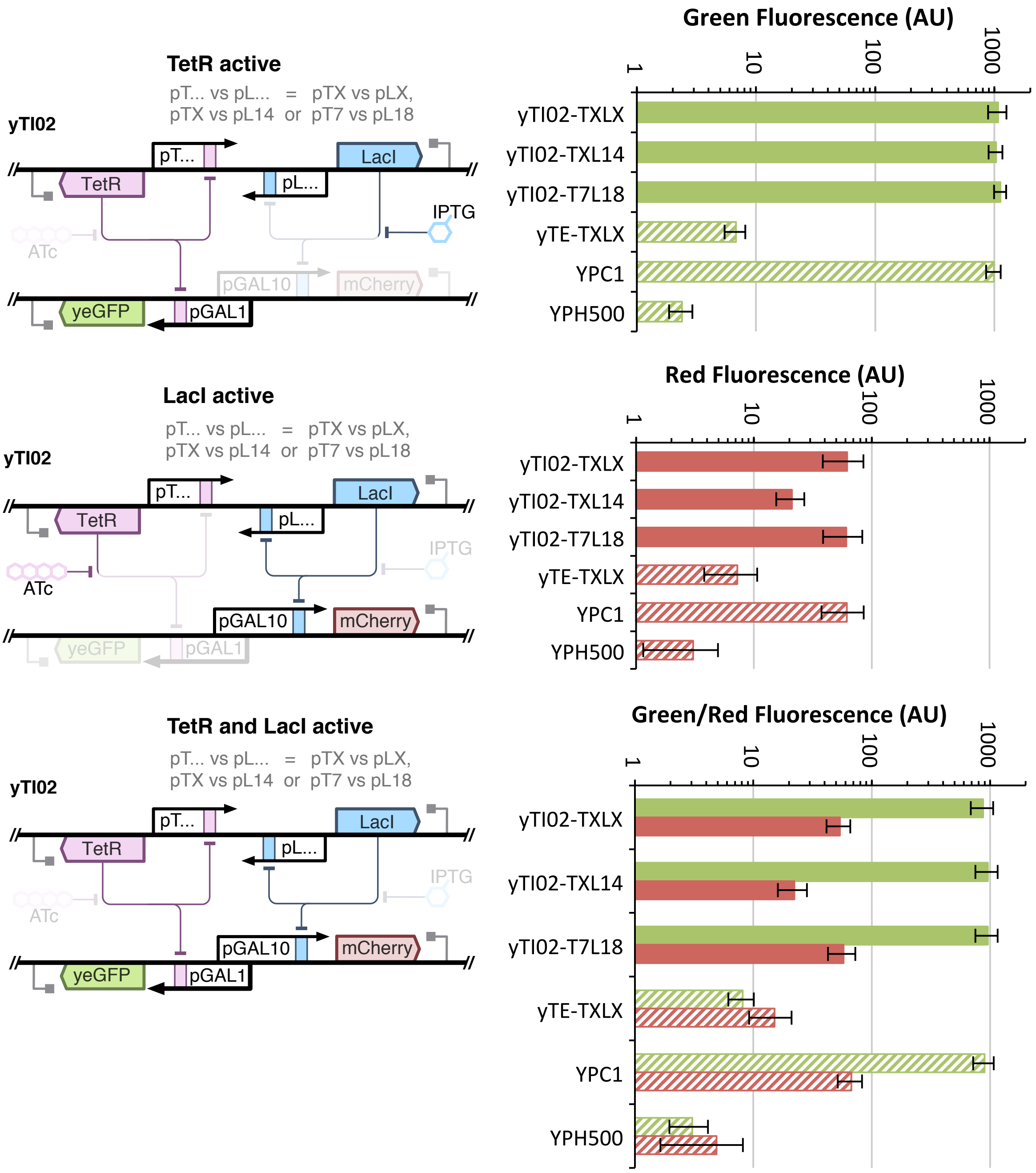


**Figure 4.7:** The response of a bistable switch circuits to IPTG and ATc inducer. Fluorescent output is shown for three circuits designed to be capable of showing bistable behaviour. The circuits contain two promoters in head-to-head configuration: pTX vs pLX, pTX vs pL14 or pT7 vs pL18. Each of the circuits was subjected to three conditions. 1) only LacI repressor active (-IPTG, +ATc). 2) only TetR repressor active (+IPTG, -ATc). 3) both repressors active (-IPTG, -ATc). Saturating concentrations of repressor inhibitor were used (250 ng/ml ATc for TetR, 10 mM IPTG for LacI). Strains were cultured for 20 hours in inducing (2% Gal) YEP media. Bars represent median green fluorescence in arbitrary units (AU) of a single clone. Data collected on a BD FACScan flow cytometer. Error bars represent median absolute deviation of a gated population around the median of forward and side scatter. Controls (hatched); 'TXLX': equivalent to the pTX vs pLX bistable circuit except with pLX directly driving TetR expression and pTX directly driving LacI, rather than the promoters facing each other in a head-to-head configuration, 'YPC1': yeGFP expressed from the strong GAL1 promoter and mCherry expressed from the medium strength GAL10 promoter, 'YPH500': non fluorescent wild-type/parental strain.

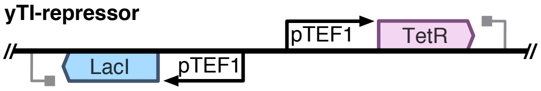
In conclusion, these experiments show that the aim of exceeding repression levels achieved in the yTE-TXLX strain was not achieved. In fact, the only strain shown to be capable of producing biologically active levels of repressor was yTI02-TXL14. Comparison of results for this strain with yTI02-TXLX indicated that transcriptional interference was playing a role in the regulation of these constructs. While most of our findings were disappointing, this was a partially encouraging result that led to further experiments devised to address some of the shortcomings in the measurements inherent to the design of the tested constructs.

### Direct measurement of fluorescent output of promoters in head-to-head configuration.

The previous experiment showed an encouraging result for the use of transcriptional interference for regulation, but it also emphasised some limitations in the data that could be obtained from theses circuits. An essential drawback of the approach is that information about the state of the system could only be obtained indirectly. Output of the promoters was converted to repressor levels, which in turn had to be converted to measurable (fluorescent) output through repression of promoters driving the fluorescence genes.

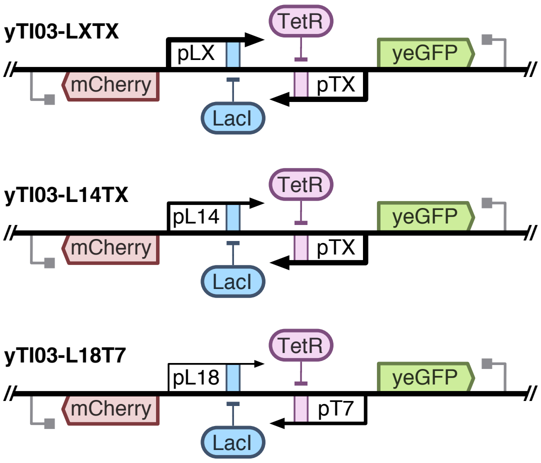
This creates inherent inaccuracies in the output levels, because the transfer function of repressor level to the fluorescent output is not designed to be a linear relationship for a repressible promoter. A more serious concern, however, was that the addition of inducer molecules ATc and IPTG, which were used to control the state of the bistable switch, also prevented the concentration of the relevant repressor to be measured.

Given that the circuits tested previously were not behaving as expected and that more precise characterisation of the state of the system in different conditions was needed, we designed a new set of strains to circumvent the issues encountered in the previous measurements. These strains contained the same sets of convergent promoters that the yTI02 strains used, but were cloned with genes for red and green fluorescence in the place of TetR and LacI, respectively. TetR and LacI were now expressed from constitutive promoters. This eliminates the indirect manner of determining the state of the system, while also preventing the addition of TetR and LacI from causing problems with data collection.



For constitutive expression of the repressor proteins, a new strain was created that served as the parental strain for the new constructs. It contained two divergently placed transcription units, driving expression of TetR and LacI from constitutive TEF1 promoters. It was constructed with a pRS405 based plasmid backbone, for integration into the LEU2 locus of YPH500. The strain was called yTI-repressor and a diagram is shown with this paragraph.

The yTI-repressor strain was used as the parental strain for the following strains, which were created to gain a better understanding of the transcriptional interference in constructs with convergent promoters. Using conventional restriction enzyme cloning, the three sets of head-to-head promoters used in previous experiments were ligated into a construct containing divergently placed mCherry and yeGFP. The vector backbone for this system was pRS406 based, allowing integration into the URA3 locus.



For cloning-related reasons, the LacI repressible promoters are now displayed facing the right and the constructs were named yTI03-LXTX, yTI03-L14TX and yTI03-L18T7. The circuits were exposed to different conditions of repressor activity to gain a better understanding of the role transcriptional interference plays in these circuits. Like in the previous experiment, three conditions were tested. In the first condition only TetR was active (+10 mM IPTG), biasing expression towards yeGFP production. In the second condition only IPTG was active (+250 ng/ml ATc), biasing expression towards mCherry expression. Finally, output of the circuits was monitored in conditions where both repressors were active (no inducer molecule) in order to gain insight into the natural balance between the promoters in repressed conditions.

These three strains were assessed by flow cytometry after overnight induction with 2% galactose in YEP media (see Materials and Methods **subsection 2.1.3** on page 58). Each of the strains was tested in each of three conditions: +10 mM IPTG (TetR active), +250 ng/ml ATc (LacI active) and plain media (both TetR and LacI active). The yTE-TXLX strain was included as a control for activity of the inducers and YPC1 was again included as a reference for maximum expression strength and galactose induction. YPH500 was included for autofluorescence reference. The measured data were analysed using Matlab and are presented in **Figure 4.8** on the following page.

From the results it quickly becomes clear that none of the strains showed fluorescence above autofluorescence levels and this is an unexpected result. In a fully functioning circuit, expression in the 'TetR active' condition would be correlated with LacI repressible promoter strength, i.e. high yeGFP expression for pLX and medium expression for the L18 promoter. In the 'LacI active' condition, expression would be correlated with TetR repressible promoter strength, i.e. high mCherry expression for pTX and medium-high expression for the pT7 promoter. In conditions where both repressors were active, we expected low expression for both fluorophores.

However, from the previous experiment we already knew that the system was not fully functioning. Many of the tested strains and conditions showed that no biologically active levels of repressors were attained. The fact that this system with direct outputs shows no increased levels of fluorescence is consistent with the earlier results. However, yTI02-TXL14 did show evidence of biologically active levels of LacI repressor. In yTI03-L14TX, this would have translated to increased levels of mCherry in the 'LacI active' condition. However, against our expectations even in that condition there is now no evidence of increased fluorescence.

In conclusion, this experiment did not yield more insight into the mechanism and performance of transcriptional interference in the regulation of these circuits and contradicted earlier positive results for the yTI02-TXL14 strain. These confounding and largely negative results prompted us to next investigate the direct outcome of our transcription, i.e. the mRNAs.

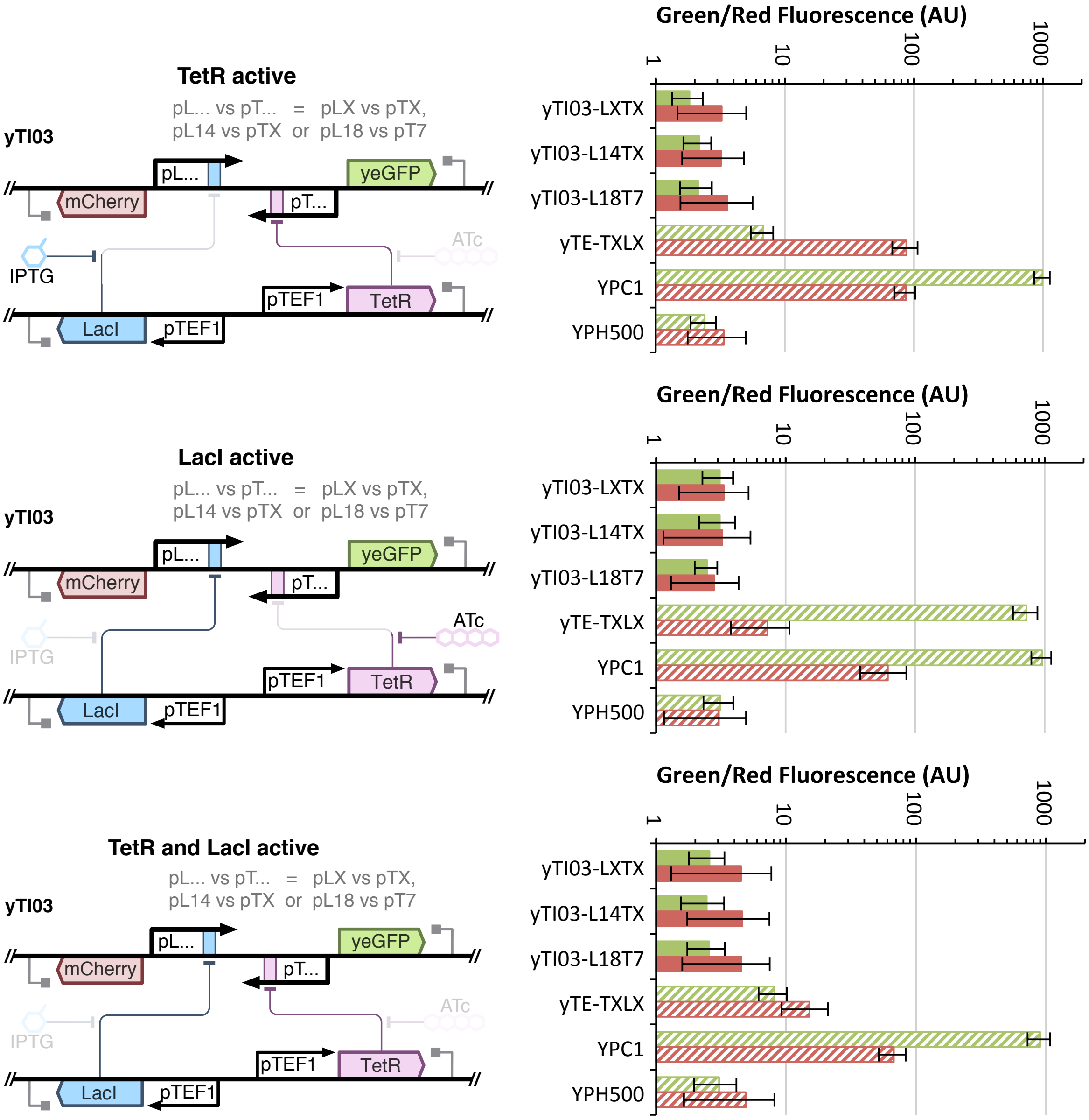
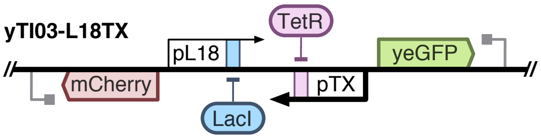


**Figure 4.8:** Direct observation of the response of convergently placed promoters to IPTG and ATc inducer. Fluorescent output is shown for three circuits containing two promoters in head-to-head configuration: pLX vs pTX, pL14 vs pTX or pL18 vs pT7. Each of the circuits was subjected to three conditions. 1) only LacI repressor active (-IPTG, +ATc). 2) only TetR repressor active (+IPTG, -ATc). 3) both repressors active (-IPTG, -ATc). Saturating concentrations of repressor inhibitor were used (250 ng/ml ATc for TetR, 10 mM IPTG for LacI). Strains were cultured for 20 hours in inducing (2% Gal) YEP media. Bars represent median green fluorescence in arbitrary units (AU) of a single clone. Data collected on a BD FACScan flow cytometer. Error bars represent median absolute deviation of a gated population around the median of forward and side scatter. Controls (hatched); 'yTE-TXLX': equivalent to the pTX vs pLX bistable circuit except with the LX promoter directly driving TetR expression and pTX directly driving LacI, rather than the promoters facing each other in a head-to-head configuration, 'YPC1': yeGFP expressed from the strong GAL1 promoter and mCherry expressed from the medium strength GAL10 promoter, 'YPH500': non-fluorescent wild-type/parental strain.

### Is mRNA produced from head-to-head promoters?

Because transcriptional interference is a process that happens at the level or mRNA production, we decided to next assess the levels of reporter mRNA produced in strains with convergently placed promoters. This would show whether some mRNA was being transcribed that is somehow not translated or whether no mRNA was produced in the first place. Answering this question was of significant importance to further development and troubleshooting.

mRNA levels were measured by quantitative Reverse-Transcription PCR (qRT-PCR). Compared to flow cytometry, qRT-PCR compares poorly in terms of cost, procedure length and throughput. For this reason, not all previously tested strains could be tested. We chose to compare the construct with two strong promoters of balanced strength (yTI03-LXTX) to a strain with a strong bias towards mCherry production as a result of a weak promoter facing pTX (yTI03-L18TX). This would increase the effect of any transcriptional interference if it were present in the system.

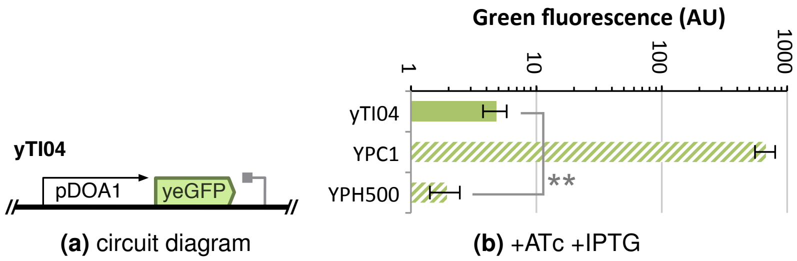


yTI03-L18TX was chosen over previously constructed strains because the larger difference in promoter strength between pL18 and pTX compared to yTI03-L14TX would emphasize the effect of transcriptional interference. yTI03-L18TX was constructed in the same manner as the other strains in the yTI03 series, which were described in the previous section. In flow cytometry experiments yTI03-L18TX was shown to be similar to the other strains in the yTI03 series, displaying fluorescence only at autofluorescence levels in all tested conditions (data not shown). A diagram of yTI03-L18TX is shown with this paragraph. Genes for constitutive expression of TetR and LacI are present in these strains, but not shown.

This experiment was designed to investigate the effect of inherent promoter strength on transcriptional interference. The effect of the two repressors in this context is also of significant importance, but is addressed in a different experiment. In order to suppress the effects of the two repressors, which were constitutively expressed from the TEF1 promoter, the strains were grown in media containing saturating concentrations of inhibitor (250 ng/ml ATc for TetR, 10 mM IPTG for LacI). Strains were grown overnight in YEPG and backdiluted to an O.D.$_{600}$ of 0.5 and grown to reach an O.D.$_{600}$ of 2.0 before further processing.

Full details of the procedure for qRT-PCR measurements are given in **section 2.3** on page 71. Briefly, the workflow consisted of extracting total RNA from exponentially growing cells. Total RNA was converted to cDNA using gene-specific primers. The cDNA was diluted 300 times and used in a qPCR reaction using the intercalating fluorescent SYBR Green dye for detection. Both the yeGFP and mCherry transcripts were amplified for detection. Primers are shown in **Table 2.8** on page 71.

After detection, the measurements had to be normalised against a reference gene. Typically the ACT1 transcript is used for this purpose, as it has a long history of use in quantitative PCR. However, based on new insights, we instead chose to use the TPI1 transcript because it has been shown to be more stable in media with alternative carbon sources[211]. Ideally, multiple reference genes are used in qPCR data normalisation, but limitations on experiment numbers prevented this.



**(a)** circuit diagram **(b)** +ATc +IPTG

**Figure 4.9:** The DOA1 promoter driving GFP expression. Expression levels of yTI04 are significantly higher than autofluorescence levels in YPH500. This proves that measured mRNA levels in the qPCR experiments would normally have given rise to detectable fluorescence levels. Double asterisks indicate significant differences (p<0.001) with explanatory measures of effect size over 0.5.

As TPI1 is a highly-expressed gene, we also measured transcript levels for the DOA1 gene. This is a gene that is constitutively expressed at levels of 0 to 5 copies per cell, with an average of 2.6 copies[212]. It was intended to form a lower boundary of mRNA levels that could still show biological activity in the form of fluorescence.

To test if the DOA1 promoter satisfied this requirement, it was cloned upstream of yeGFP in a pRS406 based plasmid. This plasmid was integrated into the URA3 locus of YPH500 and the strain was called yTI04. This strain was tested by flow cytometry after overnight growth in YEPD media (for method see Materials and Methods **subsection 2.1.3** on page 58). YPC1 and YPH500 were included as references for expression strength. The measurements were analysed using Matlab and are presented in **Figure 4.9**.

The result shows that the DOA1 promoter does indeed show a weak but detectable fluorescence that is 2.5 fold higher than autofluorecence levels, confirming its utility as a lower boundary for biologically relevant mRNA levels in qPCR experiments. As a final control, we also monitored mRNA levels for the LacI gene, which was constitutively expressed from the TEF1 promoter in all these cells.

**qRT-PCR results** The YPC1 strain, which has the bidirectional GAL1/GAL10 promoter controlling expression of yeGFP and mCherry, was used as a reference for a situation with no transcriptional interference. The GAL1 promoter driving yeGFP is a strong promoter, while GAL10 driving mCherry is a medium strength promoter. Data was collected using the Eppendorf RealPlex qPCR machine and analysed using the dd-Ct method. See **section 2.3** on page 71 for a detailed description. The results are presented in **Figure 4.10** on the following page.

Based on previous observations that the circuits in the TI03 series of strains showed no fluorescence, we anticipated two possible outcomes for this experiment. The first scenario was that no mRNA was being produced from the convergent promoters, possibly because transcriptional interference completely prevented creation of full transcripts or because the arrangement of promoters was so unnatural that the transcriptional machinery could not initiate or elongate the transcript.

The second scenario was that transcription was taking place, but certain properties of the transcript, such as its unusually long 5'UTR, prevented it from being completely translated. In this scenario, we expect transcriptional interference to cause mRNA levels for the fluorescent proteins in yTI03-LXTX to be lower than mRNA levels for yeGFP in YPC1 which is expressed

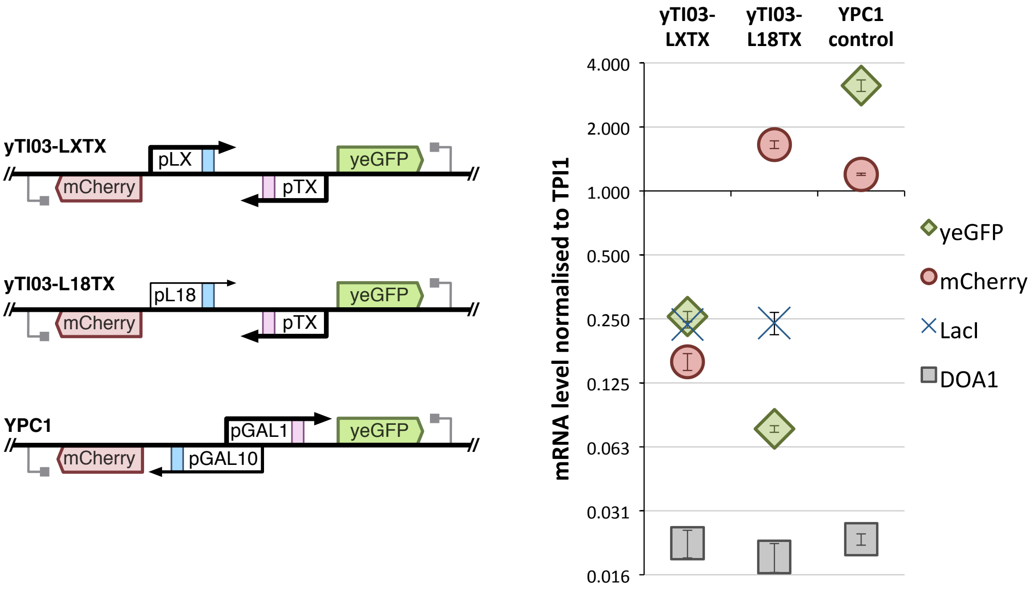


**Figure 4.10:** Demonstration of transcriptional interference at the mRNA level through qRT-PCR measurements. Two experimental strains with promoters in the head-to-head configuration (yTI03-LXTX and yTI03-L18TX) and one control strain with the divergent GAL1/10 promoter (YPC1) were tested. Left panel: circuit diagrams. Right panel: qRT-PCR results. In yTI03-L18TX, the medium strength L18 promoter replaces the strong LX promoter present in yTI03-LXTX. LacI is expressed from the moderately strong TEF1 promoter in the yTI03 strains and is absent in YPC1. The DOA1 control gene is constitutively expressed at an average of 2.6 copies per cell[212]. Results were analysed using the dd-Ct method and error bars represent the standard deviation of TPI1-normalised technical triplicates.

from a promoter of very similar strength to pLX and pTX. We further expected transcript levels for mCherry and yeGFP to be similar in yTI03-LXTX, because they are expressed from promoters of very similar strength.

The results show that the measured transcript levels closely match the expectations for the second scenario. The mRNA levels for yeGFP and mCherry in yTI03-LXTX, which are both driven by a pGAL1 promoter, are respectively 0.082 and 0.050 times the levels of pGAL1-driven transcripts in YPC1. This indicates that transcription does indeed take place, but that transcriptional interference is reducing transcript levels for these genes. mRNA levels for these genes were comparable to LacI transcript levels, which is expressed from the medium to medium-high strength TEF1 promoter. This indicates that reduced transcript levels are not the reason no fluorescence is observed in the yTI03-LXTX strain.

We expected transcript levels to change dramatically when the LX promoter was replaced by the weaker L18 promoter. Results for yTI03-L18TX show that observed transcript levels indeed changed in a way that is consistent with the assumption of transcriptional interference. Comparing yTI03-L18TX to yTI03-LXTX, the transcript level for yeGFP in L18TX is 0.30 times the level in LXTX, consistent with the use of a weaker promoter (L18 versus LX previously) driving expression of yeGFP. More remarkable, however, is the fact that mRNA levels for mCherry increase 10.5 fold. This is despite the fact that the promoter driving mCherry remains unchanged. This provides compelling evidence that the reduction of the strength of the facing promoter allows higher transcription from the first promoter because interference from the facing promoter is reduced. Expression from pTX is not fully restored to levels comparable to pGAL1 in the YPC1 strain, indicating that the L18 promoter is still interfering with transcription to some level.

Levels for DOA1 and LacI remain constant in the three strains (YPC1 does not contain the LacI gene). This indicates that the observed fluctuations in yeGFP and mCherry levels are not an artefact of fluctuations in the reference gene or otherwise unrelated to actual differences in the concentration of mRNA species in the cell.

In conclusion, this experiment shows that despite a lack of fluorescence, considerable amounts of mRNA are produced from promoters in convergent orientation. Comparison of strains with different promoter strengths confirmed our hypothesis that transcriptional interference plays an important role in the final transcription levels arising from convergently placed promoters.

### How does repressor activity affect transcript levels?

The previous experiment validated the basic principle upon which the design of an improved bistable switch was based. It showed that transcriptional interference demonstrated the required effect when the change in promoter strength was encoded into the DNA. However, in order for our design to be functional, the system also needed to respond as expected in response to changes in promoter strength brought about by the binding of repressors.

To investigate whether the binding of repressor proteins to the promoters had the required effect we subjected yTI03-LXTX to different combinations of repressor inhibitors. This strain was chosen because it contained two head-to-head promoters of equal strength. When subjected to both IPTG and ATc, neither of the promoters in this strain was repressed. From there, each of the promoters could be weakened by removing the inhibitor of the relevant repressor. This would provide insight into the effectiveness of transcriptional interference in the context of repressible promoters.

The YPC1 strain, which has the bidirectional GAL1/GAL10 promoter controlling expression of yeGFP and mCherry, was used as a reference for a situation with no transcriptional interference. In this strain the strong GAL1 promoter drives yeGFP, while GAL10 driving mCherry is a medium strength promoter. ATc was used at 250 ng/ml ATc and IPTG at 10 mM. Data was collected using the Eppendorf RealPlex qPCR machine and analysed using the dd-Ct method. See **section 2.3** on page 71 for a detailed description. The results are presented in **Figure 4.11** on the following page.

The condition where no repressors were active acted as a reference for the two conditions where one of the repressors was active. Qualitatively, this condition matched the results for yTI03-LXTX in **Figure 4.10** on the previous page, with transcript levels for yeGFP and mCherry comparable to those of LacI and yeGFP levels slightly higher than those of mCherry. Quantitatively, transcript levels are scaled differently to TPI1 and DOA1. However, we are more concerned with relative differences, rather than absolute numbers.

When we compare the initial condition to the condition where the LacI repressor is active, we find that levels of yeGFP mRNA are reduced 6.7-fold, consistent with the LX promoter driving this gene being repressed. In addition, mRNA levels for mCherry increase 2.9-fold, while no changes are made to the promoter driving expression of this gene. mRNA levels for DOA1 and LacI are constant, indicating that this effect is the direct result of transcriptional interference.

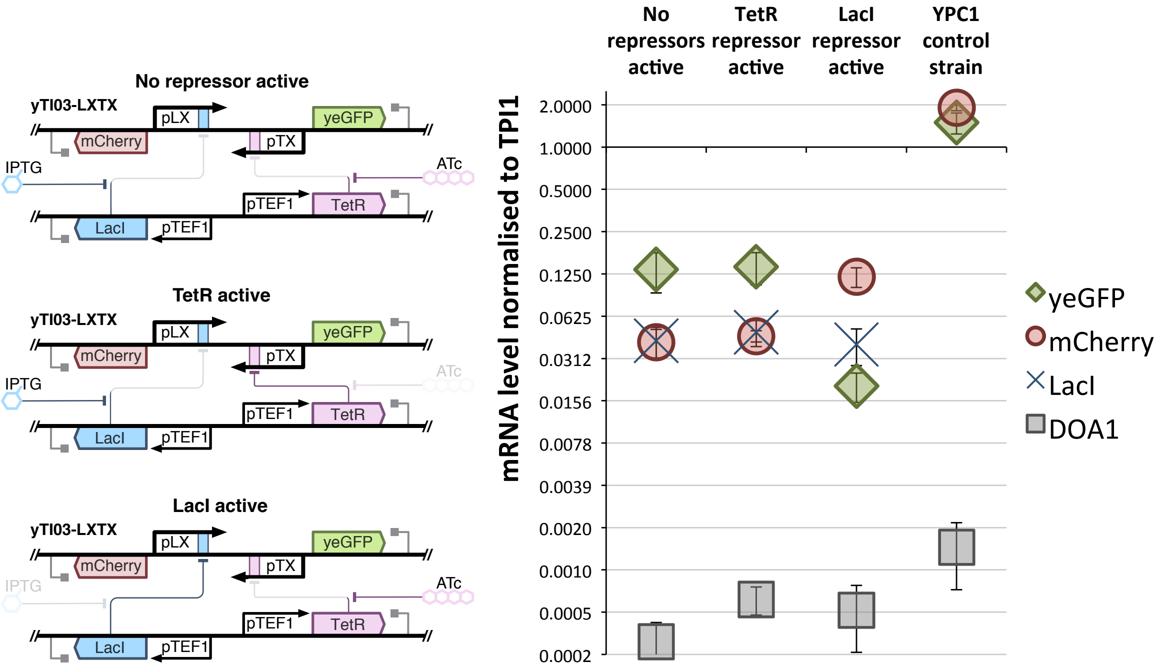


**Figure 4.11:** Demonstration of transcriptional interference at the mRNA level through qRT-PCR measurements. An experimental strain with promoters in the head-to-head configuration (yTI03-LXTX) and a control strain with the divergent GAL1/10 promoter (YPC1) were tested. Left panel: circuit diagrams. Right panel: qRT-PCR results. Promoter strength in yTI03-LXTX was modulated through the addition of inducers (ATc at 250 ng/ml ATc and IPTG at 10 mM). Three conditions were tested: both promoters at full strength ('no repressors active'), pTX repressed ('TetR active') and pLX repressed ('LacI active'). LacI and TetR are expressed from moderately strong TEF1 promoters in yTI03-LXTX and are absent in YPC1. The DOA1 control gene is constitutively expressed at an average of 2.6 copies per cell[212]. Results were analysed using the dd-Ct method and error bars represent the standard deviation of TPI1-normalised technical triplicates.

Unexpectedly, the results for the condition where only the TetR repressor is active are qualitatively identical to the condition where no repressors are active. We hypothesise that the difference in binding affinities between LacI and TetR play a role in this result and elaborate on this hypothesis in the discussion section.

Taken together, the results of the qRT-PCR experiments suggest that transcriptional interference functions mostly as expected in these circuits. Changes in mRNA levels of the output genes in response to perturbations of promoter strength are consistent with a system where TI affects transcriptional throughput. Transcript levels are at levels comparable to LacI expression, which is expressed from the moderately strong TEF1 promoter. They are also at least several times higher than the DOA1 gene, which is expressed at low levels and forms the lower boundary of protein levels that can still be detected using flow cytometry. Based on these observations, we conclude that transcript levels in our system are sufficient for detection of the protein output using flow cytometry and that the lack of flourescence must therefore be the result of issues downstream of transcription.

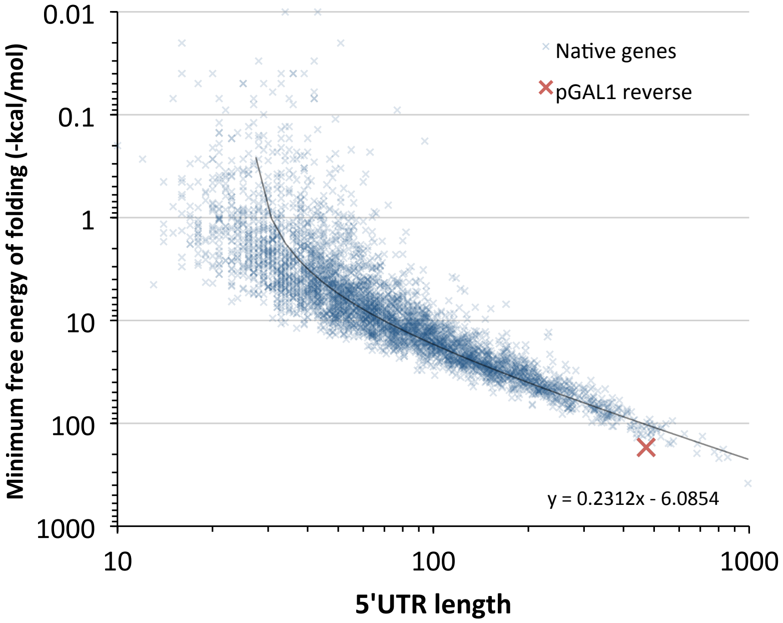


**Figure 4.12:** Plot of the predicted minimum free energy of folding in kcal/mol of all known yeast 5'UTRs versus their lengths. The red cross indicates the 5'UTR that arises from transcription through the reverse complement of the GAL1 promoter, as it is implemented in the TI cricuits. Note that all energies are negative. MFEs were calculated using RNAfold with settings for 30°C.

### 4.3.2 Does a shorter 5'UTR restore translation?

By measuring the mRNAs produced by transcription, we were able to determine that TI was working as intended, but that translation of the mRNA was failing. As we saw in the previous results chapter, the efficiency of mRNA translation in yeast can be significantly reduced when the 5'UTR is long and/or contains significant secondary structure.

Based on our experiments in this chapter and our knowledge from the previous chapter, we hypothesise that the large 5'UTR arising from transcription through the reverse complement of the various GAL1-based promoters (pTX, pLX, etc.) was leading to mRNAs that were not translated to a detectable degree. The exact mechanism by which the 5'UTR prevented translation was unclear, but it seems likely that secondary structure forming in the 5'UTR was completely preventing translation. This is further illustrated by our analysis of all known native 5'UTRs in yeast. As shown in Figure 4.12, the 5'UTR in our system is significantly longer than most and, even for its size, contains relatively strong structures.

We cannot exclude other mechanisms playing a role, such as inhibited nuclear export or specific sequences in the mRNA interacting with the ribosomal RNA in an adverse manner, however. We concluded that determining what was the exact cause of the failing translation would be complex and time-consuming. Thus, we decided that our efforts should instead be focused on reducing the length of the 5'UTR in these constructs, as this was the known differentiator between constructs that showed strong fluorescence and those that did not.

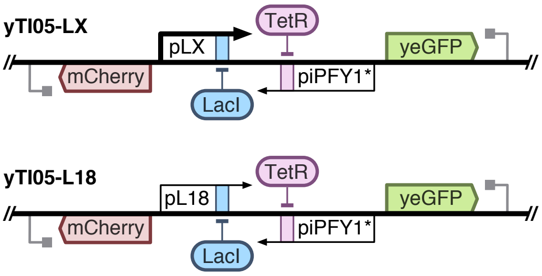
As promoter boundaries are typically poorly defined in yeast and many promoters are cloned as 500 bp sequences by default, it was challenging to select a short promoter. The minimal Cyc1 promoter is a frequently-used and well characterised promoter that is only 240 bp, however, it requires additional upstream activation sequences in order to be active, which would add complexity to the circuit and increase the length of the reverse-transcribed 5'UTR. Additionally, it contains 8 premature start codons in the reverse direction, which would need to be dealt with without affecting promoter strength. Finally, it would have had to be made repressible, further complicating the use of this promoter. For these reasons we elected not to use this promoter.

Instead we chose to use the iPFY1 promoter, which we had identified and characterised in the recent past[42]. It is a comparatively small (188 bp), medium strength promoter, that was made to be repressible by TetR. Despite its small size, it contained 4 premature start codons in the reverse direction. Thus, in order to further reduce the 5'UTR of the transcript of the facing promoter, we chose to clone the first premature start codon in-frame with the gene for green fluorescence. In this implementation, if TI leads to translated mRNAs, this design will add 44 amino acid residues to the N-terminal of yeGFP. yeGFP has frequently been used in protein fusions, so we expected this relatively small addition to have no detrimental effect on fluorescence.

Overall, the length of the 5'UTR arising from transcription through the reverse complement of iPFY1 was projected to be approximately 160 bases. This compares favourably to the 5'UTR of over 500 bases associated with the pGAL1-based promoters. Although longer than the average yeast 5'UTR of 97 bases, a length of 160 bases is not uncommon[247]. This new promoter was therefore expected to be a significant improvement.

In terms of further modifications to remove CAT motifs, no other sequence changes were needed because - remarkably - none were present in the region that would form the reverse 5'UTR. Instead, all were within the sequence that would encode the 44 amino acid polypeptide added to the reporter. This meant that as long as translation began from the first intended AUG, all others would be ignored. To ensure this occurred, the iPFY1 promoter was modified to increase recognition of this start codon, by changing the 3 bases preceding the AUG from GGA to AAA, which matches the consensus sequence in *S. cerevisiae*, where the -3 position is particularly conserved[111]. The new modified promoter was called the iPFY1* promoter.

The iPFY1* promoter was cloned in convergent orientation against a LacI repressible promoter. No PFY1-based LacI repressible promoter was available, so a pGAL1-based promoter was used. Like in previous experiments, the circuit was tested in a balanced and a biased configuration. Because the iPFY1* promoter is a medium strength promoter, the balanced configuration was achieved by using the pL18 promoter as the opposite-facing promoter. For the biased configuration, the LX promoter was used. The iPFY1* promoter was situated to drive mCherry expression, while pLX/pL18 drove yeGFP expression.



The yTI-repressor strain was used as the parent for the constructed strains, in order to allow further characterisation by changing repressor activities. Using conventional restriction enzyme cloning, the two sets of head-to-head promoters were created as shown in the accompanying diagram. The vector backbone for this system was based on pRS406, allowing integration into the

URA3 locus. The strain containing the iPFY1* promoter facing the LX promoter was named yTI05-LX, while the strain containing the iPFY1* promoter facing the L18 promoter was named yTI05-L18.

The two strains were tested by flow cytometry after overnight induction by 2% galactose in YEP media (for method see Materials and Methods **subsection 2.1.3** on page 58). Each of the strains was tested in each of three conditions: +10 mM IPTG (TetR active), +250 ng/ml ATc (LacI active) and +10 mM IPTG +250 ng/ml ATc (neither TetR or LacI active). YPC1 was included as a reference for maximum expression strength and galactose induction and YPH500 was included as an autofluorescence reference. The measured data were analysed using Matlab and are presented in **Figure 4.13** on the following page.

With the iPFY1* promoter only affecting the 5'UTR length of the yeGFP transcript, we expected any effects to manifest themselves predominantly in changes in green fluorescence levels. Moderate green fluorescence was expected for yTI05-L18 in the condition where no repressors were active, with an increase for the condition where the iPFY1* promoter was repressed by TetR. The same behaviour was expected, but at higher fluorescence levels, for the yTI05-LX strain. Green expression levels were expected to drop in conditions where LacI was active.

However, the results did not match expectations. No fluorescence was observed in any of the tested conditions. This indicates that despite the reduced length and altered sequence, the 5'UTR sequences used retained properties that prevented translation of the mRNA. The activity of the small inducer molecules was tested using a similar strain to yTE-TXLX, described previously, and these were found to be active (data not shown), which means that the results cannot be explained by assuming defective regulation of the repressors.

### Was the PFY1 promoter affected by a small modification?

During the construction of the yTI05 strains, we assumed that changing the two base pairs in the promoter to facilitate translation initiation in the reverse transcript did not affect the performance of the promoter. With the unexpected results in the previous experiment, we set out to verify that the iPFY1* promoter was performing as expected and that the small modification had not led to any unintended consequences.

In order to test the iPFY1* promoter in the most relevant context possible, the yTI05 strains were recreated, but without the LX or L18 promoter facing the iPFY1* promoter. This resulted in direct expression of mCherry from the iPFY1* promoter in an otherwise identical context. Two strains were created, one with the unmodified promoter which we named yTI06 and the second with the modified promoter named yTI07. Both strains were also tested in conditions where TetR was active, to verify that the promoters could still be repressed.

The two strains were tested by flow cytometry after overnight growth in YEPAG media (for method see Materials and Methods **subsection 2.1.3** on page 58). Each of the strains was tested in two conditions: +250 ng/ml ATc (TetR inactive) and plain media (TetR active). Glucose induction was only necessary for induction of the YPC1 control strain acting as a reference for maximum expression. YPH500 was included as an autofluorescence reference. The measurements were analysed using Matlab and are presented in **Figure 4.14** on page 135.

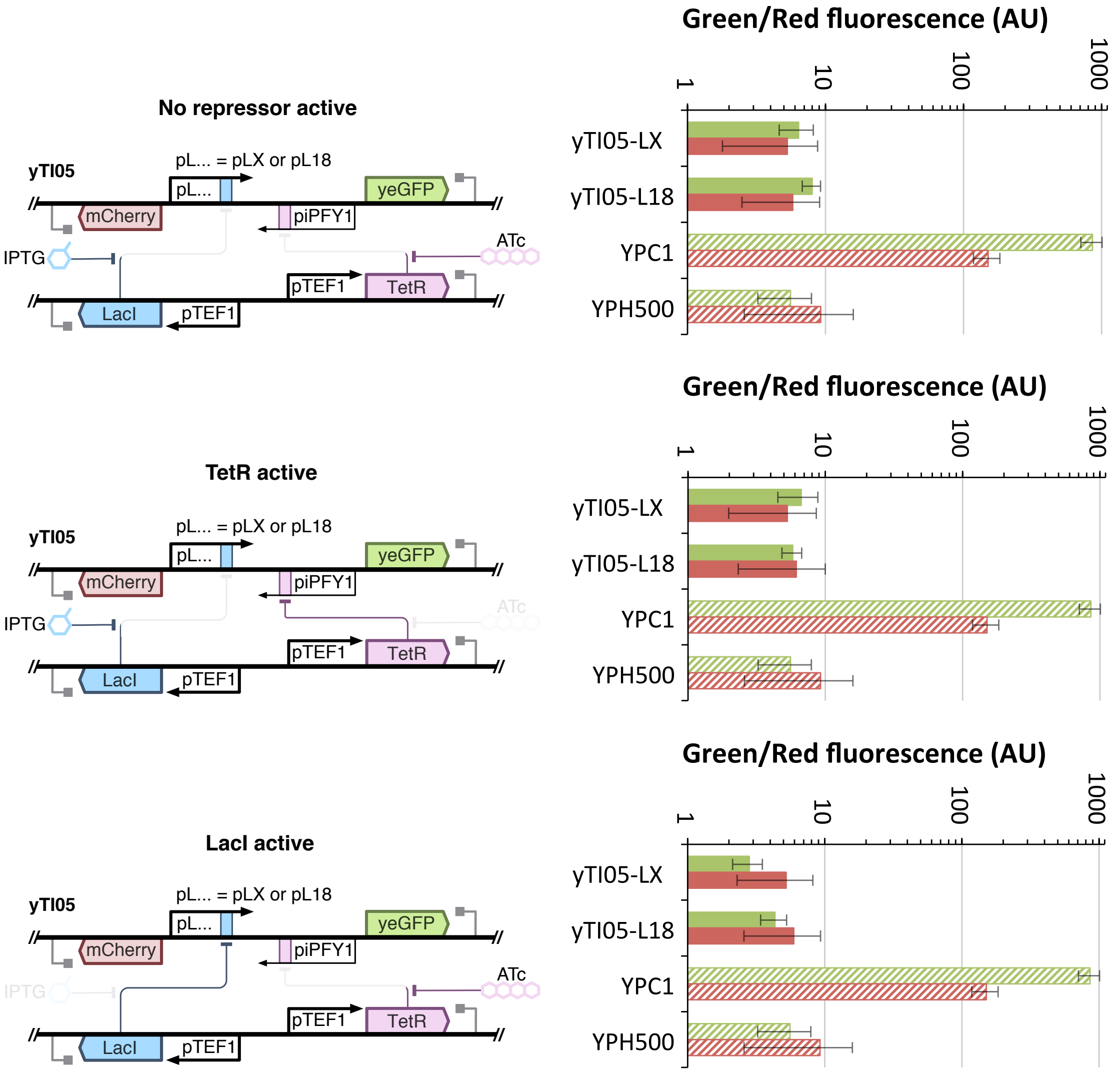
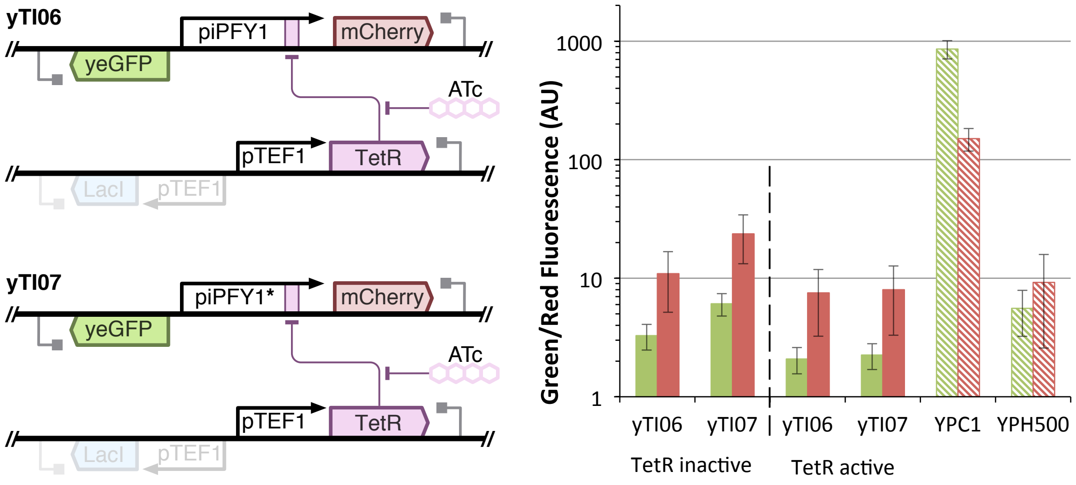


**Figure 4.13:** Direct observation of the output of convergently placed promoters with a short 5’UTR in response to small molecule inducers, as measured by flow cytometry. The short iPFY1 promoter results in a significant reduction of the 5’UTR length when transcribed by the LX/L18 promoter, compared to previously tested circuits containing a GAL1-based promoter. yTI05-LX features the strong LX promoter opposite the iPFY1 promoter, yTI05-L18 the medium strength L18 promoter. Controls (hatched); YPC1: yeGFP and mCherry expressed from the bidirectional pGAL1/10 promoter, YPH500: parental/WT strain. Left panels: circuit diagrams. Right panels: corresponding flow cytometry results. Promoter strengths were modulated through the addition of inducers (ATc at 250 ng/ml ATc and IPTG at 10 mM). All strains were tested in three conditions: both promoters at full strength (’no repressors active’), piPFY1 repressed (’TetR active’) and pLX/18 repressed (’LacI active’). Data collected on a BD FACScan flow cytometer. GAL-based promoters were induced by o/n growth in YEPAG. Error bars represent median absolute deviation of a gated population around the median of forward and side scatter.



**(a)** circuit diagram

**Figure 4.14:** The effect of a single basepair mutation in the iPFY1 promoter, as measured by flow cytometry. The promoter is modified to enhance a premature start codon in reverse direction and is designated as iPFY1* in the yTI07 strain. yTI06 contains a construct with the unmodified iPFY1 promoter. Left panels: circuit diagrams. Right panels: corresponding flow cytometry results. Promoter strengths were modulated through the addition of ATc inducer at 250 ng/ml. Both unrepressed (TetR inactive) and repressed (TetR active) conditions were tested. Controls (hatched); YPC1 control: yeGFP and mCherry expressed from the bidirectional pGAL1/10 promoter, YPH500 control: parental/WT strain. Data collected on a BD FACScan flow cytometer. Error bars represent median absolute deviation of a gated population around the median of forward and side scatter.

We expected red fluorescence for both versions of the iPFY1 promoter at levels similar to red fluorescence levels in YPC1, because iPFY1 is expected to be a promoter of moderate strength. In conditions where TetR was active, we expected significantly diminished red fluorescence. No green fluorescence was expected in any condition.

However, the results show that neither version of the promoter produces expected levels of fluorescence. Unexpectedly, the unmodified version of the promoter (yTI06) did not show red fluorescence levels that were significantly above YPH500 background. yTI07 with the modified promoter, showed slightly elevated red fluorescence, but not enough to drive biologically-relevant levels of protein in the context of a synthetic genetic circuit. We hypothesise that this is because the iPFY1 promoter is not as robust to genetic context as expected and fails to work when cloned in the construct arrangements shown here. The implementation here may differ in genetic context from the construct in which the promoter was previously characterised and the new arrangement is now somehow non-functional. The finding that PFY1-based promoters are context sensitive was also corroborated in **Figure 5.3.2** on page 178.

Unfortunately, from these findings it follows that the iPFY1* promoter is unsuitable for further transcriptional interference experiments. If the iPFY1* promoter is also inactive in the context of yTI05 constructs, which seems highly likely given these results, it means that it would have imposed no transcriptional interference on the LX/L18 promoters. In turn, this means that the conclusions from the previous experiment remain valid and that another solution must be found for the lack of translation from the mRNA produced by convergent promoters.

### 4.3.3 Can cap-independent translation initiation bypass a long 5'UTR?

The attempt to restore translation within the system by reducing 5'UTR length with a shorter promoter was unsuccessful. While this may simply be due to an unfortunate choice of a context-sensitive promoter, we decided that the TI approach would not be attractive if it only worked with very specific short promoters. Presumably the many constraints on promoter sequence would inhibit adoption of the approach by the wider community. Therefore, we instead focused our efforts on a solution that would be more generally applicable and that would allow promoters of choice to still be incorporated into the network.

Looking at the reasons why translation from mRNA in our system is not happening, the most likely candidate is that large secondary structures in the mRNA are preventing translation initiation. With a minimum free energy of -171.2 kcal/mol, the full structure of the 5'UTR of the mRNA molecules in our system with the Gal1-based promoters is several times stronger than the structures that were shown to be able completely suppress translation in the previous chapter (which were -45 kcal/mol). To bypass this potential problem, we chose to next explore the use of cap-independent translation initiation so that ribosomes could begin translation without being inhibited by 5'UTR structures that are encoded by the reverse-complement of the interfering promoters.

Cap-independent translation initiation is typically achieved through Internal Ribosome Entry Sites (IRESs). These sequences interact with ribosomal subunits and/or eukaryotic initiation factors (eIFs) to load the translation initiation machinery onto the mRNA without relying on scanning from the 5' cap as is normally required for translation. This method bypasses potential inhibitory effects from secondary structures by loading the ribosome onto the mRNA at a location close to the start codon, regardless of upstream sequence.

As briefly mentioned in the thesis introduction (see **subsection 1.5.2** on page 43), three types of IRES have been identified in *S. cerevisiae* and these are listed below:

- Short sequences that recruit the ribosome through direct complementarity with rRNA in the 18S ribosomal subunit.
- Native sequences with low secondary structure, that rely on adenine-rich stretches and require poly-A binding protein (PAB1) for their function.
- Sequences, primarily viral in origin, that rely on intricate secondary and tertiary structure for interaction with translation initiation machinery.

We selected the most active and well characterised representative from each of the categories to test in our system and these will be described below.

For the first category we selected one of the three sequences that were found to be most active in a screen of thousands of randomly generated 18 bp sequences[248]. Hits in this screen had been rigorously tested in a bicistronic-based assay including promoterless and upstream hairpin containing controls to eliminate false positives. We used hit number 47, which was confirmed to contain an uninterrupted stretch of 11 base pairs complementary to residues 689 to 699 of 18S rRNA. This was named IRES-A in our experiments.

In the second category we selected the best performing sequence in a screen of native IRESs involved in invasive growth[249]. The 5'UTRs of the selected genes had been subjected to a thorough test of IRES activity, including controls with an upstream hairpin and the reverse complement of the selected 5'UTR to exclude false positives. Of the reported hits, we selected the YMR181c IRES, which was the top performing sequence in one of the reported assays. In the same paper, it was also shown that cap-independent translation initiation increased when only the 60 bp adjacent to the start codon were used. This minimal YMR181c-based IRES was named IRES-B in our experiments. In order to preserve the native context, we cloned the first 38 bp of the native ORF in frame with the reporter gene, for a total of 98 bp.

Despite the rigorous testing, the YMR181c promoter was named in other reports as possessing cryptic promoter activity[250,251]. For this reason we included a second native IRES in our experiments: one that was identified in the 5'UTR of the GIC1 gene[249]. Although not among the strongest IRESs, it was favourable for other reasons. It was shown to be active in vegetatively growing cells[252]. This contrasts to other IRES sequences which are especially active only when cells are stressed. Additionally, according to current information it does not fall into any of the listed categories. Together, these attributes made it worthwhile to include this sequence in our experiments and it was named IRES-C.

Finally, a viral sequence was selected as IRES-D. Viral IRESs have mostly been characterised in mammalian systems. However, some, including the Hepatitis-C Virus (HCV) IRES have been shown to work in yeast[119]. A direct comparison of different viral IRESs in yeast was not available. For this reason, we chose the HCV IRES because of its long history and frequent appearance as positive control in the mammalian IRES field.

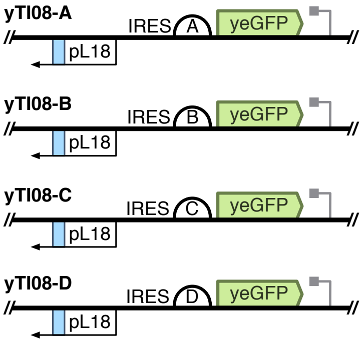
In the characterisation of various native IRES sequences, it has been reported that some IRESs are inhibited by eIF2A[250]. In order to potentially increase the activity of the selected IRESs, a deletion strain was constructed of YPH500 where eIF2A was replaced by the KanMX marker. This was done using the method described in **Figure 2.3** on page 57. This strain was used as the parental strain for all strains created in this section.

### Do the selected IRES sequences exhibit cryptic promoter activity?

In the discovery and characterisation of IRES sequences, it is challenging to distinguish between a true IRES and a sequence that contains a cryptic promoter. This is a recurring problem in the literature[250,251]. For this reason, we first set out to measure the levels of expression associated with these sequences with no promoter driving expression.

The IRES sequences were cloned upstream of yeGFP. In order to preserve the context in which these parts would later be used, they were cloned downstream of an inverted pL18 promoter. This promoter is facing away from the ORF and was shown in previous experiments to be strictly unidirectional.

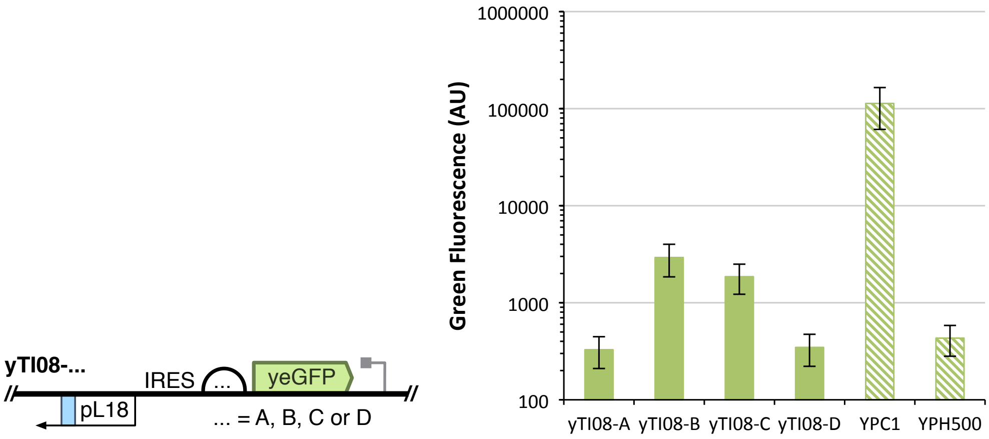
Using conventional restriction enzyme cloning, the IRES control circuits were created as shown in the accompanying diagram. The vector backbone for this system was based on pRS406, allowing integration into the URA3 locus of the eIF2A knockout strain. The four strains were named after their respective IRES: yTI08-A, yTI08-B, etc.



The four strains were assessed for green fluorescence by flow cytometry after overnight induction with 2% galactose in YEP media (see Materials and Methods **subsection 2.1.3** on page 58). Measurements were taken with the Attune NxT flow cytometer. YPC1 was included as a reference for maximum expression strength and galactose induction and YPH500 was included as an autofluorescence reference. The measured data were analysed using Matlab and are presented in **Figure 4.15**.

If the tested IRES sequences contained no cryptic promoter activity, we would have expected all strains to display autofluorescence levels comparable to the YPH500 background. However, for yTI08-B and yTI08-C we observed significantly increased expression levels, indicating that these sequences contained cryptic promoter activity. This is especially unexpected for IRES-B, since it is only 60 base pairs. It is surprising that this had not been found in previous reports. A possible explanation is that these sequences contain a core promoter region that is activated by the opposite-facing UAS of the pL18 promoter, but had gone unnoticed in other contexts.

Despite the unexpected fluorescence, the expression levels of yTI08-B and yTI08-C were only 0.023 and 0.015 fold of the expression levels of the pGAL1 promoter in YPC1. These levels were low enough that the IRESs could still be of use if they showed very high cap-independent translation initiation efficiencies. The following experiment was designed to investigate if this was the case.



**(a)** Circuit diagram

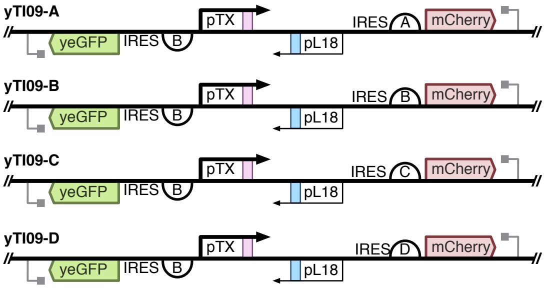
**Figure 4.15:** Characterisation of cryptic promoter activity in 4 IRES sequences as determined by flow cytometry. IRESs are situated directly preceding the yeGFP ORF, down stream of the L18 promoter facing away from the ORF. See text for the identity of the A, B, C, and D IRES. Left panel: circuit diagram. Right panel: corresponding flow cytometry results. Controls (hatched); YPC1: yeGFP and mCherry expressed from the bidirectional pGAL1/10 promoter, YPH500: parental/WT strain. The yTI08 strains are deficient for eIF2A, to increase potential activity of the IRES sequences. Data collected on the Attune flow cytometer. GAL1-based promoters were induced by o/n growth in YEPAG. Error bars represent median absolute deviation of a gated population around the median of forward and side scatter.

### Can an IRES sequence restore translation in a convergent promoter system?

To test whether IRES sequences could bypass problems associated with a long 5'UTR and initiate ORF expression directly from a location close to the start codon, we inserted the selected IRES sequences into constructs similar to those tested previously (and particularly similar to the construct in yTI03-L18TX). The number of convergent promoter configurations was limited to speed up the testing process. To maximise the mRNA output of the promoter, we chose to use the biased system, where a strong promoter (pTX) is facing a substantially weaker promoter (pL18). The strong promoter drove mCherry expression and each of the selected IRESs were placed directly upstream of the mCherry ORF.

The opposite promoter, pL18, drove yeGFP expression. IRES-B was placed directly upstream of it for symmetry and for potential further testing of the circuit (subject to results obtained from this experiment). Note that the choice for the selection of IRES-B was made before the results from the previous experiment were available. However, green fluorescence levels were not the focus of this experiment anyway, since the weaker L18 promoter was expected to be overwhelmed from transcription arising from the TX promoter.

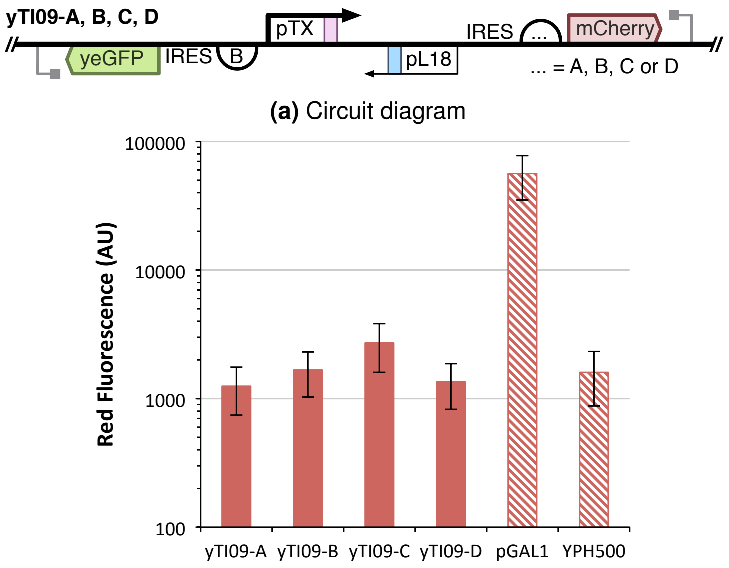
Using conventional restriction enzyme cloning, the biased convergent promoter circuits with different IRES sequences were created as shown in the accompanying diagram. The vector backbone for this system was based on pRS406, allowing integration into the URA3 locus of the eIF2A knockout strain. The four strains were named after their respective IRES: yTI09-A for the strain with IRES-A cloned directly upstream of mCherry, yTI09-B for the strain with IRES-B cloned directly upstream of mCherry, etc.



The four strains were assessed for fluorescence by flow cytometry after overnight induction with 2% galactose in YEP media (see Materials and Methods **subsection 2.1.3** on page 58). Measurements were taken with the Attune NxT flow cytometer. The p714 control strain expressing mCherry from the strong GAL1 promoter was included as a reference for maximum expression strength and galactose induction. YPH500 was included as an autofluorescence control. The measured data were analysed using Matlab and are presented in **Figure 4.16** on the next page.

From the qPCR experiments in **Figure 4.10** on page 128, we knew that the mRNA output of the TX promoter in the context of facing the L18 promoter is very comparable to the mRNA output of the GAL10 promoter in YPC1. Any IRES sequence that completely restored translation would increase mCherry expression to comparable levels with the YPC1 mCherry control, which is expressed from the GAL10 promoter.

Against our expectations, none of the tested IRES sequences showed expression levels comparable to YPC1. IRES-B and IRES-C showed increased expression levels over YPH500. Comparing these levels to the results obtained in the previous experiment, we can conclude that this is likely just due to cryptic promoter activity of these IRESs, rather than cap-independent translation initiation. Overall, this experiment shows no evidence of cap-independent translation



**(a)** Circuit diagram

**Figure 4.16:** Application of 4 IRESs in a circuit with convergently placed promoters with direct fluorescent outputs, as measured by flow cytometry. The facing promoters are biased towards strong expression of mCherry by pTX. The four IRESs are situated directly upstream of the mCherry ORF, as shown in the circuit diagram. See text for the identity of the A, B, C, and D IRES. Controls (hatched); pGAL1: yeGFP and mCherry expressed from the strong pGAL1 promoter, YPH500: parental/WT strain. The yTI09 strains are deficient for eIF2A, to increase potential activity of the IRES sequences. Data collected on the Attune flow cytometer. GAL1-based promoters were induced by o/n growth in YEPAG. Error bars represent median absolute deviation of a gated population around the median of forward and side scatter.

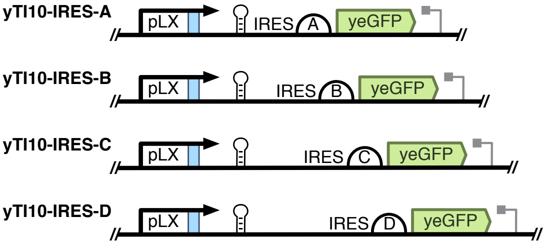
initiation from the tested IRES sequences. This was quite an unexpected result, so we decided to investigate whether the IRESs were capable of showing activity in a context without convergent promoters.

### Are IRES-A, B, C and D capable of initiating cap-independent translation?

The previous experiment called into question whether the tested IRES sequences could in fact initiate cap-independent translation. In this experiment we subjected the selected IRES sequences to a more rigorous characterisation. We created a transcription unit containing the IRES in question in front of the yeGFP ORF and upstream of that a strong RNA hairpin sequence. After transcription, the hairpin in the 5'UTR of the mRNA would completely prevent translation unless the IRES sequence allowed for cap-independent translation initiation. This is an approach that is more stringent than the commonly applied bicistronic design, which has been criticised for its inability to differentiate between IRES sequences and cryptic promoters[251–253].

Using conventional restriction enzyme cloning, the four IRES sequences were placed in between the yeGFP ORF and a strong hairpin sequence. For the hairpin sequence we used a hairpin from the qPCR experiment in the previous chapter. We used the strong tetraloop containing hairpin that was characterised in **Figure 3.8** on page 97 and was shown to suppress expression levels to near-autofluorescence levels. This assembly was placed under the control of the LX promoter.

The vector backbone for this system was based on pRS406, allowing integration into the URA3 locus of the eIF2A knockout strain. As shown in the accompanying diagram, the four strains were named after their respective IRES: yTI10-IRES-A for the strain with IRES-A cloned directly upstream of yeGFP, yTI10-IRES-B for the strain with IRES-B cloned directly upstream of yeGFP, etc.



For comparison, the yTI10 set of strains was recreated with spacer sequences instead of the IRESs. This would allow a precise assessment of the effect of the IRES sequences. The spacer sequences were made to be of identical length as their corresponding IRES. To eliminate unintended effects, the chosen spacer sequences did not contain start codons and were chosen from sections of the Zymocin resistance ORF, in order to reduce the chance that they contained any regulatory sequences or cryptic promoters.

The vector backbone and other construction details were identical to those for the yTI10-IRES series of strains. The spacer-containing constructs were named as follows: yTI10-spacer-A for the 18 bp spacer of the same length as IRES-A. yTI10-spacer-B for the 98 bp spacer of the same length as IRES-B. yTI10-spacer-C for the 212 bp spacer of the same length as IRES-C. And finally yTI10-spacer-D for the 334 bp spacer of the same length as IRES-D inserted in its place.

In addition to the yTI10-IRES and yTI10-spacer strains we created another series of strains that did not contain the hairpin inhibiting translation of the mRNA. This series of strains was called yTI11 and it shared the same naming convention as the yTI10 series e.g. yTI11-IRES-A is a strain with pLX driving expression of yeGFP with the IRES-A sequence directly upstream of it while yTI11-spacer-B is a strain with pLX driving expression of yeGFP with a spacer that is the same length as the IRES-B sequence directly upstream of it. This series of strains would allow us to see if the inclusion of an IRES sequence showed any benefits over a random sequence.

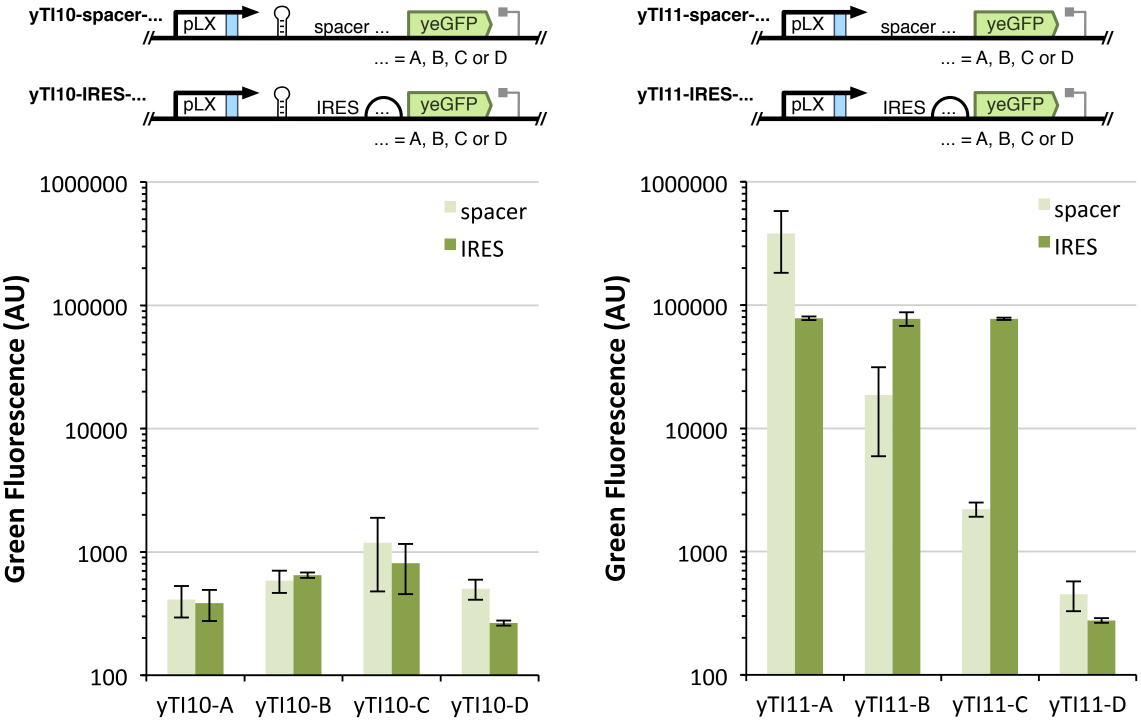


**Figure 4.17:** Stringent characterisation of 4 IRES sequences, as measured by flow cytometry. Flourescence levels were determined for the yTI11-IRES strains, that contain 4 IRES directly preceding the yeGFP ORF, driven by the strong LX promoter. In the same panel, these are compared to equivalent constructs where the IRES sequence is replaced by a spacer of the same length (yTI11-spacer). In the left panel, a corresponding set of constructs is shown that contain a strong hairpin structure in the '5UTR (yTI10-spacer and yTI10-IRES strains). See text for the identity of the A, B, C, and D IRES and spacer sequences. The yTI10 and yTI11 strains are deficient for eIF2A, to increase potential activity of the IRES sequences. Data collected on the Attune flow cytometer. GAL1-based promoters were induced by o/n growth in YEPAG. Error bars represent the standard deviation of the median fluorescence of 6 individual clones.

This is a less stringent assay than the series with a hairpin, but it was included because of the discouraging results exhibited by the IRESs in the previous experiment.

The sixteen strains were tested by flow cytometry after overnight induction with 2% galactose in YEP media (for method see Materials and Methods **subsection 2.2.3** on page 67). Six isolates were tested per strain. The measured data were analysed using Flowjo and are presented in **Figure 4.17**.

In the scenario where all IRES sequences are fully active, we would have expected expression in yTI10-IRES strains to be rescued from inhibition by the hairpin sequence, compared to expression in yTI10-spacer strains. Given the results in the previous experiment, it was not surprising that this was found to not be the case. In fact, none of the tested IRES sequences with hairpin showed a significant increase in expression compared to the respective construct containing a spacer sequence. This indicates that none of the tested IRESs are capable of completely cap-independent translation initiation in our hands.

In the less stringent assay (strains yTI11), a clear trend emerges in the strains with a spacer sequence. For every significant increase in spacer length (there is an increase of approximately 100 bases between every subsequent spacer) the expression drops by an order of magnitude.

This trend is broken, however, when the spacer sequences B and C are replaced by IRESs of the same length. This shows that these IRESs are biologically active in some way that increases expression levels compared to random sequence of the same length. This may explain why they have been reported in the literature as IRES sequences. IRES D, a sequence from the Hepatitis-C virus, does not show this behaviour. This is particularly unexpected and is further examined in the discussion section.

Together, these measurements show that some of the tested IRESs may possess biological activity. Indeed, some authors argue an intermediate mechanism exists: *5' cap-assisted* internal ribosome entry[251,252]. However, it is clear that these IRESs do not fulfil the requirement of complete cap-independent translation initiation required for restoration of translation in head-to-head promoter circuits. Therefore, we shifted our focus to a different solution in the next section.

### 4.3.4 Can inclusion of a long 5'UTR in an intron restore translation?

Conceptually, introns offer an elegant solution to the problem of translation inhibition caused by uncommonly long 5'UTRs. As detailed in the introduction (see **subsection 1.4.3** on page 39), introns are sequences in the precursor mRNA that are cotranscriptionally removed to yield a mature mRNA molecule that does not contain the intron sequence. This process, called splicing, is directed by a short signal sequence at the 5' end of the intron sequence, together with two signal sequences at the 3' end of the intron sequence. The sequence in between the 5' and 3' splice sites is not known to contain sequence elements important for splicing efficiency. This offers the possibility of introducing the 5' and 3' splice sites into an arbitrary sequence in order to remove it from the mature mRNA.

We applied this concept in an attempt to reduce the length of the 5'UTR produced in a system with two convergent promoters. By including the 5' and 3' splice sites at strategic locations in the construct, we could theoretically produce mature mRNA that was identical to the mRNA produced from a system where the promoter was placed directly upstream of the reporter gene.

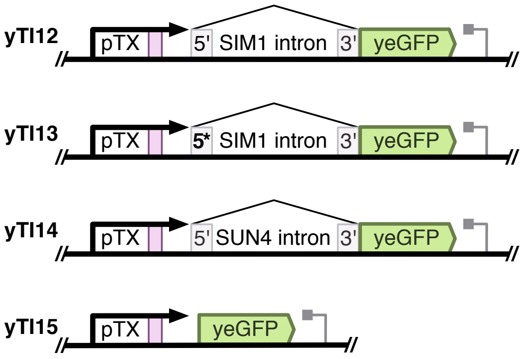
In order for this concept to work, the intron needed to span the length of the LacI and TetR repressible promoters, which is approximately 450 bp. Limited information is available on what features are important in the efficiency of intron splicing, apart from the presence of the three regulatory sequences. In order to achieve the highest possible efficiency, we chose to use introns that were as similar as possible in their native context to the intended use.

The two features that we identified as important selection criteria were intron length and location in the 5'UTR. Intron length in yeast is bimodal, with a peak around 100 bp and another around 400 bp. A significant proportion of yeast introns is of the longer type, since the second peak accounts for approximately two thirds of the total number of introns identified in yeast[77].

Selection by location in the 5'UTR was more stringent, since less than 10% of yeast introns are contained in the 5'UTR. Depending on the source, between 24 and 33 introns contained within the 5' UTR of yeast genes have been reported[77,254,255]. Encouragingly, about 80% of these are larger than 300 bp. Thus from the remaining list of 20 long introns contained in the 5'UTR of yeast genes, two were arbitrarily chosen for further characterisation. These were the 345 bp SUN4 intron and the 486 bp SIM1 intron.

### Do the SIM1 and SUN4 introns perform as expected in a non-native context?

Since context effects can have a significant impact on the performance of biological parts, we performed a characterisation of the SIM1 and SUN4 introns in a non-native context. The two introns were placed in the 5'UTR of a construct with the TX promoter driving expression of yeGFP. This would inform us if the selected introns were capable of inducing efficient splicing of highly expressed genes.



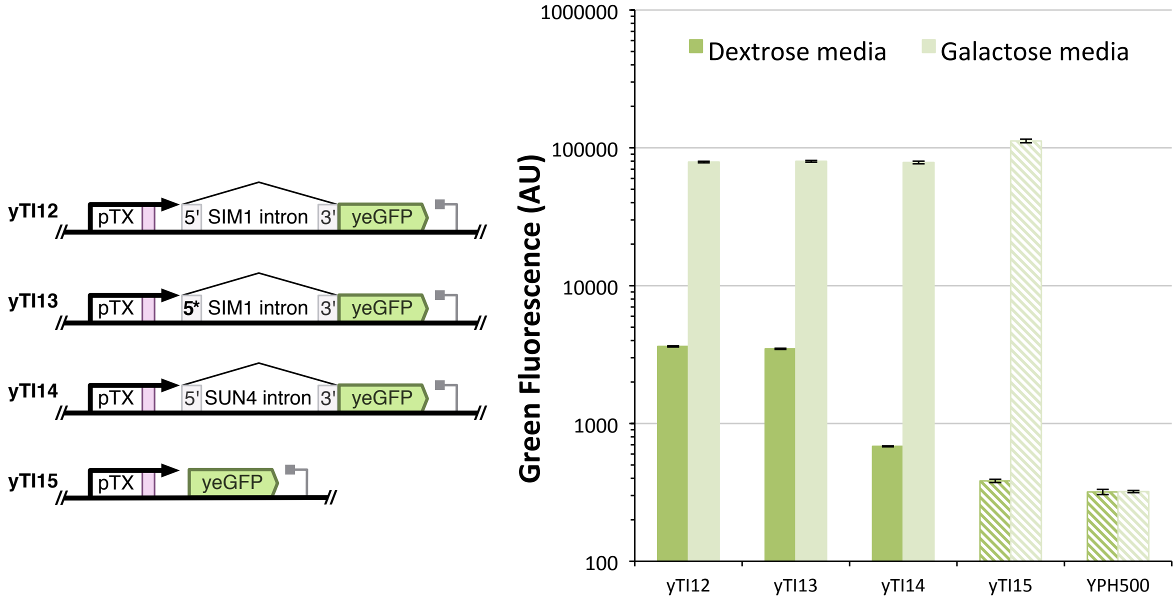
Using conventional restriction enzyme cloning, the selected introns were introduced into the pTX-yeGFP transcriptional unit, as shown in the accompanying diagram. The vector backbone for this system was based on pRS406, allowing integration into the URA3 locus of YPH500. The native SIM1 intron contained a non-canonical 5' splice site sequence. Downstream cloning limitations required that this be changed to the canonical sequence also present in the SUN4 intron. To test whether this one base pair change had an impact on the performance of the intron, it was included as a separate construct in the experiment. As a reference for expression efficiency without an intron, we also included the unmodified pTX-yeGFP transcriptional unit.

As shown, the construct containing the native SIM1 intron was named yTI12. The construct with the canonical 5' splice site in the SIM1 intron was named yTI13. The construct with the SUN4 intron was named yTI14, while the pTX-yeGFP reference strain was called yTI14.

The four strains were assessed by flow cytometry after overnight growth in GAL1 promoter repressing conditions (2% dextrose in YEP media) and overnight induction with 2% galactose in YEP media (see Materials and Methods **subsection 2.2.3** on page 67). YPH500 was included in both conditions as an autofluorescence reference control. Four measurements were performed per transformant. The measurements were analysed using Flowjo and are presented in **Figure 4.18** on the following page.

We expected expression levels for yTI12 through 14 to match the expression levels of yTI15 if the introduced introns were highly efficient. For the induced (galactose) condition, expression levels for yTI12-14 were consistently 0.70 times the levels found in yTI15. At such high expression levels, this is a modest and very acceptable reduction, indicating that the introns are capable of efficiently inducing splicing in these mRNAs. It further proves that the introduction of the canonical 5' splice site in the SIM1 intron did not have any detrimental effects on translation.

Unexpectedly, the expression levels in GAL1 promoter repressing conditions (dextrose), did show deviations from yTI15. Expression levels in repressed conditions for the SUN4 intron (yTI14) were 1.78 times the expression level of yTI15, while the SIM1 intron showed expression levels 9.5 times higher. This could indicate cryptic promoter activity or a direct interference with glucose-based repression of the TX promoter. The SIM1 intron was therefore not an ideal candidate for implementation in the final version of the circuit. Nonetheless, both introns were sucessfully characterised and shown to be capable of efficiently inducing splicing in highly expressed genes, regardless of a small modification to the SIM1 intron.



**(a)** Circuit diagrams

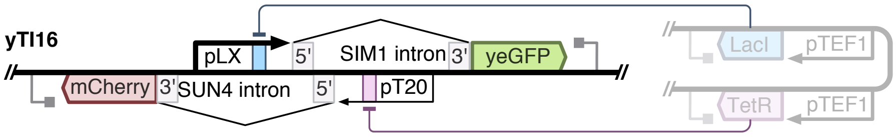
**Figure 4.18:** Determination of splicing efficiency of unusually long native introns, as determined by flow cytometry. yTI12 and yTI14 strains contain the native SIM1 and SUN4 introns in the 5'UTR of the yeGFP gene, respectively. yTI13 is identical to yTI12 with the exception of a single base-pair mutation to make the 5' splice site canonical. yTI15 (hatched) does not contain an intron and acts as a reference for maximum expression strength of the TX promoter. Data collected on the Attune flow cytometer. Strains were grown o/n in YEPAG and YEPAD for inducing and non-inducing conditions. YPH500 control (hatched): parental/WT strain. Error bars represent the standard deviation of the median fluorescence of 4 repeated measurements of a single clone.

### Implementation of introns in a circuit with convergent promoters

Despite the shortcomings of the SIM1 intron, the two characterised introns were implemented in a circuit containing promoters facing each other head-to-head. Using the SUN4 intron twice would have been a possibility, however, we prioritised fewer inverted DNA repeats in our constructs over low expression in dextrose. Presumably, if the introns performed well, we could characterise more introns to find a suitable replacement in the future.

Like previously, the convergent promoter strengths were chosen to be biased in order to maximise the output of one of the sides. In this case, the LX and T20 promoters were used in order to create the strongest bias. The SIM1 intron was cloned downstream of pLX, so as to benefit the strongest promoter in case the increased leakiness in dextrose that this intron imparted would benefit the 'strong side'.

In order to incorporate the facing promoter into the SIM1 intron, the native sequence present between the 5' splice signal and the downstream sequences (3' splice signal and branch point) was replaced by the reverse complement of the T20 promoter and vice versa for the LX promoter and the SUN4 intron. The construct is shown in the diagram below.



The circuit was constructed using conventional restriction enzyme cloning. The vector backbone for this system was based on pRS406, allowing integration into the URA3 locus. The circuit was transformed into the yTIrepressor strain, allowing the promoter strengths to be further modified through the addition of small inducer molecules. The final strain created for this experiment was named yTI16.

We tested the output of the circuit in response to small molecules modulating the activity of the repressors that were constitutively expressed in yTI16. The strain was tested in each of three conditions: +10 mM IPTG (TetR active), +250 ng/ml ATc (LacI active) and +10 mM IPTG +250 ng/ml ATc (neither TetR or LacI active). This would allow more insight into the performance of the circuit.

The strain was tested for fluorescent protein expression by flow cytometry after overnight induction with 2% galactose in YEP media (see Materials and Methods **subsection 2.1.3** on page 58). YPC1 was included as a reference for maximum expression strength and galactose induction and YPH500 was included as an autofluorescence reference. The measured data were analysed using Matlab and are presented in **Figure 4.19** on the following page.

In the biased circuit (yTI16) we primarily expected a high output of green fluorescence. The highest output was expected in the condition where only TetR was active, repressing the already weak T20 promoter. In the case where only LacI was active we expected a drop in green fluorescence and a corresponding increase in red fluorescence, because of the changed balance in transcriptional interference. For the condition where neither repressor was active, we expected a state in between the aforementioned states.

The results show, however, that none of the tested conditions resulted in significantly increased fluorescence over autofluorescence. This suggests that the intron sequences introduced into the head-to-head promoter circuit were ineffective. It is possible that splicing in this set-up is not occurring, or does not occur fully, leaving mRNA molecules that are incompatible with efficient translation, much as was the case in previous experiments.

Further experiments would be needed to determine why the intron-based approach was not a viable solution, but considering that yet again a strategy to enable TI to be applied to translated mRNAs had been unsuccessful, we decided to next look at other ways that the promising TI results we had seen at the RNA level (via qPCR) could still be linked to gene regulation without translation.

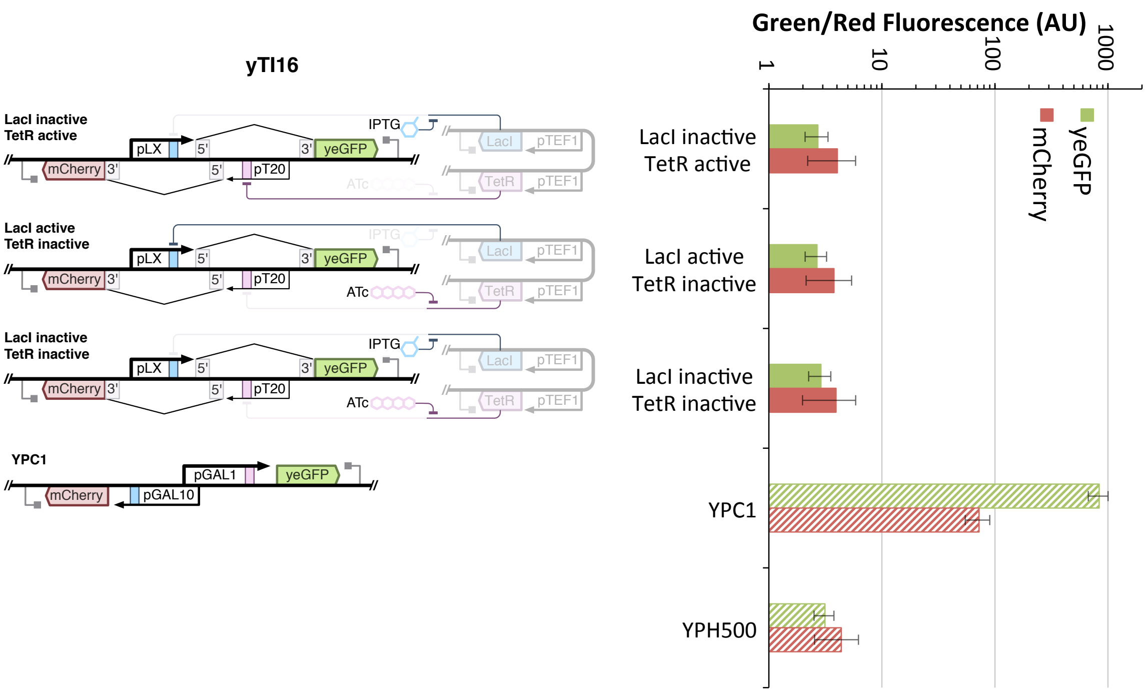


**Figure 4.19:** Direct observation of the output of convergently placed promoters enclosed in introns in response to small molecule inducers, as measured by flow cytometry. Expression strength of the promoters is biased towards yeGFP production by the strong LX promoter. The intron downstream of pLX consists of the splice-sites from the SIM1 intron, the intron downstream of pT20 consists of slice-sites from the SUN4 intron. Left panel: circuit diagrams. Right panel: corresponding flow cytometry results. Promoter strengths were modulated through the addition of inducers (ATc at 250 ng/ml ATc and IPTG at 10 mM). yTI16 was tested in three conditions: pT20 repressed ('LacI inactive, TetR active'), pLX repressed ('LacI active, TetR inactive') and both promoters at full strength ('LacI inactive, TetR inactive'). Data collected on a BD FACScan flow cytometer. GAL-based promoters were induced by o/n growth in YEPAG. Controls (hatched); YPC1: yeGFP and mCherry expressed from the bidirectional pGAL1/10 promoter, YPH500: parental/WT strain. Error bars represent median absolute deviation of a gated population around the median of forward and side scatter.

### 4.3.5 Using dCas9 to link RNA output to protein output

We envisaged that the ideal use for transcriptional interference would be to generate an extra layer of feedback loops between two opposing transcription factors in a bistable switch design. Unfortunately, while we are able to see TI at the RNA level, unless the produced mRNAs are translated, this design cannot be implemented. However, given the plethora of RNA-based regulation methods in eukaryotic gene expression and RNA-based tools in synthetic biology, we next look to see if the observed effects of TI could be linked through an RNA-level mechanism to control changes in protein expression.

To do this we chose to link the RNA outcome of a TI circuit to the control of the programmable DNA-binding protein dCas9. By having the RNA products of the TI circuit encode guide RNAs, we could program dCas9 to bind to different promoters elsewhere on the genome and repress (or potentially activate) the expression of other genes. This strategy relies on the protein component (dCas9) being constitutively expressed from elsewhere in the yeast genome and complexing with the gRNA that is produced from the TI circuit. For the creation of a TI-based bistable switch, this

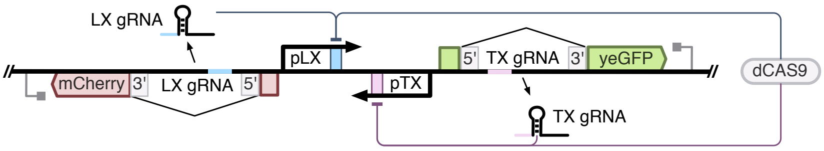


**Figure 4.20:** Bistable switch circuit design incorporating repression by dCas9. In addition to transcriptional interference from the convergently placed promoters, mutual repression is augmented by mutually repressing gRNAs encoded within introns of the opposing ORFs.

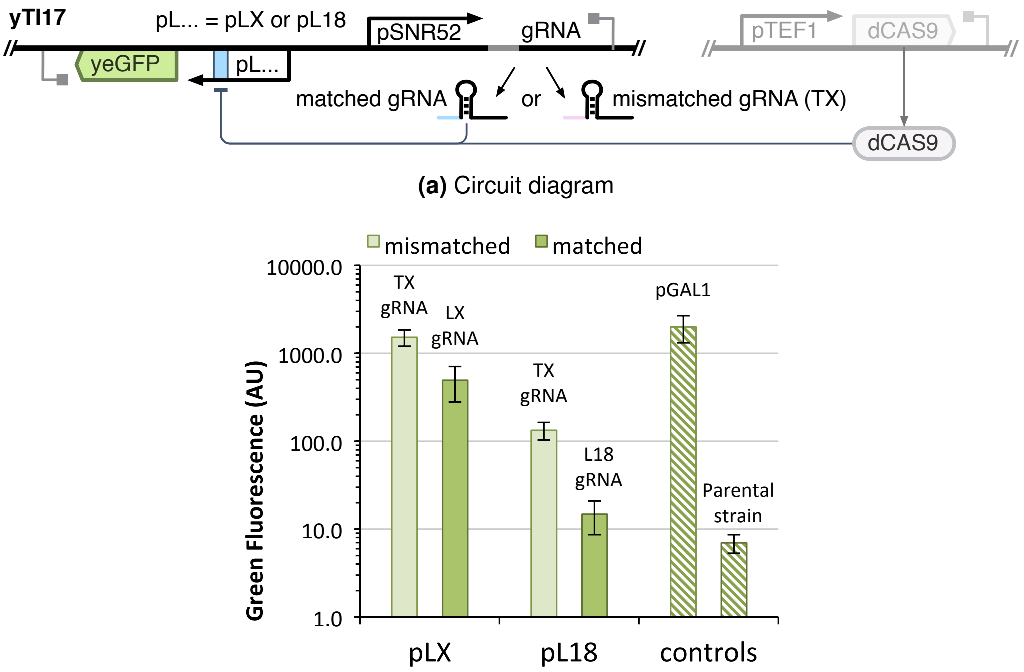
requires that two head-to-head repressible promoters (e.g. pLX and pTX) lead to the production of opposing gRNAs that compete to target dCas9 to each other and cause steric-hindrance based repression. This design is shown in **Figure 4.20**

To realise dCas9-mediated repression from the TI circuit, a major challenge first needed to be overcome. In eukaryotes, the production of uncapped RNA that acts in the nucleus is performed by RNA Pol I and Pol III, yet the regulated promoters that need to be used in transcriptional circuits, are transcribed by RNA Pol II. This therefore raises a problem - how can a regulated promoter be used to trigger the production of guide RNAs without these being capped and exported to the cytosol where they are not functional? Typically in the use of dCas9 or Cas9 in yeast, the guide RNAs are produced from Pol III promoters like the SNR52 promoter that result in large amounts of unmodified nuclear gRNA. Recent work in mammalian cells has described an approach that could solve this challenge. By placing the gRNA sequence within an intron of an mRNA transcribed by Pol II it becomes possible to produce an mRNA from a regulated promoter while in parallel producing guide RNAs in the nucleus[256]. The guide RNAs are released co-transcriptionally when the intron is spliced. Given the many similarities at the basic level between yeast and mammalian cell gene expression, we hypothesised that this approach could also work in *S. cerevisiae*.

The first step was to demonstrate that guide RNAs could be designed to target dCas9 to the promoters we intended to repress and that the repression was strong and specific. To do this we assembled a construct where the gRNA is expressed from the SNR52 Pol III promoter and the pLX and pL18 promoters express yeGFP. Constitutive expression of dCas9 from the TEF1 promoter is also provided from DNA inserted elsewhere in the genome.

We tested the output of this circuit when the gRNA sequence is designed to match the core region of the LX and L18 promoters or was designed to be a mismatch and instead target the core of the TX promoter which was not present in the constructed strains. Four constructs were produced. In the first two, yeGFP was expressed from the LX promoter and included gRNAs targeting the pLX and pTX core regions. These strains were named yTI17-LX-match and yTI17-LX-mismatch. The second two strains yeGFP expressed from the L18 promoter plus gRNAs targeting the pL18 and pTX core regions. These were named yTI17-L18-match and yTI17-L18-mismatch.

These strains were tested for yeGFP protein expression by flow cytometry after overnight induction with 2% galactose in YEP media (see Materials and Methods **subsection 2.1.3** on page 58). YPC1 was included as a reference for maximum expression strength and galactose induction and YPH500 was included as an autofluorescence reference. The measured data were analysed using Matlab and are presented in **Figure 4.21** on the next page.



**(a)** Circuit diagram

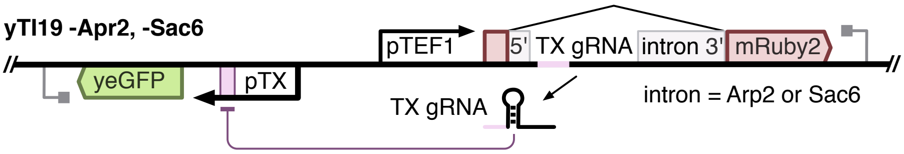
**Figure 4.21:** Characterisation of repression strength and orthogonality by targeting of dCas9 to the core promoter, as determined by flow cytometry. Four experimental conditions are shown: the LX promoter driving yeGFP expression targeted by a matching gRNA, the L18 promoter targeted by a matching gRNA and both promoters with a mismatched gRNA targeting the TX promoter. gRNA was produced from a SNR52 Pol III promoter construct included in the assembly. Data collected on a BD FACScan flow cytometer. GAL1-based promoters were induced by o/n growth in YEPAG. Controls (hatched); pGAL1: yeGFP expressed from the strong pGAL1, parental strain: YPH500 strain. Error bars represent median absolute deviation of a gated population around the median of forward and side scatter.

The results of this first dCas9 experiment were promising; the guide RNAs designed to target the core promoter regions were able to repress yeGFP expression significantly, while the mistmatched gRNA control did not lead to any repression. When the target promoter was relatively weak when in the ON state (pL18) the targeted repression by dCas9 was very effective, leading to a 9.0-fold decrease in expression down to levels close to autofluoresence. However, the repression from the much stronger LX promoter was not as effective, with only a 3.1-fold repression level seen. This suggests that while gRNA-mediated dCas9 repression is possible, it is best-suited for repression of medium to weak promoters.

With the RNA-guided method of repression confirmed in yeast, the next challenge was to ensure that our guide RNAs could be produced from Pol II transcribed promoters through the intron-splicing system used previously by others working in mammalian cells[256]. To demonstrate this, we constructed a total of 4 different genetic circuits in yeast, each containing yeGFP expressed from the strong TX promoter.

Two of the circuits were controls. Similar to the previous experiment, we included a gRNA expressed from the SNR52 promoter, targeting the TX promoter as a positive control for repression. This strain was named yTI18. The second control consisted of yeGFP driven by the TX promoter paired with a gene where pTEF1 drives mRuby2 expression to give constitutive red and green fluorescence. This strain was named yTI20.

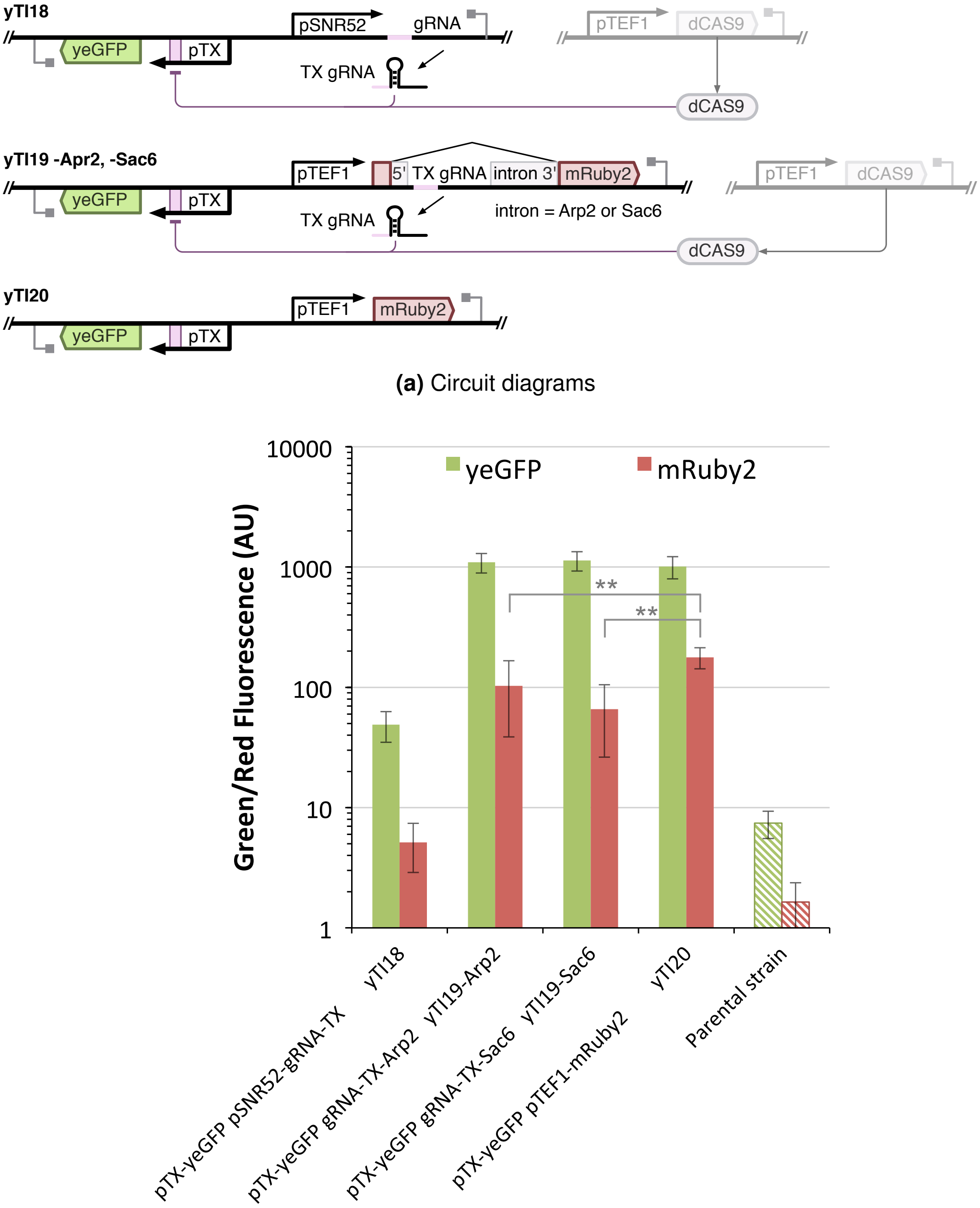
The two test constructs were designed to assess whether the intron-encoded gRNA strategy was viable. In these constructs, the pTX-targetting gRNA sequence was placed within introns within the mRuby2 gene expressed from pTEF1. Two different introns were used as the basis for the design. The *S. cerevisiae* Sac6 and Arp2 introns were selected as these are both short introns found in natural yeast genes and are found in genes that are annotated as highly-expressed genes in at least two databases. After the lack of success with the intron work in the previous section, we paid attention here to use introns known to work with genes expressed from relatively strong promoters, especially as in this work we were using pTEF1. The key splicing motifs for these introns at their 5' and 3' ends were maintained in the design, but between these motifs was placed the pTX-targeting gRNA sequence. The strains were named yTI19-Sac6 and yTI19-Arp6, after their respective original introns. A diagram of these constructs is shown below. Where relevant (i.e. in the yTI18 and yTI19 strains), dCas9 was expressed from the TEF1 promoter.



The four constructed yeast strains were tested for yeGFP and mRuby2 protein expression by flow cytometry after overnight induction with 2% galactose in YEP media (see Materials and Methods **subsection 2.1.3** on page 58). YPH500 was included as an autofluorescence reference. The measured data were analysed using Matlab and are presented in **Figure 4.22** on the next page.

Unfortunately the results from this experiment were discouraging. The control construct where the guide RNA is produced by pSNR52 showed excellent repression of pTX (22-fold repression). In all constructs with introns, we saw significantly less, but still sufficient expression of mRuby2, consistent with adequate expression of this gene and effective splicing of the intron within. However, in neither the Arp2 nor the Sac6 intron design, did we see any decrease in yeGFP expression consistent with dCas9-mediated repression of pTX. In fact the yeGFP levels remain unchanged compared to the positive control.

These results indicate that despite medium-to-high gene expression and effective intron splicing, no guide RNAs are being produced that can be used by dCas9 in our yeast cells. Could the strategy of embedding gRNA sequences within introns be fundamentally unsuitable for use in yeast? Or is more work needed to optimise the sequences within the introns so that the spliced RNA can be used as a guide RNA? Ultimately at this point the strategy looked too complicated to continue with and at this point further work on TI was halted.



**(a)** Circuit diagrams

**Figure 4.22:** Characterisation of repression strength by gRNAs expressed from introns, as determined by flow cytometry. yTI18: strain expressing a matched gRNA to the TX promoter from the SNR52 Pol III promoter. yTI19-Arp2: strain expressing a matching gRNA from the Arp2 intron embedded in mRuby2 expressed from the TEF1 promoter. yTI19-Sac6: strain expressing a matching gRNA from the Sac6 intron embedded in mRuby2 expressed from the TEF1 promoter. yTI20: control strain expressing yeGFP from the strong TX promoter and mRuby2 from the moderately strong TEF1 promoter. Data collected on a BD FACScan flow cytometer. GAL1-based promoters were induced by o/n growth in YEPAG. Parent strain control (hatched): YPH500 wild-type parental yeast strain. Error bars represent median absolute deviation of a gated population around the median of forward and side scatter. Double asterisks indicate significant differences (p<0.001) with explanatory measures of effect size over 0.5.

## 4.4 Discussion

In this chapter, we set out to implement a new form of regulation - transcriptional interference - in order to generate a bistable switch circuit. Using qPCR, we showed that mRNA levels are indeed affected as expected in response to the transcriptional interference, but are not suppressed completely. However, despite sufficient mRNA levels being present, protein expression from these transcripts was completely abolished. Our first hypothesis regarding this issue was that the large 5'UTR that is introduced in the head-to-head promoter design was responsible for this issue, and so we attempted a variety of solutions. PFY1-based promoters were used, which are shorter and allow translation to start within the reverse-complement of the opposite promoter. Both of these qualities reduce the 5'UTR length significantly, yet did not yield significant protein expression. Subsequently, we implemented Internal Ribosome Entry Sites (IRESs) to allow ribosomes to bypass the 5'UTR completely. However, this attempt was unsuccessful as we were unable to show that the IRES sequences we chose were actually functional. Similarly, our attempts to resolve the problem by using intron splicing to shorten the 5'UTR after transcription also did not lead to a functioning circuit with protein expression. A final experiment where the RNA component of the switch TI was designed to generate guide RNAs for dCas9-regulated expression was also unsuccessful.

From the variety of performed experiments our conclusion is that head-to-head based transcriptional interference has the expected effects on mRNA levels, but getting these transcripts translated into protein products is not as straightforward as expected. Important new techniques and information have become available since the inception of this project, offering potential new directions for extending this work forward, which we discuss below. We also take a closer look at the experiments that were performed for this project and discuss their outcomes and implications.

### 4.4.1 Transcription and translation in a bistable head-to-head circuit design

Over several initial experiments we tested both mRNA levels and expression levels of mutually-repressive genetic constructs with head-to-head promoters. These expressed the LacI and TetR repressors and were designed to be bistable switch circuits with an added layer of feedback provided by the transcriptional interference component.

While we never saw significant protein expression from these experiments, our qPCR measurements showed that the mRNA levels in the head-to-head situation were reduced compared to promoters without interference, demonstrating that the TI system had an impact (see **Figure 4.10** on page 128). In the case where the strength of the promoters was matched, the mRNA levels were reduced to levels comparable to the strong TEF1 promoter. While this is less than the strength of expression we would expect for the promoters we used when they are unhindered, this should still be sufficient for easily-detectable protein levels. When the strength of one of the two promoters in our system was reduced, not only did the corresponding mRNA levels drop as expected, but the mRNA levels of the opposing promoter increased significantly, consistent with expectations.

In a second experiment, these results were repeated, but instead of changing the promoter strengths by using an inherently weaker promoter, we changed them through repression with LacI and TetR (see **Figure 4.11** on page 130). In the case of LacI repression, this resulted in comparable results to the first experiment. However, for TetR there was no change in behaviour. While this was initially surprising, we feel that this may be explained by the fact that the binding rates of TetR and LacI differ significantly, as determined by *in vitro* experiments. The *in vitro* measured affinities (**Kd**) of LacI[257–259] and TetR[260,261] differ by two to three orders of magnitude and TetR is expected to be the weaker binder ($K_d = 1 * 10^{-12}$ to $2 * 10^{-13}$ for LacI versus $K_d = 5.2 * 10^{-10}$ to $1.8 * 10^{-10}$ for TetR). It is conceivable therefore that TetR is constantly displaced by RNA polymerase from the opposing promoter and does not re-bind the DNA quickly enough to confer its repressive effect, unlike LacI which exerts strong binding and can act as a more forceful repressor.

In our work looking at protein expression from the potential bistable circuits and test-circuits where the repressors were replaced by fluorescent proteins, we only ever saw protein expression in strain yTI02-TXL14. In this strain we saw evidence of biologically-active levels of LacI repressor (see **Figure 4.7** on page 122). However, yTI03-L14TX, which is the equivalent circuit but with fluorescent reporters instead of repressors, did not show fluorescence. How this observation is possible remains unclear. To confirm the finding that yTI02-TXL14 produces LacI, the measurements were repeated on a new flow cytometer and led to identical results (data not shown) indicating that this was not an experimental error. It would have been interesting to include both of these strains in the qPCR experiments, but this discrepancy was not noticed until after strains had been selected for qPCR analysis and so following-up on this one inconsistency was not prioritised.

Based on the hypothesis that it is the length and structure of the 5'UTR that is causing the lack of translation, it is unclear why the iPFY1 promoter-swap did not work. Despite the indications that this promoter is very context-dependent, the experiment should have worked because translation was still expected from transcript arising from the opposing strong GAL1-based promoter, regardless of whether iPFY1 promoter was functioning. Was the 5'UTR still too long for efficient translation? The design we used gave a length of 160 bp which is not uncommon in yeast so this would not be expected. However, there may still be unidentified features in these natural 5'UTRs that allow efficient translation, which were not present in the reverse complement section of iPFY1. Alternatively, another potential explanation is that the way that the reversed iPFY1* promoter leads to the production of GFP with an N-terminal peptide leader fusion may end up preventing any green fluorescence. The functionality of the GFP with this fusion was never directly tested, so this remains an unknown aspect although highly unlikely to be the cause considering the widely-known tolerance of GFP for a wide variety of fusions.

In the previous chapter we also saw the inhibitory effect on translation initiation due to the formation of hairpins in the 5'UTR regions of mRNAs, especially when they are very close to the AUG start codon. In the designs tested here, there were indeed hairpin sequences within the generated 5'UTRs. The binding sites for TetR and LacI are palindromic so these sequences would presumably form hairpin RNA structures when transcribed. These sequences were found in both the GAL1-based and PFY1-based promoters and so could potentially be responsible for the lack of mRNA translation. However, the minimum free energy of folding for these hairpins are -22.54 and -22.61 kcal/mol respectively. Based on our findings in the previous chapter, these structures are not strong enough to have a severe impact on expression, if any.

Our investigations into the lack of translation focused almost entirely on changing the sequences of the 5'UTR regions encoded by the head-to-head promoter design in order to reduce length and folding. However, other considerations may need to be taken if this work is to be explored further. For example, if the 5'UTR is not actually the issue it may be that nuclear export is somehow affected by this design. Diminished or abolished 5' capping of the mRNA may result in the high transcript levels as determined by the qPCR, but would see almost no mRNA in the cytosol and a lack of translation. Alternatively, our head-to-head design could be significantly affecting the positioning of nucleosomes and the modification of histones at the DNA level around the promoters, thereby affecting their functioning. Intuitively, if this was the case we would expect this to manifest itself in issues at the transcriptional level as well as the translation levels, which was not the case in our results. However, it is hypothesised that some nucleosomes and the modifications of their histones play an active role during transcription in recruiting or preventing capping proteins and other post-transcriptional modification enzymes. Could it be that the new layout of the promoters in our design leads to chromatin modifications that signal the cell to not export and/or translate the resultant transcripts?

### 4.4.2 Internal Ribosome Entry Sites

The use of Internal Ribosome Entry Sites provided an attractive solution to our lack of translation and would also be a valuable tool for further yeast synthetic biology work - e.g. for making polycistronic mRNAs. However, in our efforts we were unfortunately not able to find a functioning IRES that we could use in our designs. Our experiments showed evidence that the tested IRES sequences could increase translation from long 5'UTRs, but translation of these still appeared to be 5'cap dependent, rather than an independent binding of the ribosome to a location within the mRNA. A point of major concern was that many of the sequences we tested as IRES actually showed cryptic promoter activity. In the case of the YMR181c IRES (IRES-B), we later found published evidence from others that confirmed this to be the case[250,251]. However, in the same paper that reveals this not to be an IRES, they do not mention the other sequence GIC1 (IRES-C), which we also found to have cryptic promoter activity.

The fact that we find so many of our sequences to act like cryptic promoters may be related to the way we implemented the IRES sequences. In our system, they are cloned close to the UAS of the GAL1 promoter and we tested these constructs by induction with galactose. It is therefore possible that the UAS activated a cryptic core promoter in the IRES that would have otherwise been left inactivated. Given the ability of strong UAS sequences to work with many different core promoters in past yeast synthetic promoter library work, this is one possibility. However, for this to occur would likely require the IRES sequences to also contain TATA-box like sequences within them. Interestingly, both IRES-B and C contain one or more sequences that differ by only a single nucleotide from the consensus sequence TATAWAWR. These sequences could possibly contribute to the observed cryptic promoter activity.

Perhaps the biggest mystery in the results obtained with the IRES sequences in this chapter was that the widely-used HCV IRES was not active in our hands. Activity of this IRES is undisputed. This potentially indicates that the test construct and assay we devised for the IRES work has some fundamental problem. However, it is difficult to identify this without considerable further work. Since we have not conclusively shown that IRES sequences cannot solve the issue of no translation, it would be interesting to test more IRES motifs in order to find one that works. In many of the papers on IRES motifs, the NCE102 IRES is mentioned as a particularly strong sequence, and as yet we haven't tested in our system. Additionally, a recent publication systematically tested a large number of IRES sequences in human cells and this could be the basis for a selection of further potential strong IRES motifs to test in our system[120].

### 4.4.3 Introns

The use of introns to remove the 5'UTR regions encoded by the reverse promoters, was a more complex solution to using IRES sequences, but one that we were confident could work. By including the head-to-head promoters within the introns of the opposite transcripts, the reverse complement of the promoter sequence would be spliced out and the mature mRNA would have the same 5'UTR as one produced in a normal non-TI design. However, once again, this apparent solution did not restore translation in the tested designs. Because of the size of the promoters, the introns that were required needed to be significantly larger than the average intron length in yeast, which could have been the cause. Although we tested and used unusually large introns naturally found in the yeast genome in our design, we still had to also replace most of this native intron sequence with sequence encoding promoter DNA. Although we did this in a way designed to not change any of the known splice site sequences, it is possible that the modification of the internal sequence within intron sequence rendered it non-functional. This would be the case if the natural intron contains unnannotated sequences features within the middle of the intron that are important for efficient splicing. In support of this, there is evidence in mammalian cells that long introns contain secondary structure that brings the 5' and 3' splice sites physically close together, in order to facilitate efficient splicing[77]. It is quite possible that such mechanisms also play a role in yeast and that by changing the sequence within the intron we abolish its function.

The work done with introns in this chapter only showed that they did not work to restore translation efficiency. If more time was available it would have been worthwhile to also investigate their effect at the RNA level. Having direct evidence for the functioning of splicing would be an ideal addition to this work, to confirm whether the lack of translation in this study was indeed due to the intron not splicing as expected. It would also be worthwhile mutating the sequences within the intron in the initial SIM1 and SUN4 experiments. For example, these constructs could be used for a siding-window mutation study to look at the role of the intervening sequence (and secondary structure) of RNA within a yeast intron. One further possible option would be to instead look at introns from yeast ribosomal proteins. These are the only essential introns in yeast and are overrepresented in both 5'UTR introns and long introns and so would be ideal for further study.

### 4.4.4 dCas9-based regulation

The use of dCas9 to direct regulation has greatly accelerated work on synthetic gene circuits due to the ease with which the (orthogonal) promoter target can be changed. In bacterial systems in particular, CRISPR/dCas9 regulation is increasingly common-place and represents the state-of-the-art. In eukaryotes application of this system is complicated by the fact that the guide RNA needs to be produced by an alternative mechanism to most genes because the product needs to remain in the nucleus to exert its effect.

When an appropriate RNA pol I or RNA pol III transcribed promoter is used to produce the guide RNA in eukaryotic cells, the system works reasonably well. The ability of dCas9 to orthogonally repress targeted promoters was demonstrated by our experiments here when the SNR52 promoter was used. For weak promoters (such as pL18), the steric block of dCas9 binding to the core promoter was sufficient to almost complete shut down expression. For much stronger promoters (e.g. pLX), repression is also seen but this is not complete (see **Figure 4.21** on page 149). Theoretically, if more stringent repression of expression was required, fusion of a domain that promotes chromatin condensation (e.g. Mxi) to dCas9 would be a relatively straightforward solution. A further option to change the degree of repression is to vary the DNA site targeted by the guide RNA and how strong the RNA-DNA base pairing interactions are. It was noticeable that the pTX-targeting guide RNA led to significant more repression of its target promoter in **Figure 4.22** on page 151, compared to the repression seen with the pLX-targeting guide RNA: in both cases the targets are very strong promoters. This is likely due to differences in the target sequence and how well the gRNA can bind this DNA; a characteristic that could be further optimised in the future.

Given its relative simplicity, flexibility and proven effectiveness, it was attractive for us to consider how dCas9-based regulation could be interfaced with our demonstrated RNA-level transcriptional interference. The requirement for translation from the RNA at the TI level could be removed and this also would further open up the possibility of using different promoters in future TI-based systems, as there would be no restricted sequences. Sequences like CAT acting as premature start codons would not need to be removed. Unfortunately, however, we were unsuccessful in demonstrating that we could produce guide RNAs from regulated Pol II transcribed promoters - a necessity for implementing our system. Given that it had been previously described in mammalian cells, the intron-excision based method for generating nuclear gRNAs from mRNAs was seen as the obvious route. However, this failed to work with the 2 yeast introns that we tested here. Ideally we probably should have tested more than 2 introns, especially if at least 2 would be needed to implement a TI-based toggle switch that uses dCas9. A full screen of 10+ different intron sequences naturally found in *S. cerevisiae* may reveal alternatives that do generate gRNAs as intended. However, an alternative explanation - e.g. that spliced RNA is processed in yeast in a way that makes it unusable as a guide RNA - is equally likely and could mean that all future work on this would be in vain.

Recent new research now provides an alternative method of generating the guide RNA from mRNAs transcribed from Pol II promoters. In this work, two self-cleaving RNA sequences (ribozymes) are incorporated into the mRNA sequence to excise a central gRNA region before the mRNA is exported from the nucleus[262]. This approach could enable RNA-level TI regulation to generate gRNAs to control downstream gene expression but is perhaps not as elegant as the as yet unsuccessful intron-based method, especially as each transcript produced needs to incorporate long ribozyme encoding sequences at both 5’ and 3’ ends.

## 4.5 Conclusion

This chapter sought to develop a new form of regulation at the RNA-level that could be used to introduce extra layers of feedback inhibition into genetic circuits without requiring the addition of any new parts. By placing regulated promoters in an opposing head-to-head arrangement, the strong expression of one will directly repress the other by transcriptional interference and this could be of particular use for robust bistable switch circuits. Unfortunately, despite many attempts and different strategies, this system could never be realised. When promoters were placed head-to-head they indeed did repress the RNA production from one another in a predictable and engineerable way. But once in this format it became impossible for the produced mRNAs to be translated to make expressed proteins. Even when IRES and intron-based designs were tested these failed to work, for as yet unknown reasons. Unless a fix for the lack of translation can be achieved through further work it seems that the only route forward is to couple the RNA-level system to expression through other RNA-binding mediators such as the popular CRISPR/dCas9 system.

# 5. Simultaneous Transcription Activation and Repression

## 5.1 Introduction

In the previous chapter, we looked at stabilising a bistable switch by adding mutually exclusive transcriptional interference to the circuit. However, other ways exist to increase the robustness of a bistable circuit. Adding positive auto-regulation to a network of mutually-repressive elements dramatically increases robustness and flexibility in bistable systems[263]. The simplest way to achieve this requires single transcription factors that can both activate and repress, depending on the context. Although pervasive in nature, the type of transcription factor that can both activate and repress, is rarely applied in synthetic biology circuits. Therefore in this chapter, we aim to create a modular, synthetic transcription factor system that is capable of both activation and repression of target promoters.

### 5.1.1 Benefits of simultaneous transcription activation and repression

In **subsection 1.1.1** on page 10 the challenge of complexity in synthetic biology was introduced, noting how forward engineering becomes exponentially more difficult as the number of transcription factors in a system increases. In **section 1.6** on page 47, we described theoretical efforts that have found that some of the most robust systems utilise transcription factors that are both activating and repressing, depending on the context[144].

In **Figure 5.1** on the next page we show the potential of utilising transcription factors that are both activating and repressing in designs for robust bistable switch genetic circuits. The same fundamental circuit topology is implemented twice in this diagram: once using standard TFs that can either only activate or repress (mono-regulatory) and once using TFs capable of performing both operations. Diagrams for an oscillator circuit are also shown. The impact of TFs that can both activate and repress is dramatic, reducing the number of required transcription factors by 50% in the case of a robust bistable switch. In addition to the advantages this confers with regards to modelling and forward engineering, this also reduces the burden on the host by reducing the number of proteins that need to be expressed to achieve the circuit function.

Given the potential of transcription factors that can both activate and repress in different contexts, it is not surprising that these are pervasive in nature. Perhaps the most relevant example is the RAP1 transcription factor. It is named for its function as a Repressor Activator Protein. It is involved in the regulation of many genes, and has been studied in particular for its

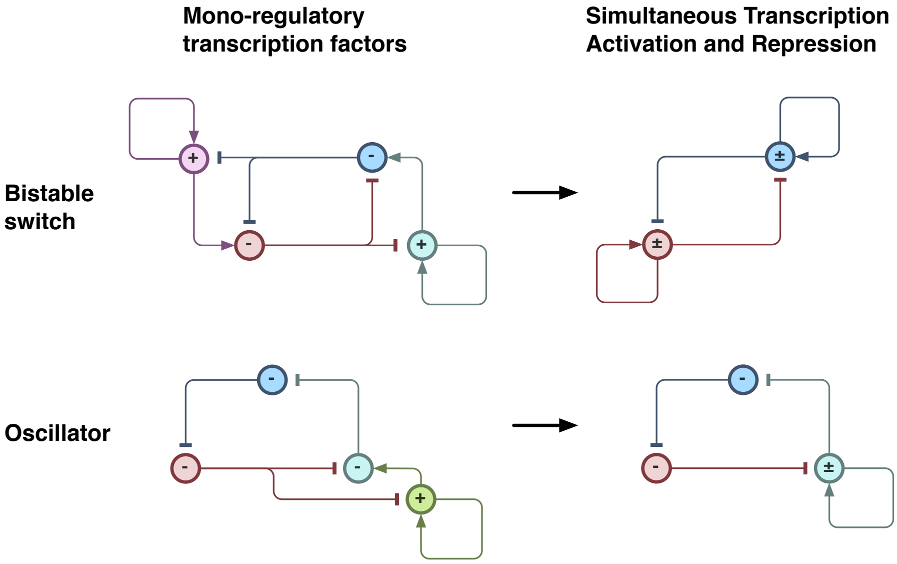

Mono-regulatory transcription factors
Simultaneous Transcription Activation and Repression
Bistable switch
Oscillator
+
-
±

**Figure 5.1:** The use of simultaneous transcription activation and repression in the simplification of genetic circuits. Diagrams show robust bistable and oscillator circuits where regulation is achieved by transcription factors binding to promoters. Left panels show minimal circuits with mono-regulatory TFs and right panels show the corresponding functionality implemented with TFs capable of both activation and repression. Arrows indicate activation, bars indicate repression.

involvement in the yeast mating-type bistable switch circuit[264].

The yeast transcription factor Mot1p is another example that associates with transcriptionally active promoters and inhibits association of a repressor that forms a transcriptionally-inactive complex with TBP to physically block formation of the PIC[265,266]. However, Mot1p was first identified as a repressor through its ability to dissociate TBP-DNA complexes and hence repress basal transcription[267]. The zinc-regulated Zap1p transcription factor is another example of a TF capable of gene induction and repression in various contexts[268].

In fact, the majority of transcription factors in the *Yeastract* database have documented interactions that are both activating and inhibiting[269]. In the vast majority of cases, a gene that is activated by a particular transcription factor is not also inhibited by it, indicating that the interactions are unique and specific. We can conclude from these data that the concept of transcription factors having both activating and repressive functions is widespread in nature. Because of this, we believe that the use of transcription factors that both activate and repress different promoters is an important and underexplored concept in synthetic biology. In this chapter we therefore propose and construct from the bottom-up the first Simultaneous Transcription Activation and Repression (STAR) transcription factors for use in yeast synthetic biology.

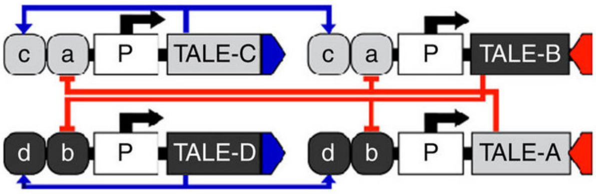


**Figure 5.2:** Implementations of a bistable circuit design using non-cooperatively binding transcription factors. This design is implemented using separate TAL-effectors carrying an activation domain (blue) or repression domain (red). Figure adapted from Lebar *et al.* 2014[270].

### 5.1.2 Requirement for cooperative binding in regulatory circuits

As we have discussed previously (see **subsection 4.1.1** on page 108), the use of programmable synthetic transcription factors is at odds with design requirements that stipulate the use of cooperativity in the creation of robust biological circuits. As we described, this requirement can be circumvented by the use of additional feedback loops in these designs. Here, we highlight an example that is particularly relevant to the intended application. **Figure 5.2** shows the implementation of a bistable switch circuit exclusively relying on non-cooperatively binding TAL-effectors. This design is analogous to the mono-regulatory TF bistable switch design presented in **Figure 5.1** on the preceding page.

The design was shown to results in robust bistable behaviour, demonstrating that the required behaviour can indeed be attained through the use of TFs incapable of cooperative binding[270]. At the same time, the figure shows the complexity of the approach, with four transcription units needed to implement the core functionality and several more to implement output and input functionality. Here, we aim to reduce the complexity in an equivalent bistable circuit that implements TFs capable of simultaneous transcription activation and repression, halving the number of transcription units required to implement the core functionality of the switch.

### 5.1.3 TAL-effectors

In **section 1.6** on page 47, we briefly discussed how Transcription Activator-Like Effectors (TALEs) have offered an attractive solution to the orthogonality challenge that limited synthetic biology in its first decade. In biology, many DNA binding proteins exist, but the overwhelming majority have evolved in a very bespoke manner. Based on their structure it is very challenging to predict what sequence they bind and it is similarly very ambitious to try to construct one that binds to a predetermined sequence. Zinc-finger proteins have offered some improvement in this respect, binding 3-bp sequences in a modular fashion with some degree of predictability. However, the real breakthrough in programmable targeting of DNA came with TAL-effectors.

As is shown in **Figure 5.3** on the following page, TALEs are extensively modular proteins that interact with DNA with single nucleotide resolution. They consist of a series of repeated modules that are capped by an N- and C-terminal domain. Each of the modules interacts with a single nucleotide of DNA, conferring DNA binding specificity to that nucleotide only. The complete TAL-effector protein assumes the helical shape of the DNA, with the modules reaching into the major groove of the double helix to interact with the bases of the DNA.

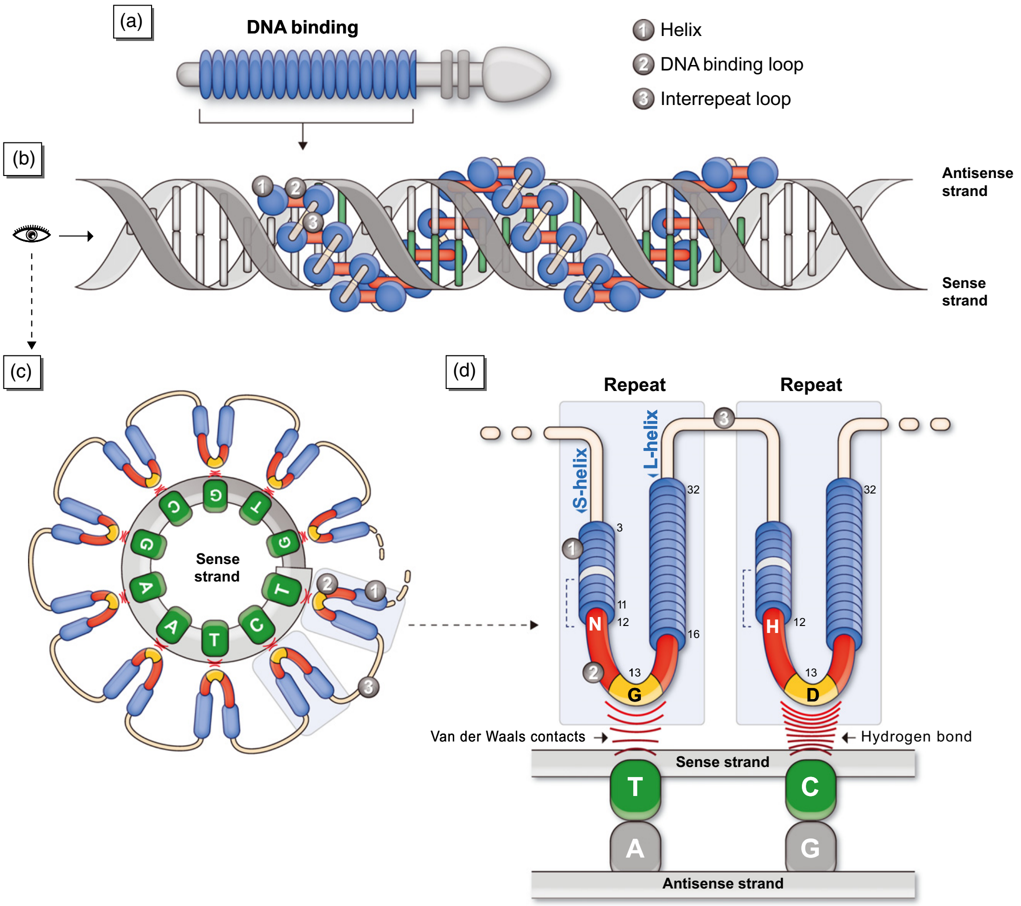


**Figure 5.3:** General structure of TAL-effectors and their mechanism of DNA binding. Note how the TALE forms a super-helical structure that nests into the major groove of the DNA. Blue regions represent identical regions of the repeated modules. The red and yellow regions indicate amino acid residues conferring the specificity of binding to nucleotides. Interaction takes place exclusively with the bases that are part of the sense strand of the DNA. Figure taken from De Lange *et al.* 2014[271].

Within the 34 residue repeated module, only one amino-acid is responsible for interacting directly with the nucleotide in question and conferring binding specificity to it. Another stabilises this interaction in a bespoke manner. These two amino acids are referred to as the Repeat Variable Domains (RVDs). In **Figure 5.4** on the next page, different RVDs are shown with their corresponding nucleotide binding specificity. It is immediately obvious that many RVDs show some binding affinity to multiple nucleotides. G, C and A residues have at least one RVD that is specific with a high affinity. Thymine, on the other hand is generally bound with only modest affinity and some aspecificity to other bases. The reason for this is that hydrogen bonds can be formed with some of the bases, but no RVDs have been reported that form a hydrogen bond with thymine. Instead the interaction relies on the relatively weak Van der Waals interactions. However, since targeted sequences typically do not consist purely of thymine residues this is not a major problem. Yet this does need to be taken into account when selecting TALE targeting sequences.

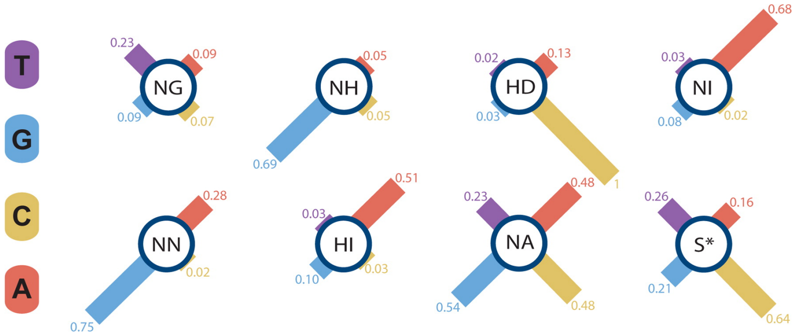


**Figure 5.4:** Illustration of the binding specificities of various RVDs to the four nucleotides. The affinities are normalized to the strongest reported affinity which is between RVD HD and the Cytosine base. The NA RVD has high specificity to all nucleotides allowing it to be incorporated when the exact base at a specific position in a sequence is not known. Figure from Moore *et al.* 2014[272].

By stringing together the right combination of RVDs, a TAL-effector can be created that targets the desired sequence. While this is a simple concept, creating the required ORF to encode a TALE is a considerable effort as the DNA sequence of a TALE gene is highly repetitive due to the repetitive nature of the protein. This makes TALE gene synthesis almost impossible and any assembly technique involving PCR or Gibson Assembly very challenging. To solve this DNA assembly challenge, several kits and techniques have been developed. In the work herein, we used a kit developed by the Voytas lab that uses a Golden Gate-based method for DNA assembly, and was one of the first kits to become widely available and popular in the early days of TAL-effector technology[18]. More information on the construction of TAL-effectors can be found in **subsection 2.1.4** on page 60 in the Materials & Methods chapter.

### 5.1.4 Orthogonal repression by TAL-effectors through steric hindrance

TAL-effectors were originally discovered as bacterial pathogen proteins that could activate target genes in plant cells. However, in our proposed system TALEs need to be able to both activate and repress transcription, and they need to be able to do so in yeast, rather than plants. We defined the basis for this work in 2012 when we showed that targeting a TAL-effector to the core promoter region of a yeast constitutive promoter leads to repression[42]. This result is shown in **Figure 5.5** on the next page and in this implementation the repression relies on steric hindrance of the pre-initiation complex by the TAL-effector due to tight sequence-specific binding at the core promoter region.

In this past work published at the start of this thesis, the promoter used was the yeast PFY1 promoter. This was shown to be an ideal constitutive promoter because it is stably expressed in a large variety of conditions. As described earlier (see **section 1.3.2** on page 34) the PFY1 promoter is a TATA-less promoter, which reduces the possibility that it is subject to unknown regulation. For these reasons, it was selected as a base promoter for further engineering and taking the concept one step further, a set of orthogonal repressors for PFY1-derived promoters were produced. This unpublished work was done by Dr B.A. Blount who characterised a set of 5 PFY1 promoter variants and generated a corresponding set of 5 TAL-effector orthogonal repressors (TALORs) specific for these.

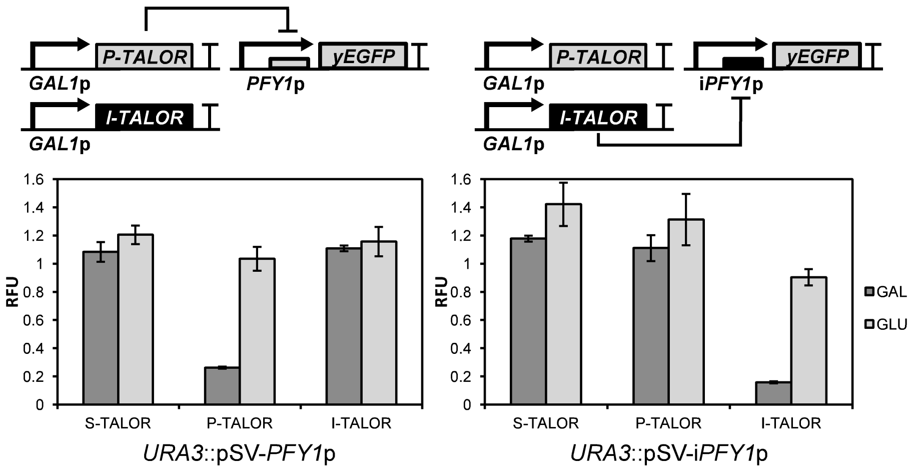
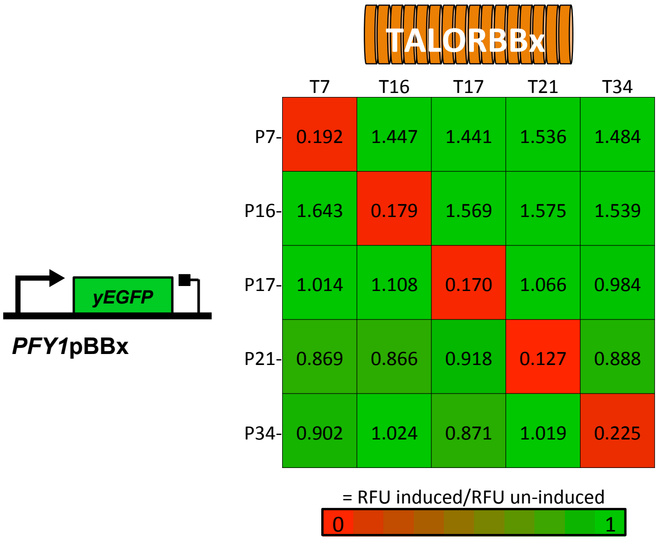


**Figure 5.5:** Orthogonal repression of PFY1 promoter variants by engineered TAL-effectors, shown as TALORs (Transcription Activator-Like Orthogonal Repressors). In this experiment, different TAL-effector variants were tested against two versions of the PFY1 promoter. One version contained a P-TALOR binding site, the second an I-TALOR binding site. The results show that only the corresponding promoter is repressed by the matching TALOR protein. The proteins remain unaffected by non-matching TALORs, proving that they are in fact orthogonal. Figure from Blount *et al.* 2012[42].

To generate these promoter-TALOR pairs, a region of 10 bp in the core promoter was randomised and shown to have little impact on the promoter expression strength. This region of sequence variation could then be targeted by specific TALORs to inhibit expression. To verify the orthogonality of the promoter-TALOR pairs, combinations of the different TALORS and promoters were tested in yeast and the unpublished results are shown in **Figure 5.6** on the next page. This shows that very little cross-reactivity is observed between the different variants, which is desirable to ensure specificity and orthogonality. For the folliwing work in this chapter, we arbitrarily chose the P7 and P21 promoters, with their corresponding TAL-effectors TAL7 and TAL21.

### 5.1.5 Engineering transcriptional activation and repression in yeast

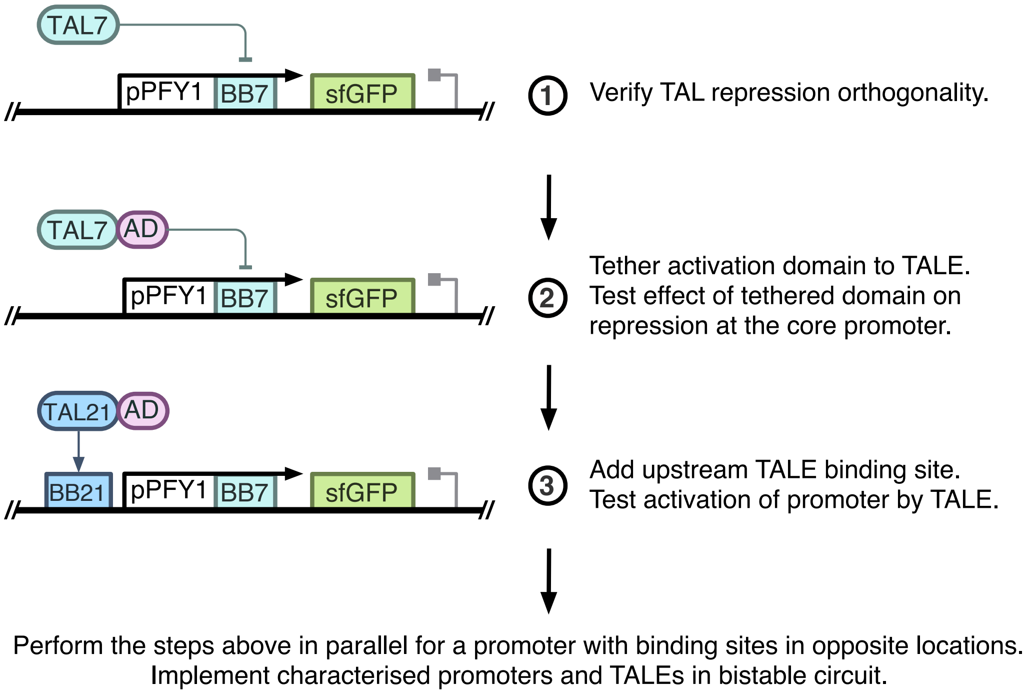
In order to activate gene expression, a transcription factor needs to bring transcription machinery into a promoter. To do this, a number of Activation Domains (ADs) can be paired with DNA-binding domains. In fact, there is a long history in yeast research of engineering DNA-binding domains to activate gene expression. In 1989, Fields and Song pioneered a technique that would become known as the yeast two-hybrid screen, where the GAL4 activation domain (GAL4-AD) was targeted to a DNA binding domain through protein-protein interactions, thereby activating the reporter gene[273]. The DNA binding domain itself binds to the upstream region of a promoter, in similar fashion to native activating transcription factors. Since its inception, this technique has been used thousands of times, so it is reasonable to state that GAL4-AD tethering is a thoroughly validated way of activating transcription. Alternative widely used activation domains are the VP16 viral activation domain and the VP64 domain which consists of four tandem copies of VP16. These have also been used extensively to activate gene expression in yeast[243,274].



|  | T7 | T16 | T17 | T21 | T34 |
|---|---|---|---|---|---|
| P7- | 0.192 | 1.447 | 1.441 | 1.536 | 1.484 |
| P16- | 1.643 | 0.179 | 1.569 | 1.575 | 1.539 |
| P17- | 1.014 | 1.108 | 0.170 | 1.066 | 0.984 |
| P21- | 0.869 | 0.866 | 0.918 | 0.127 | 0.888 |
| P34- | 0.902 | 1.024 | 0.871 | 1.019 | 0.225 |

**Figure 5.6:** Characterisation of a set of TAL-effectors and a set of promoters with a unique binding sequence embedded in the core promoter region. Each TALE is designed to only bind its corresponding sequence, laid out on the diagonal of this matrix. A lower score and red colour indicate repression of the promoter. Scores are relative to maximum expression of the unrepressed promoter, which is set to 1. Experiment by B.A Blount, unpublished.

Repression is not as commonly applied in yeast gene regulatory circuits compared to activation. When used, it is typically done through steric hindrance by DNA binding directed to the core promoter. Many domains have been identified in yeast that repress promoter activity when attracted to the promoter region through tethering to a DNA binding domain. However, they have not yet been widely applied in synthetic circuits. A possible exception is the Mxi1 repression domain, which has recently been applied in a variety of CRISPR/Cas9-based projects[19,275].



① Verify TAL repression orthogonality.

② Tether activation domain to TALE. Test effect of tethered domain on repression at the core promoter.

③ Add upstream TALE binding site. Test activation of promoter by TALE.

Perform the steps above in parallel for a promoter with binding sites in opposite locations. Implement characterised promoters and TALEs in bistable circuit.

**Figure 5.7:** Strategy to achieve orthogonal activation of promoters and repression of other promoters using TALE DNA-binding proteins fused to activating domains. Diagram shows design stategies for the PFY1 promoter activating expression of yeGFP and targeted by two orthogonal TALE domains

## 5.2 Aims and strategy

In this chapter we aim to create a system for simultaneous transcription activation and repression (STAR) of different promoters by combining the elements described earlier. The purpose of this system is to unite activating and repressing properties into a single transcription factor, and depending on the location of its binding, it will act as a transcriptional activator or repressor. This enables less convoluted, more robust and more advanced circuit designs that would theoretically impart less metabolic burden on yeast and could be extended for use in other eukaryotes. We intend to illustrate this approach in the context of the bistable switch circuit. Apart from its fundamental importance as a building block in more complex circuits, theoretical work has also found that TFs that can both activate and repress are of crucial importance when cooperative binding of transcription factors is not feasible[244].

In the proposed approach, TAL-effectors are used for their ability to be directed to target sequences. They bind the core promoter region of a series of similar promoters, and through the steric hindrance caused by their binding the promoters can be repressed. The repression can be made orthogonal between the promoters, since each TAL-effector can be targeted to a unique sequence in each of the promoters. As shown in **Figure 5.6** on the preceding page, this has already been implemented for the PFY1 promoter.

In order to allow activation, we propose to tether activation domains to the selected TALEs. Then through addition of an upstream binding site on the DNA, this TAL-effector will be able to activate transcription from the PFY1 promoter. This way, the location where the TAL-effector binds dictates its function: either activation from upstream binding or repression from binding

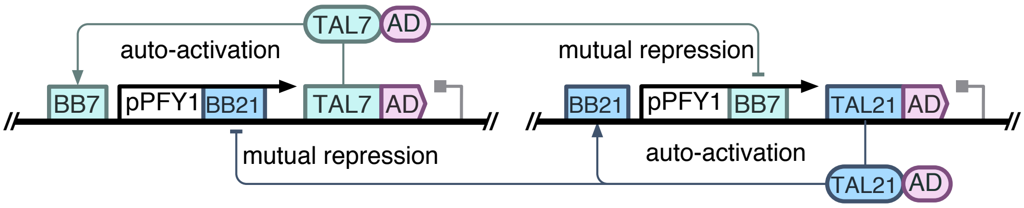


**Figure 5.8:** A robust bistable genetic switch designed to utilise simultaneous transcription activation and repression by engineered TAL-effector proteins. Each TAL-effector induces its own expression via auto-activation loops. Additionally, each represses the other TAL-effector through repressive binding at the core promoter.

to the core promoter. The discrete steps towards the realisation of this design as they were mentioned are shown in **Figure 5.7** on the previous page. A TALE will never be designed to bind to both locations on the same promoter, rather TALEs that activate one promoter by binding upstream, will repress different promoters where their target sites are in the core region.

In the context of a bistable switch, two TALEs are required, each activating its own promoter and repressing the other promoter. A diagram of this layout is shown in **Figure 5.8**. A crucial condition for this circuit to work as intended is that the repressive action of the TALE is not affected by the fusion to an activation domain. It is conceivable that the activation domain partially or completely neutralises the repressive effect of the steric hindrance at the core promoter. Thorough characterisation experiments will be necessary to identify potential issues arising from binding with the activation domain at the repression site.

It is worth noting, that although TAL-effectors were used because of their programmable nature, they did not prevent the promoter from having to be engineered in the first place. In order for the STAR system to be scalable, closely related promoter variants need to be created differing only in the position of the TALE binding sites. In one promoter, for example, the TAL21 binding site is at the activating (upstream) location while the TAL7 binding site is in the repressing (downstream) location. In a second promoter, these locations need to be opposite. This can not be achieved solely by changing the binding specificity of the TALE, without modifying the promoter. This is especially true when the system is extended beyond two promoters and two types of binding sites. In an ideal scenario, the promoters do not have to be changed at all and only the binding specificity of the TAL-effectors is changed. However, as we have described, the fundamental design of the STAR system does not allow that.

## 5.3 Results

We next describe the results of our efforts to create a transcription factor that is capable of both transcriptional activation and repression. The work of Robert Chen (from here on abbreviated with RC), an undergraduate project student in our lab, was instrumental in this effort and his contributions are attributed in the relevant experiments. According to the workflow outlined in the introduction, we started this work by doing part verification and promoter testing. Early on in this project, it became evident that the PFY1 promoter was particularly susceptible to variation in genetic context. The PFY1 promoter was abandoned altogether and exchanged for a set of TATA-box promoters. In addition we performed several optimisations in the design of the system and ultimately arrived at a promoter that could both be induced and repressed by a TAL-effector, depending on the binding location.

### 5.3.1 Part verification

The ultimate goal of synthetic biology is to be able to design a complex genetic network in silico and know that it will function as intended in vivo. Inconveniently, in the vast majority of cases a genetic circuit will not work or will not function as intended if built purely from an in silico design without any prior testing. Troubleshooting becomes more complex with increasing complexity of the circuit, and with this in mind we opted to start with basic circuits and increase complexity in incremental steps, while troubleshooting continuously.

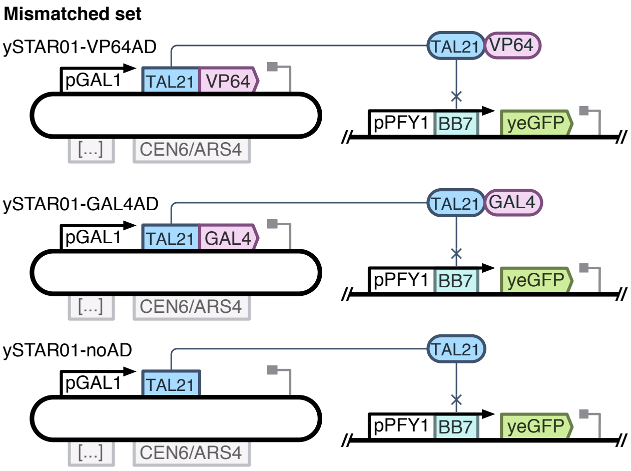
Orthogonal repression of a set of modified PFY1 promoters by a set of TAL-effectors had already been demonstrated by unpublished work in our lab by Dr B. A. Blount. The promoters and TALEs from his work were used as a foundation for more complex networks. Initial experiments focussed on verifying the orthogonality of the TAL-effector/promoter pairs and assessing the impact of the fusion of activation domains to the TAL-effector. In a purely repressive function, the addition of an activation domain is unnecessary. However, the central premise of this work is that the TAL-effector can simultaneously drive activation of a second promoter and thus needs an activation domain for that capacity.

Although the system was designed with complex networks in mind, it was impractical to perform all tests on a large number of TAL-effectors and all their corresponding binding sites. We therefore selected two constructed TALEs and worked with their respective binding sites. Given their modular design and short binding sequences, we assume that introduction of different TAL-effectors and binding sites in the future will not lead to major difficulties provided they be selected from the characterised set of orthogonal repressors due to be published by Dr B.A. Blount.

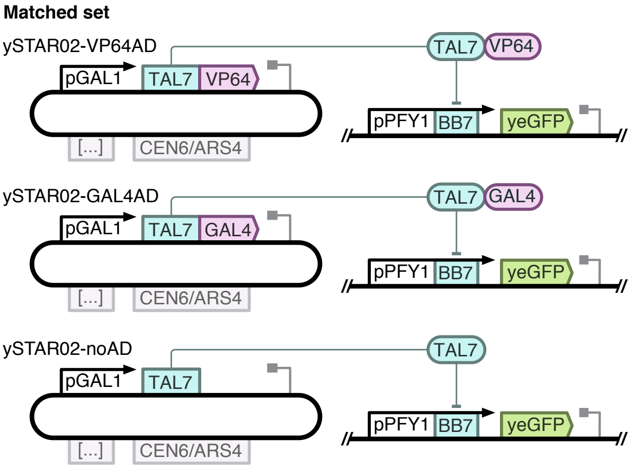
The two TALEs used in this and all further experiments are TAL7 with its corresponding BB7 binding site and TAL21 with its corresponding BB21 binding site. Assembly of the TALEs was performed using a Golden Gate based kit made available by Cermak et al.[18]. The destination vector was modified to support expression from the strong galactose-inducible GAL1 promoter. This vector also contains the CEN6/ARS4 for stable maintenance in yeast at low copy number. TALE expression is perfromed by growth in galactose.

### PFY1 promoter repression verification

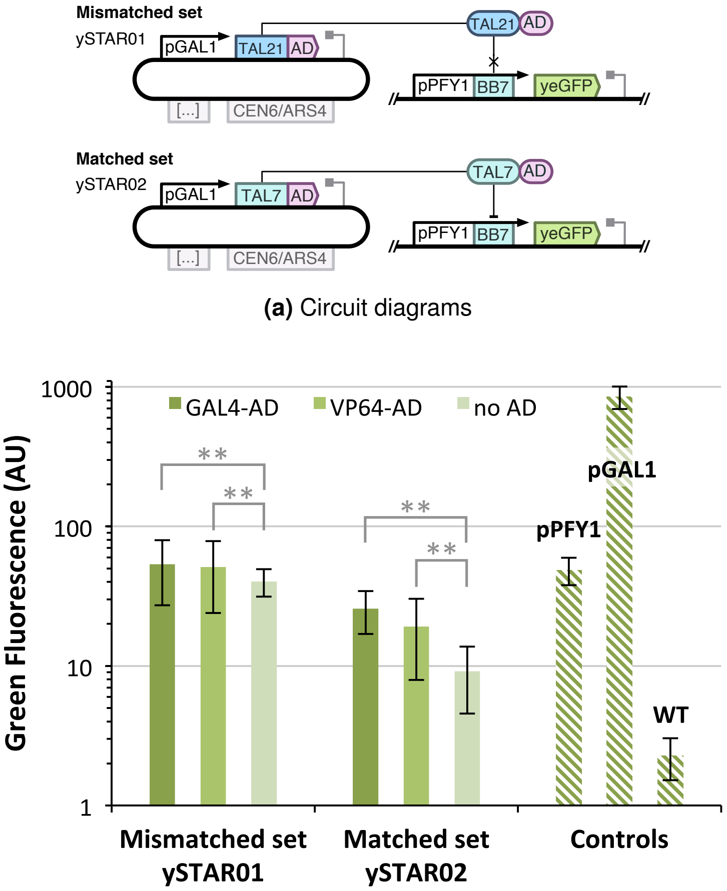
To verify the past work described earlier, the orthogonality for the selected TALEs was first checked. For this verification, two sets of yeast strains expressing TALEs with and without fused activation domains were created. One set was called ySTAR01, where TAL21 was expressed from the GAL1 promoter on a plasmid maintained extrachromosomally with the CEN6/ARS4 ori. The TALE was fused to the VP64 activation domain (ySTAR01-VP64AD), the GAL4 activation domain (ySTAR01-GAL4AD), or to no activation domain (ySTAR01-noAD). This plasmid was complemented in yeast by a construct with yeGFP expressed from a PFY1 promoter with the BB7 binding site placed in the core promoter region. This second construct was integrated into the genome in single copy. Together, this set of strains is referred to as the mismatched set, as there is a mismatch between the TALE (TAL21) specificity and the binding site sequence in the core promoter region (BB7).



The second set, ySTAR02, is a set of yeast strains that is identical to the first except TAL21 is replaced by TAL7 as the DNA-binding protein. This makes it the matched set, as TAL7 can bind to the BB7 binding site in the core promoter region of the PFY1 promoter. These two sets of strains were tested for the resultant yEGFP gene expression by flow cytometry after overnight induction by galactose in minimal media (for method see Materials and Methods **subsection 2.1.3** on page 58). As controls we also tested yeGFP expression from the original PFY1 promoter, from the strong GAL1 promoter, and from the parental untransformed strain. The resulting data were analysed using FlowJo and are presented in **Figure 5.9** on the following page.



This experiment essentially combines step 1 and 2 as defined in the workflow for this project (see **Figure 5.7** on page 165). As expected, the results show that there is no significant change in promoter output in the mismatched set when the PFY1-BB7 promoter is subjected to the presence of TAL21-noAD. This follows from the observation that the ySTAR01-noAD strain reaches the same expression level as the pPFY1 control. Expression of ySTAR02-noAD, however, is considerably lower than the pPFY1 control. This confirms that the TAL-effectors can specifically bind and orthogonally repress promoters containing their respective binding sequences, despite only a 6bp difference between the BB21 and BB7 binding sites.



**(a)** Circuit diagrams

**Figure 5.9:** Validation for TAL-effector induced repression and orthogonality. TAL effectors were targeted to a binding site (BB7) in the core promoter region of the PFY1 promoter. Activation domains (ADs) were included to evaluate interference with repression characteristics. Tested domains are the GAL4 and VP64 activation domains in addition to a TAL-effector with no AD attached. Mismatched versions of these TAL-effectors targeting a different binding site (BB21) were included as controls. Bars represent median green fluorescence in arbitrary units (AU) of a single clone. Error bars represent median absolute deviation of a gated population around the median of forward and side scatter. Double asterisks indicate significant differences ($p<0.001$) with explanatory measures of effect size over 0.5. Controls (hatched): 'pPFY1': yeGFP expressed from the PFY1 promoter, 'pGAL1': yeGFP expressed from the high strength GAL1 promoter, 'WT': wild type/parental strain BY4741.

In addition, the behaviour of the promoter was tested in the presence of activation domains fused to the TALEs. For intended system behaviour, the addition of the ADs to the TALEs should not change the repression of the promoter. However, the results show that this is not the case. Expression is slightly but significantly increased in the mismatched set when TALEs fused to an activation domain are expressed. This potentially reflects a non-specific interaction of the TALEs, leading to a general increase in expression levels across the genome.

In the matched set, the fusion of the AD led to even more dramatic effects. Repression with the VP64-fused TALE (ySTAR02-VP64AD) is significantly less strong as seen with the original TALE (ySTAR02-noAD) and repression with the GAL4 fusion (ySTAR02-GAL4AD) appears weaker still, although not significantly so. From this we conclude that activation domains do interfere with the ability of a TALE to repress when bound to the core promoter region. In this case, VP64 appears preferable to GAL4 as an activation domain as it may interfere to a lesser degree. Interference from the activation domain may need to be reduced in further optimisation experiments, since strong repression is essential for the function of many genetic circuits.

### Alternative approach to construct assembly

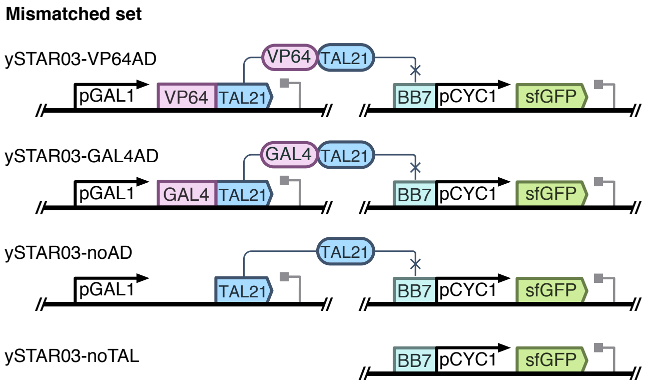
During the initial experiments, it became clear that constructs that have activation domains attached to the TALEs also show a severe negative impact on yeast growth rate and that this effect appeared to be especially pronounced in the VP64 constructs (data not shown). We speculate that this is due to non-specific binding and subsequent activation of arbitrary genes in the genome. This can have a detrimental effect on the essential processes of the cell and hence negatively affect its growth rate. To mitigate this effect, we chose to reduce the effective concentration of the TAL-effectors in our experiments.

Two ways of reducing the effective concentration of the TALEs were reduction of its promoter strength and changing from low-copy plasmids to genomic integration. Because galactose inducibility was an essential feature of the system, we opted for a genomic integration approach. In order to keep up with advancing insights in protein engineering, the activation domain was also moved from the C-terminus of the protein to the N-terminus[276]. To accommodate these changes and to allow for fast modifications in further work, at this point we adapted our assembly method to fit with the Yeast ToolKit system[8]. For more information on construct assembly using the YTK system see **section 2.2** on page 62.

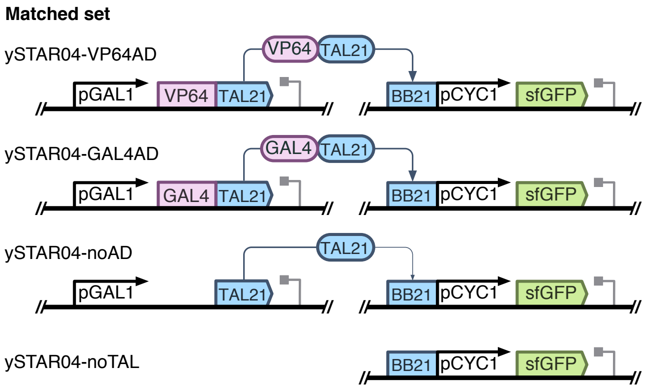
### Activation domain verification and selection

In order to assess the efficacy of the relocated VP64 and GAL4 activation domains, TALE fusions expressed from genome-integrated cassettes were targeted to a binding site located immediately upstream of the minimal CYC1 promoter. This is a commonly used method for synthetic transcription factor creation and testing (see **subsection 5.1.5** on page 163) and was considered a preferred control to the PFY1 promoter, which had never previously been used with synthetic activation and could therefore not be guaranteed to work.

Two sets of yeast strains were created for this experiment. The first set was called ySTAR03, where TAL21 was expressed from the GAL1 promoter on a construct integrated into the URA3 locus on the genome. The TALE was fused to the VP64 activation domain (ySTAR03-VP64AD), the GAL4 activation domain (ySTAR03-GAL4AD), or to a four glycine-serine (4xGS) repeat (ySTAR03-noAD). The 4xGS repeat acts both as a negative control for the activation domain and as a flexible linker connecting the main body of the TAL-effector to the nuclear localisation signal (NLS) that must be present at the N-terminus of all TALEs.



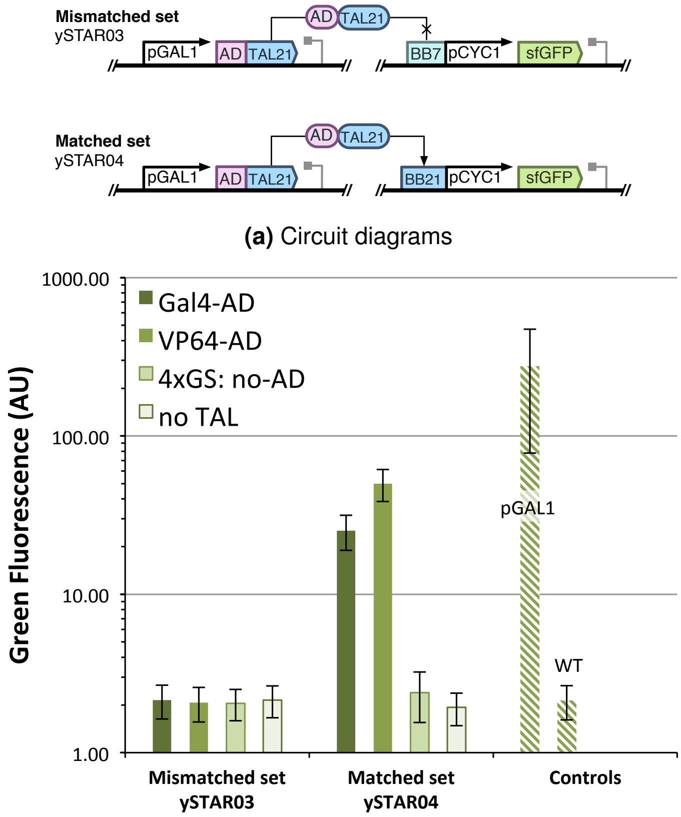
This first construct was complemented in yeast by an output construct with superfolder GFP (sfGFP) expressed from a core CYC1 promoter with the BB7 binding site placed upstream. This output construct was integrated into the genome at the LEU2 locus. To assess the basal expression of the uninduced promoter, a strain called ySTAR03-noTAL with only the second construct was also included in the set. Together, the first set of strains is referred to as the mismatched set, as there is a mismatch between the TALE (TAL21) and the upstream binding site (BB7).



The second set, ySTAR04, is a set of yeast strains that is identical to the first except the upstream binding site BB7 is replaced by the BB21 site sequence. This set is referred to as the matched set, since TAL21 can bind to the BB21 binding site upstream of the CYC1 promoter. Together, these two sets of strains and controls were tested for sfGFP expression by flow cytometry after overnight induction with galactose in minimal media (for method see Materials and Methods **subsection 2.1.3** on page 58). As controls we also tested yeGFP expression from the strong GAL1 promoter, along with the parental BY4741 strain. The measurements were analysed using FlowJo and are presented in **Figure 5.10** on the next page.

The results confirm that both the VP64 and GAL4 activation domains are capable of inducing the minimal CYC1 promoter to a moderate degree. This is demonstrated by higher expression levels of ySTAR04-GAL4 and ySTAR04-VP64 compared to ySTAR04-noAD and ySTAR04-noTAL, which are between 10 and 25 times higher. As expected, binding of the TALE with no activation domain does not lead to a significant induction of the promoter (compare ySTAR04-noAD and ySTAR04-noTAL). Additionally, lack of increased expression in ySTAR03-GAL4 and ySTAR03-VP64 confirms the specificity of the TAL21 to the BB21 binding site.

In this experiment, the activation observed using the VP64 domain fusion (ySTAR04-VP64) is 2.0-fold higher than that seen with the GAL4 domain (ySTAR04-GAL4). Taken together with the earlier result that the VP64 domain interferes less with repression when the TALE is bound to the core promoter region of the PFY1 promoter, we concluded that in this context its performance is superior to the GAL4 AD. Hence VP64-AD was selected as the activation domain to be used in further experiments.



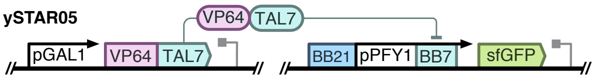
**(a)** Circuit diagrams

**Figure 5.10:** Proof of concept for TAL-effector induced promoter activation through activation domain (AD) fusion. TAL-effectors were targeted to the BB21 binding site immediately upstream of the minimal CYC1 promoter (pCYC1m). Tested domains are the GAL4 and VP64 activation domains in addition to a TAL-effector with no AD attached. Mismatched versions of these TAL-effectors targeting a different binding site (BB7) were included as controls (first set of bars). A strain with no integrated TAL-effector was included for comparison. Bars represent median green fluorescence of a single clone in arbitrary units (AU). Error bars represent median absolute deviation of a single population gated around the median of forward and side scatter. Controls (hatched); 'pGAL1': yeGFP expressed from the high strength GAL1 promoter, 'WT': wild type/parental strain BY4741. Cloning partially by RC.

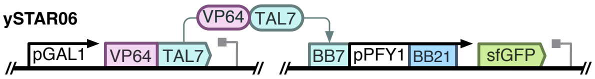
### 5.3.2 Inducible and repressible PFY1 promoter testing

With proof of concepts demonstrated for repression and activation using TAL-effectors, we were in a position to combine these two aspects to make a single dual-regulated promoter. To do this, TALE binding sites were cloned upstream of the PFY1 promoters to allow for induction in addition to the repression previously demonstrated. This resulted in two new PFY1-based promoters: one with the BB21 binding site upstream and a BB7 site in the core promoter region (called BB21-PFY1-BB7) and a second with the BB7 site upstream and the BB21 site at the core (BB7-PFY1-BB21). Four new yeast strains were built containing either of these two new promoters.

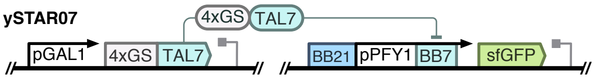
The first strain, ySTAR05, was designed to confirm that the new PFY1 promoters could still be repressed. It contained the VP64-TAL7 fusion under control of the GAL1 promoter



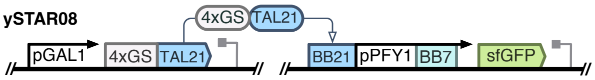
integrated into the URA3 locus of BY4741. In addition it contained the BB21-PFY1-BB7 promoter driving sfGFP expression, integrated into the LEU2 locus. TAL7 expression in this strain should lead to repression of sfGFP expression, through binding to the BB7 site in the core promoter region.



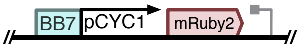
The second strain, ySTAR06, was designed to show inducibility of the modified PFY1 promoter. It contained the VP64-TAL7 fusion under control of the GAL1 promoter integrated into the URA3 locus of BY4741. In addition it contained the BB7-PFY1-BB21 promoter driving sfGFP expression, integrated into the LEU2 locus. TAL7 expression in this strain should lead to induction of sfGFP expression, through binding to the BB7 site upstream of the promoter.



The third strain, ySTAR07, was designed to show repression of PFY1 by a TALE with no activation domain. It contained the 4xGS-TAL7 under control of the GAL1 promoter integrated into the URA3 locus of BY4741. In addition it contained the BB21-PFY1-BB7 promoter driving sfGFP expression, integrated into the LEU2 locus. TAL7 expression in this strain should lead to repression of sfGFP expression, through binding to the BB7 site in the core promoter region. Based on the earlier results in this chapter, the repression is expected to be stronger than for ySTAR05, since the VP64 activation domain is replaced by the inert 4xGS linker.



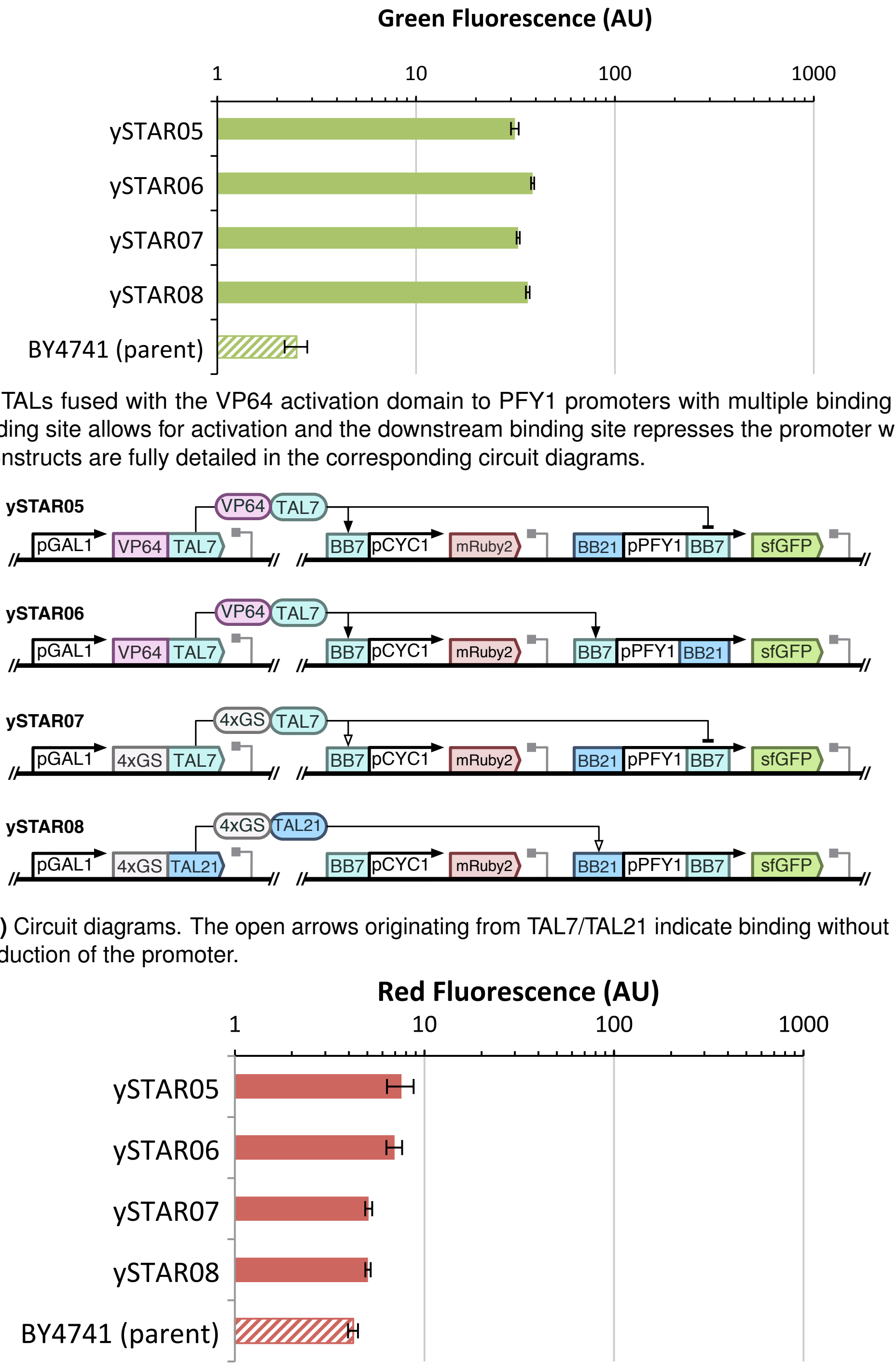
The fourth strain, called ySTAR08, was designed to show the effect of TAL-effector binding upstream of the PFY1 promoter when it is not fused to an activation domain. It contained the 4xGS-TAL21 under control of the GAL1 promoter integrated into the URA3 locus of BY4741. In addition it contained the BB21-PFY1-BB7 promoter driving sfGFP expression, integrated into the LEU2 locus. TAL21 expression in this strain should not lead to a change in expression of sfGFP, since the binding TALE does not have an activation domain and it is not binding to the core promoter.



To aid in troubleshooting of the circuit, a TAL7-inducible CYC1 promoter driving expression of a red fluorescent protein (mRuby2) was included in all strains. This is a positive control for the expression and activity of the TALE and its activation domain. These expression cassettes were cloned upstream of the new PFY1-based promoters driving sfGFP.

These four strains were tested for sfGFP and mRuby2 expression by flow cytometry after overnight induction with galactose in minimal media (for method see Materials and Methods **subsection 2.1.3** on page 58). The parental strain BY4741 was included as a negative control. The measured data were analysed using FlowJo and are presented in **Figure 5.11** on the next page.

Upon inspection of the results for green fluorescence, it became clear that none of the four strains are showing expression levels that are consistent with expectations. Qualitatively, strains ySTAR05 and ySTAR07 do match expectations in the sense that they show lower expression



**(a)** Targeting TALs fused with the VP64 activation domain to PFY1 promoters with multiple binding sites. The upstream binding site allows for activation and the downstream binding site represses the promoter when bound. The tested constructs are fully detailed in the corresponding circuit diagrams.

**(b)** Circuit diagrams. The open arrows originating from TAL7/TAL21 indicate binding without induction of the promoter.

**(c)** Targeting TALs with the VP64 activation domain (AD) to a corresponding sequence upstream of the minimal CYC1 promoter. The tested constructs are fully detailed in the corresponding circuit diagrams.

**Figure 5.11:** Characterisation of activation and repression by TAL-effectors within a single chassis. In strain ySTAR07, VP64 is replaced by a 4xGlycine-Serine linker as a negative control for the activation domain. Construct ySTAR08 contains TAL21 which is expressed but does not target the BB7 binding site. Bars represent the average of the median green or red fluorescence of three individual clones in arbitrary units (AU). Error bars represent the standard deviation of the median fluorescence of the three clones. Controls (hatched); 'BY4741 (parent)': wild type/parental strain. Cloning and data collection by RC, analysis by TW.

levels compared to ySTAR06 and ySTAR08, but the quantitative effect is minimal. In addition, ySTAR07 is expected to be a better repressor because it lacks the activation domain. However, it does not show a lower expression level than ySTAR05. Equally, ySTAR06 would be expected to give stronger induction of the promoter than ySTAR08, since it has an activation domain. However, it does not show any appreciably higher expression compared to ySTAR08.

Results for the red fluorescent positive controls also show similarly low effects. ySTAR05 and ySTAR06 were expected to show similar induction strengths to those shown in **Figure 5.10** on page 172. However, fluorescence for these constructs is only marginally higher than for ySTAR07 and ySTAR08 and likely not to be statistically significant. Given that these are positive controls, our only choice was to conclude that the results from all of this experiment are not reliable. Because of various issues during experimentation and cloning, we proposed to first perform some optimisations on the circuit design prior to attempting to replicate this experiment.

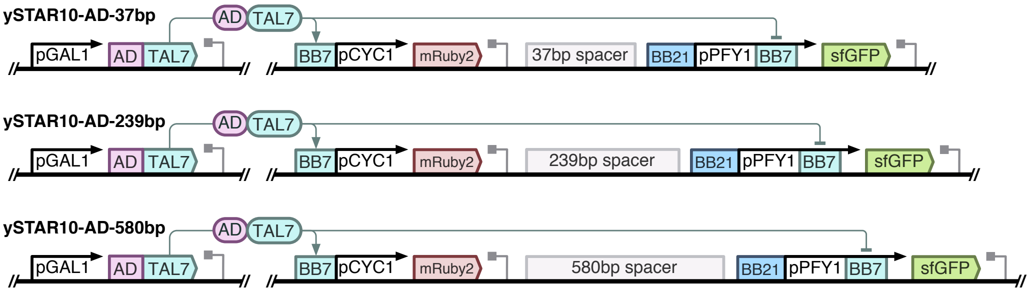
### PFY1 promoter context optimisation

With the unexpected results obtained in **Figure 5.11** on the previous page, the question was raised as to how amenable the PFY1 promoter is to the introduction of foreign sequences such as the BB21 and BB7 binding sites upstream of its sequence. It is known that regulatory elements such as promoters can be highly sensitive to their surrounding DNA sequence[11]. The fact that the promoter started showing unpredictable behaviour when it was placed in a different genetic context with the TAL binding sites placed upstream raised our awareness to this possible issue.

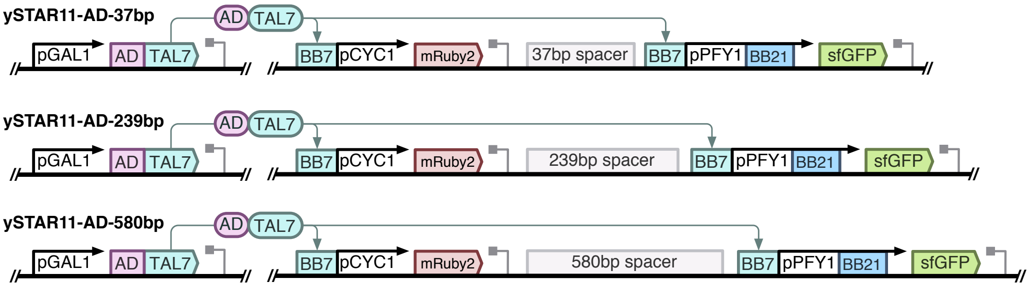
To determine if local genetic context plays a role in the expression from the PFY1 promoter, we created a series of strains similar to those in the previous experiment but with upstream sequences changed. In accordance with new insights described in **subsection 5.3.3** on page 179, the VP64 activation domain was replaced by the GAL4-AD in all of the TALE fusions.

The first set, called ySTAR10, was designed to show the context dependence of the PFY1 promoter in a repressed situation. It contained the pGAL1 promoter driving expression of the GAL4 activation domain fused to TAL7 and was integrated into the URA3 locus. As a positive control for TAL-effector activity, it contained the BB7 binding site upstream of the CYC1 promoter driving mRuby2 expression. Lastly, the PFY1 promoter driving sfGFP expression was also included, using a variant with the BB21 binding site located upstream and the BB7 site included in the core promoter.

The two reporter transcription units in this multigene construct were separated by spacer sequences. Three different spacer sequences were cloned into this design and tested, each providing three genetic contexts upstream of the PFY1 promoter. These spacers had lengths of 37, 239 and 580 base pairs. The combined multigene construct with spacers was integrated into the LEU2 locus of BY4741. The resulting 3 strains with the different spacer sequences were called ySTAR10-AD-37bp, ySTAR10-AD-239bp and ySTAR10-AD-580bp, as shown on the following page.



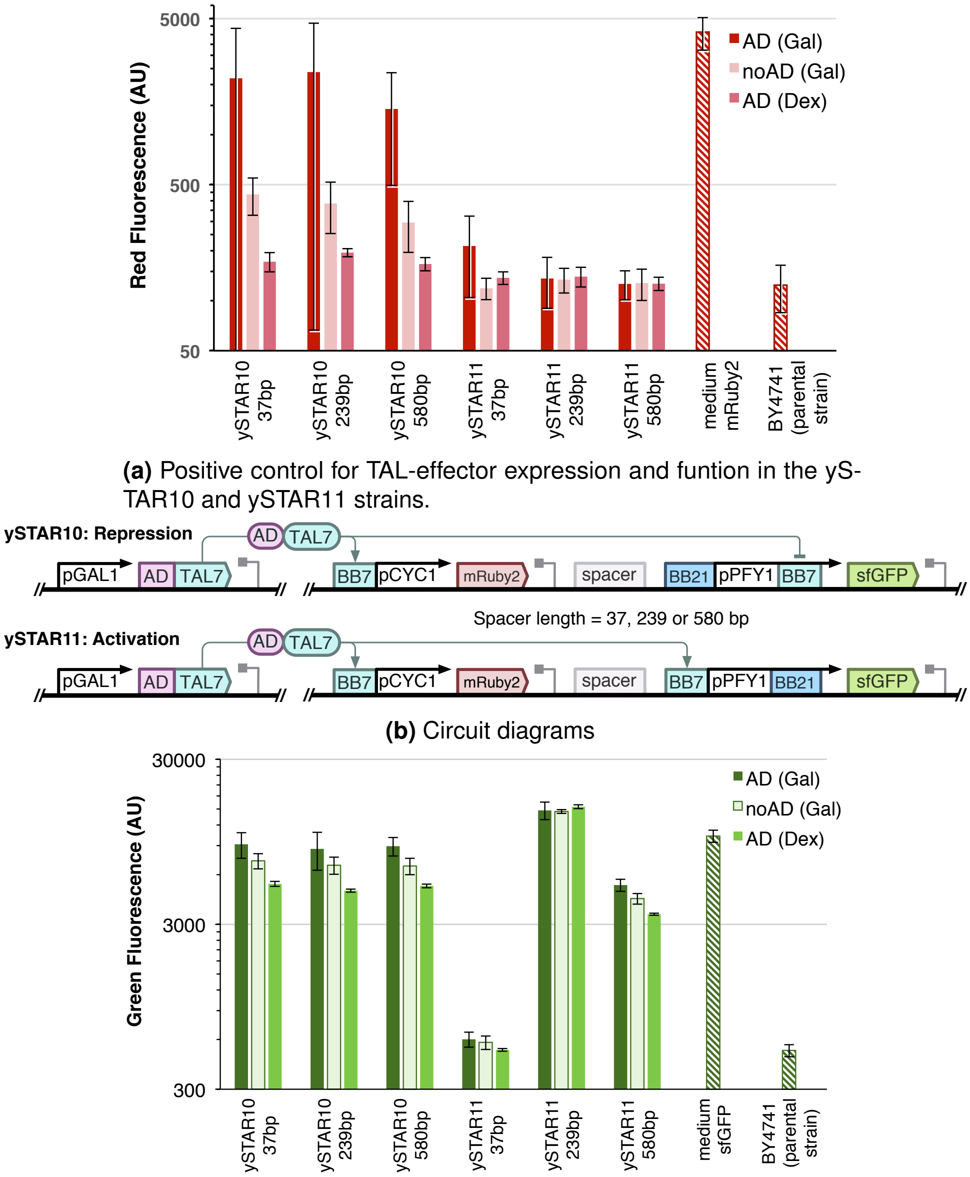
A similar set of strains was created for the characterisation of induction of the PFY1 promoter through binding to the TAL-activator to an upstream binding site. This set was identical to the set described above, except that the BB7 binding site in the core promoter and the BB21 site upstream of the PFY1 promoter were swapped to yield a promoter with the BB7 binding site upstream and the BB21 site in the core promoter. The GAL4-TAL7 present in these strains will therefore activate, rather than repress it. The 3 created strains were called ySTAR11-AD-37bp, ySTAR11-AD-239bp and ySTAR11-AD-580bp. It is worth noting that the difference between the BB7 and BB21 binding sites is variation in only 6 base pairs. Consequently, the difference between the ySTAR11 strains and their corresponding ySTAR10 counterpart is simply a total of 12 base pairs per strain.



We had determined previously that the activation domain could interfere with repression when its fused TAL-effector binds to the core promoter region. To control for this we included a version of the construct that had the activation domain replaced by a non-activating 4xGS linker. The corresponding strains were called ySTAR10-noAD-37bp, ySTAR10-noAD-239bp, ySTAR10-noAD-580bp, ySTAR11-noAD-37bp, ySTAR11-noAD-239bp and ySTAR11-noAD-580bp.

These twelve strains were assessed by flow cytometry after overnight induction with galactose in minimal media (for method see Materials and Methods **subsection 2.1.3** on page 58). Measurements were performed on the BD LSR-Fortessa X-20 cell analyzer. As controls we also included a strain with the medium strength YRA1s promoter driving expression of sfGFP, a strain with the medium strength RPL18B promoter driving expression of mRuby2 and the parental strain BY4741. As additional controls the ySTAR10-GAL4 and ySTAR11-GAL4 series of strains were also measured after overnight growth in minimal media with glucose. This suppresses expression of the TAL-effector and shows construct behaviour without any TALEs present. The measurements from this experiment were analysed using FlowJo and are presented in **Figure 5.12** on the following page.

The results show that in the ySTAR10 series of strains, the TAL-effector is being expressed and is inducing the mRuby2 positive control circuit. When the TALE is expressed without the



**(a)** Positive control for TAL-effector expression and funtion in the ySTAR10 and ySTAR11 strains.

**(b)** Circuit diagrams

**(c)** Effect of spacer length and binding site configuration on induction and repression of the PFY1 promoter by a TAL-effector.

**Figure 5.12:** Impact of spacer length and genetic context on the properties of engineered versions of the PFY1 promoter. In each TALE-expressing construct the GAL4 acitvation domain (AD) was replaced with the 4xGS linker sequence to generate a non-activating control (noAD). In addition, the uninduced expression levels were obtained in each of these constructs by incubation in glucose media (indicated with '(Dex)' in the legend, as opposed to '(Gal)' for induction in galactose media). Measurements were performed on the BD Fortessa X-20 flow cytometer. Bars represent the average of the median green or red fluorescence of cells from eight individual clones and are given in arbitrary units (AU). Error bars represent the standard deviation of the median fluorescence of the eight clones. Controls (hatched); 'medium mRuby2' and 'medium sfGFP' are expressed from the medium strength RPL18B and YRA1s promoters, respectively. 'BY4741 parental strain': autofluorescence control. Cloning and data collection by RC, analysis by TW.

activation domain, the red fluorescence drops down to near autofluorescence. This proves that the TAL-effector is functioning as expected.

According to our expectations and previous experiments, the expressed TALE should have a repressive effect in the ySTAR10 strains. The results show, however, that the opposite is true. We compared ySTAR10-AD strains in conditions favouring expression of the TAL-effector (overnight growth in galactose) to conditions where the TAL-effector is not expressed (overnight growth in glucose). This shows that expression of sfGFP is actually higher when the TAL-effector is present. This is unexpected, because previous experiments have shown that the TALE represses the PFY1 promoter when targeted to the core promoter region.

Similarly, even the set of ySTAR10-noAD strains shows higher sfGFP expression than the ySTAR10-AD strain in conditions repressing TALE expression. Expression is not as high as ySTAR10-AD in inducing conditions, but higher than absent TALEs nonetheless.

Furthermore, there is no significant effect of spacer sequence on the ySTAR10 strains. Results for both the red and green fluorescence are very similar for the 37, 239 and 580 bp spacers separating the red and green reporter genes. This cannot also be said for the ySTAR11 strains, however. Despite the relatively small difference of only 12 bases changed between the PFY1 promoters in ySTAR10 and ySTAR11 strains, the effects of the spacer sequences in ySTAR11 strains is remarkable, yet it is absent in ySTAR10 strains.

The choice of spacer seems to have a severe impact on the basal expression level of the PFY1 promoter in the ySTAR11 strains. The short spacer eliminates virtually all promoter activity, resulting in expression levels comparable to untransformed cells. The 239 bp spacer boosts basal expression to levels comparable to the medium-strength YRA1s promoter and higher than the original PFY1 promoter. Finally, the 580 bp spacer takes expression to a level that falls in between the previous two cases and lower than seen for the original promoter.

The effect of TAL-effector expression in the ySTAR11 strains could not be determined, however. There are no large differences in sfGFP expression levels between the TALEs with activation domain, without activation domain and the samples where TALE expression was repressed. It is unclear in these cases if the promoters have become insensitive to TAL-effector binding, because the positive control for TALE expression failed in all of these strains. This can be concluded from the fact that there is no detectable red fluorescence from the ySTAR-AD strains that have been incubated overnight in galactose media.

From these further unexpected results, we were forced to conclude that the PFY1 promoter is especially context dependent to a degree that makes it unworkable for this system. This confirms earlier experiments where we had already observed indications that this might be the case (see **Figure 4.3.2** on page 135). Results for the ySTAR10 strains show that addition of the BB21 binding site upstream of the PFY1-BB7 promoter makes it insensitive to repression by TAL7 at the core promoter. Indeed, the results now show weak activation from binding TAL7. Results for the ySTAR11 strains show that addition of the BB7 site upstream of the PFY1-BB21 promoter in conjunction with different spacer sequences heavily influences basal expression of the promoter. Depending on the spacer sequence expression was either reduced to autofluorescence levels or increased to medium strength promoter levels.

These observed behaviours are unique, but in the context of synthetic biology are highly undesirable. If the characteristics of a promoter change every time the upstream binding site is modified, it undermines the modularity of the system and makes engineering of the system intractable. These considerations led us to refocus our efforts on more tractable promoters. We therefore decided to investigate TATA-box containing promoters, such as the CYC1 promoter which had already been shown to be activatable using our TALEs. Thus, rather than make repressible promoters activatable as we planned for PFY1, the strategy now requires adding repression to activatable TATA box containing promoters like CYC1.

### 5.3.3 TAL-effector optimisations

The results for the modified PFY1 promoters were difficult to interpret, largely because the PFY1 promoter showed severe context dependency. However, another significant problem seen in the previous experiments was that the induction of the CYC1 minimal promoter through upstream TAL-AD binding was not consistently successful.

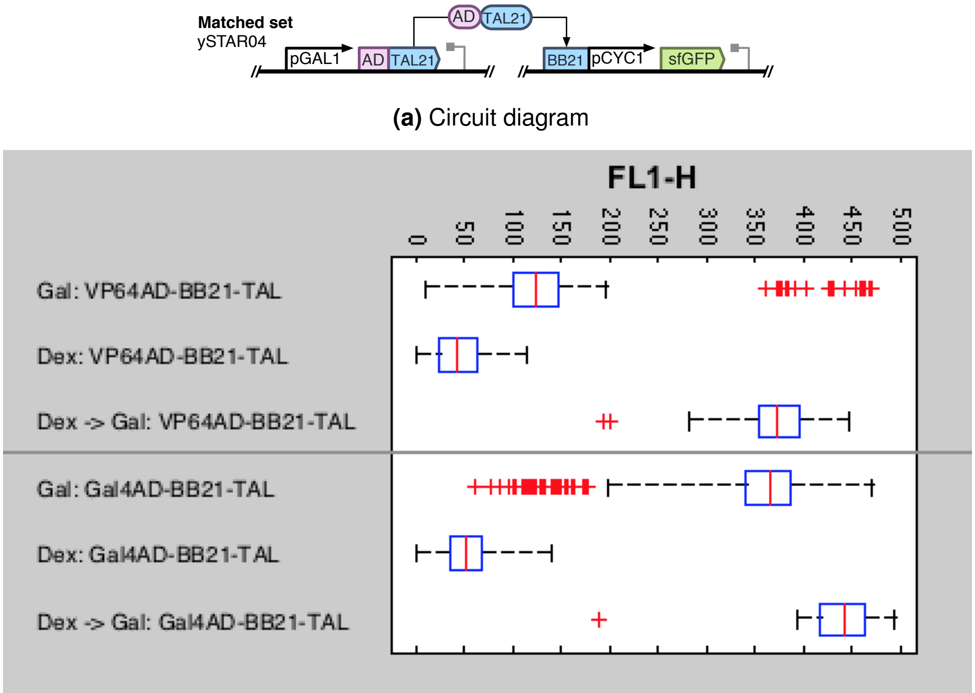
There are many causes that could underlie the issue, and in this context, the most likely candidates are genetic context effects and/or problems with evolutionary stability. Below we describe the experiments that led to this view. Furthermore we take steps to alleviate these problems, such as the reduction of DNA sequence repetition and the introduction of appropriate spacer regions. In addition we investigate if the TAL-effectors can be modified to give a stronger activation signal which will allow for better detection of signal.

#### Effect of the activation domain on evolutionary stability of TAL-effectors

The goal of the following experiment was to gain insight into the negative selection pressure imparted by the activation domain when it is fused to a TAL-effector. If there is a strong negative selection against a genetic construct a growing population will quickly be overtaken by mutants that have inactivated the detrimental effect of the particular element. This genetic element is thus said to be **evolutionary unstable**.

In order to measure the evolutionary stability, the mutants need to spread through the growing population. In a typical scenario, mutants have 12 to 24 hours to spread through the populations when a culture is inoculated overnight for flow cytometry measurements the next day. In YEP-galactose media, doubling time under standard growth conditions is roughly 2.5 hours. This equates to 5 to 10 generations for the mutants to dominate under typical conditions. To gain a more definitive result, we extended the experiment to 100 generations. If by this time the population was not taken over by a majority of mutants, we accepted that the construct was sufficiently evolutionary stable for our purposes. In practice, this meant that the 5 ml cultures were diluted 1,000,000-fold into fresh media every 2 days to ensure a period of exponential growth for 10 consecutive days.

To avoid unnecessary complexity, we chose to compare ySTAR04-VP64 and ySTAR04-GAL4 from a previous experiment. These strains have TAL21 fused to an AD controlled by the GAL1 promoter, targeting the BB21 binding site upstream of the CYC1 promoter driving sfGFP expression. For more information see **Figure 5.10** on page 172.



**(a)** Circuit diagram

**Figure 5.13:** Activation domain TAL-effector fusion stability assay in yeast. A fusion with the VP64 activation domain (top three rows) is compared against a fusion with the GAL4 activation domain (bottom 3 rows). Figure shows box-plots of green fluorescence (FL1-h, arbitrary units) after 10 days of growth with dilutions. Plots represent 10,000 events with outliers marked as red crosses. Cultures were diluted 1,000,000-fold every two days, for an estimated total of 100 generations. Cells were grown in YEP medium with indicated carbon source: galactose (Gal) for inducing, glucose (Dex) for repressing conditions. Dex-▷Gal strains were grown in glucose media for 9 days and in galactose media for 1 day directly prior to data collection.

Because we hypothesise that it is the expression of the TALE fusion, not the presence of the genetic construct that is causing the growth defect and the selection of mutants, we also included a second growth condition. In this condition, the strains were grown in YEP-glucose for 9 days and only transferred to galactose media 12 hours before flow cytometry measurements. We expect these populations to develop very few mutants, since there is only a selective pressure due to protein expression during the 12 hours prior to data collection.

Finally, we also included the strains after growth in glucose for the entire duration of the 10 days, in order to compare fluorescence levels of the other conditions to a condition with no detectable expression of fluorescent proteins. For every condition and strain 10,000 events were collected by flow cytometry. The results were analysed in Matlab. Box plots were created to better analyse the distribution of the mutants and unmutated constructs within the populations. The results are shown in **Figure 5.13**.

The figure details the expression levels seen from the constructs over 100 generations and confirms our hypothesis that there is a selection against TALE-AD fusions. This follows from the fact that strains that have only been induced for 12 hours prior to data collection (Dex-▷Gal) show markedly higher sfGFP expression levels than those that have been grown in galactose for 10 days (Gal). In fact, expression in ySTAR04-VP64 after induction over 100 generations is only marginally higher than basal expression of the minimal CYC1 promoter. A small minority of cells, marked by the red crosses, have retained expression at the level seen with the original construct. This means that the original population has been outcompeted by mutants that have completely or largely abolished the transcription activating ability of the TAL-effector.

Expression in ySTAR04-GAL4 after induction over 100 generations is generally lower than that from the original system. However, only a small minority of cells (indicated by the red crosses) are at the level of uninduced minimal CYC1 promoter expression. In fact, 25% of the population have expression levels that fall within the range of the original construct. This indicates that the selective pressure against the GAL4 AD is significantly less than the pressure against VP64 when fused to the TAL-effector.

It is important to note, however, that evolution is inherently a stochastic process. The 1,000,000 fold dilutions employed in this experiment further exacerbate the stochastic nature of the mutations. This could have consequences for the conclusions drawn from this experiment. In the discussion (**subsection 5.4.3** on page 197), the implications of this notion in the context of this experiment are examined further.

Evolutionary stability of elements used in synthetic circuits is of crucial importance for reliable function of the circuit. Instability not only hampers characterisation of the construct and any intermediates, but is especially important for the intended applications. Any circuit that has reduced or abolished functionality in 10 to 100 generations is not suitable for all but a few applications. For this reason, and despite other favourable characteristics of the VP64 activation domain (better repression when the TALE is targeted to the core promoter region and higher induction strength), we now chose to use the GAL4 activation domain for the experiments in **Figure 5.12** on page 177 and the remainder of this project.

#### Linker optimisation

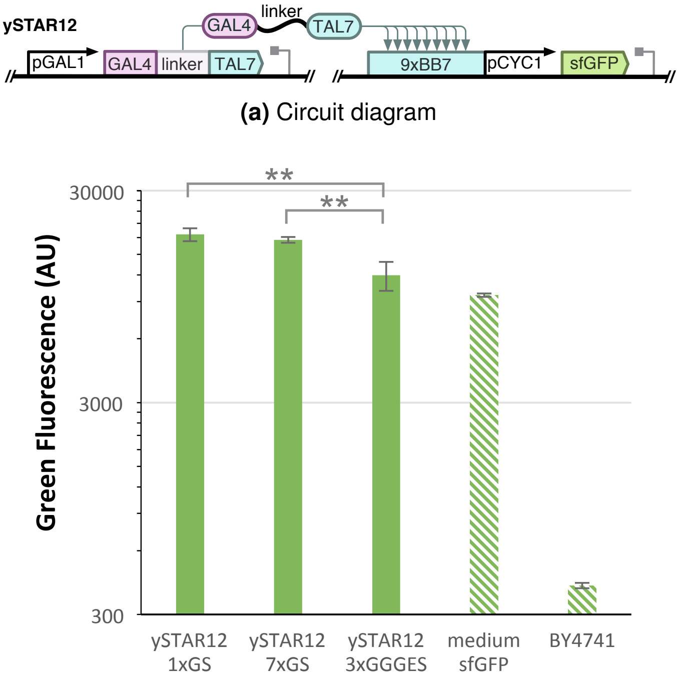
Despite its improved evolutionary stability, the results for the GAL4 activation domain still showed some selective pressure against the domain being maintained over time. In addition, we had shown previously that GAL4 AD does not induce the same activation levels as VP64 (see **Figure 5.10** on page 172). With this in mind, we next investigated if the GAL4-AD-based TALE construct could be optimised to boost its performance. VP64 activation levels could potentially be matched or exceeded, but more importantly, optimisation could also allow the TAL-effector fusions to be expressed at a lower level. This would likely lead to a reduced selection pressure against the construct.

During the described constructions of TAL-effectors with activation domains, no effort was made to optimise the linker region between the two domains. It is known that linker sequence can have a significant impact on the activity of the domains in protein fusions[277] and so this led us to set up an experiment that would test if performance could be optimised by changing the linker sequence connecting the GAL4 activation domain to the TAL-effector.

A set of strains called ySTAR12 were created to test the efficacy of different linkers. These contained a reporter unit integrated into the LEU2 locus of BY4741, consisting of an array of 9 BB7 binding sites upstream of the minimal CYC1 promoter driving sfGFP expression. An array of 9 TAL-effector binding sites was shown previously to lead to higher expression than a single site and was selected for this reason (we return to this topic in the next optimisation experiment).

The effector unit that acts upon this reporter unit, consists of the GAL1 promoter controlling expression of the GAL4 activation domain fused to TAL7 via a selection of different linkers. This unit is integrated into the URA3 locus on the genome. The linkers were selected to vary in length and flexibility. The original linker consisted of just a glycine-serine sequence (ySTAR12-1xGS). Linkers consisting predominantly of glycine and serine residues are considered to be flexible[277]. A longer flexible linker consisting of a repeat of seven glycine-serine residues (ySTAR12-7xGS) was also included. The increased length may reduce steric hindrance from the fused TALE protein or allow the activation domain to be more accessible to the transcription initiation machinery. Finally, a semi-flexible triple repeat of the glycine-glycine-glycine-glutamic acid-serine sequene (ySTAR12-3xGGGES) was also tested. This linker contains a charged residue, which reduces the propensity for the linker to fold back on itself, effectively reducing its flexibility.

The three strains of this set were tested for sfGFP expression by flow cytometry after overnight induction with galactose in minimal media (for method see Materials and Methods **subsection 2.2.3** on page 67). A strain with the medium strength YRA1s promoter driving expression of sfGFP and the parental strain BY4741 were included as controls. Measurements were performed on the BD Fortessa X-20 flow cytometer, analysed using FlowJo and are presented in **Figure 5.14**.



**(a)** Circuit diagram

**Figure 5.14:** Assessing the role of the peptide linker sequence between the GAL4 activation domain and the TAL-effector protein. The original glycine-serine (1xGS) linker was compared against a repeat of 7 glycine-serines (7xGS) and a repeat of 3 glycine-glycine-glycine-glutamic acid-serine linkers (3xGGGES). Bars represent the average median green fluorescence of cells sampled from 8 individual transformants in arbitrary units (AU). Error bars represent the standard deviation of the median of the 8 isolates. Controls (hatched); 'medium sfGFP': sfGFP expressed from the YRA1s medium strength promoter, 'BY4741': wild type/parental strain BY4741. Double asterisks indicate significant differences ($p<0.001$) as determined by a two-sided t-test. Cloning and measurements by RC.

The results show that the effect of changing the linker peptide sequences is very limited. The 7xGS linker shows an insignificant decrease in expression, while the 3xGGGES linker shows a significant decrease, compared to the 1xGS default. This showed there was no merit in changing the linker with the aim of achieving higher expression strengths.

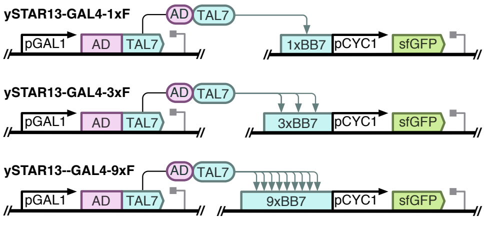
### TALE binding site array optimisation

Another way to optimise activation from the TALE is to consider its binding sites on its promoter pair. It is known that adding multiple binding sites for an activating transcription factor increases maximum expression promoter, and so we investigated here new promoter designs that can boost the activation strength further. We designed a series of constructs with 1, 3 and 9 repeats of the TALE binding site upstream of the core promoter. Although adding more than 9 binding sites would potentially result in increased induction, we expected returns to diminish with increasing binding site number. More importantly, genetic stability could also become an issue as repeat numbers increase, as this means that the promoters end up with repetition of DNA sequences and multiple repeats of the same sequence are prone to recombination *in vivo*. This means the constructs are unstable, which can lead to problems in downstream applications, such as characterisation by flow cytometry. The issue is less severe in cloning strains of *E. coli*, because these have typically had major recombinase enzymes inactivated. However *S. cerevisiae* is a favoured model organism specifically for its efficient homologous recombination abilities and so in yeast this is a major concern.

In nature this problem is resolved by the promiscuity of transcription factors. Subtle sequence variations in the binding site do not prevent the TF from binding its targets, but reduce direct homology between target sites. The single base specificity of TAL-effectors prevented us from using the same approach here. Instead, we chose to introduce non-repeating spacer sequences between every instance of the binding site. This does not completely solve the problem, but the relatively short length of the repeating sequence (20 bp) and the reduced proximity between them will lower the risk of recombination[278]. The increased spacing can also potentially reduce steric hindrance between the TAL-effector proteins bound to this stretch of DNA.

In accordance with standard practice in protein engineering, the activation domain was moved to the N-terminus of the TAL-effector early on in the project[276]. This moved the activation domain to the side of the TAL-effector distal to the core promoter region. We hypothesised that the activation strength could be increased by reversing the orientation of the binding site and thereby allowing the TALE to bind with the activation domain proximal to the core promoter. With this in mind, we designed a set of characterisation constructs with TALE binding sites in reverse orientation compared to the regular binding sites.

Strains were created with a variety of binding site arrays upstream of the minimal CYC1 promoter. All contained TAL7 fused to the GAL4 activation domain. Expression of the TALE was driven by the inducible GAL1 promoter and this assembly was integrated into the genome at the URA3 locus. The reporter construct consisted of sfGFP driven by the minimal CYC1 pro-



moter with the inducing TALE binding site array upstream, integrated into the LEU2 locus of BY4741. Three strains were created with the default forward orientation: ySTAR13-GAL4-1xF, ySTAR13-GAL4-3xF, ySTAR13-GAL4-9xF, containing repeats of one, three and nine BB7 binding sites respectively.

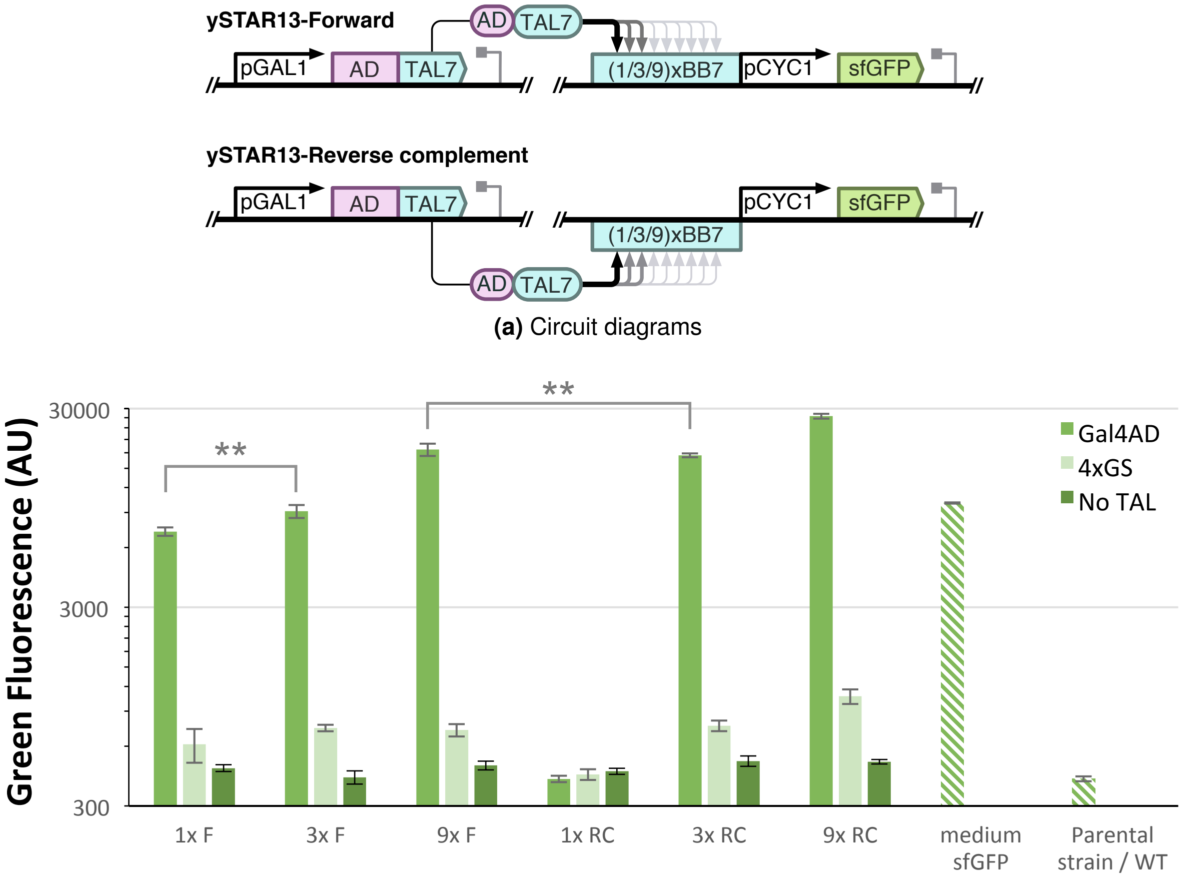
Another three strains were built to mirror the first three, including the same binding site arrays except reverse-complemented: ySTAR13-GAL4-1xRC, ySTAR13-GAL4-3xRC, ySTAR13-GAL4-9xRC. Finally, for each of the six strains, a control strain was created with the GAL4 activation domain replaced by a non-activating 4xGS linker: ySTAR13-4xGS-1xF, ySTAR13-4xGS-3xF, etc.

These strains were tested for green fluorescence by flow cytometry after overnight induction with galactose in minimal media (for method see Materials and Methods **subsection 2.2.3** on page 67). Measurements were performed on the BD Fortessa X-20 flow cytometer. The 6 ySTAR-GAL4 strains were also measured after incubation in media suppressing TALE expression (minimal dextrose media). A strain with the medium strength YRA1s promoter driving expression of sfGFP and the parental strain BY4741 were included as controls. The measured data were analysed using FlowJo and are presented in **Figure 5.15** on the following page.

The experiment confirms earlier results that increasing the number of binding sites upstream of the promoter increases the activation of the promoter. Induction in ySTAR13-GAL4-3xF is moderately, but significantly, higher (1.24 fold) than the expression level in the strain with a single binding site. Induction in ySTAR13-GAL4-9xF is 2.2 times as high as the strain with a single binding site. Low expression or no expression was observed for controls where the TAL-effector was expressed without the activation domain or in conditions where the TALE was not expressed, respectively. This is consistent with expectations.

Interestingly, we also find that the results for induction from sites in the reversed direction are consistent with our hypothesis that reversal of the binding site can lead to increased activation by orientating the activation domain towards the core promoter region. Specifically, induction in ySTAR13-GAL4-3xRC is more than twice as strong as for its forward equivalent, ySTAR13-GAL4-3xF. Induction in ySTAR13-GAL4-9xRC is stronger still, at 3.9 times the expression level of a single binding site in forward orientation. There is one exception, however. ySTAR13-GAL4-1xRC, the strain with a single binding site in reverse orientation, shows no activation at all, which is not a result that was expected.

If induction strength had been the only consideration for selecting the binding site array, then the 9xRC configuration would have been selected to take forward from this point. However, genetic stability is another variable that needed to be taken into account. Thus, because the 3xRC configuration retains over 0.6 fold of the activity of the 9xRC configuration, we selected this reversed array of 3 binding sites for further experiments. We judged this to be an optimal compromise between activation strength and protection from homologous recombination and other forms of genetic instability.



**(a)** Circuit diagrams

**Figure 5.15:** Optimisation of the promoter binding site array for strong activation. The figure shows results for arrays of 1, 3 and 9 TAL-effector binding sites located upstream of the minimal CYC1 promoter. Binding to both the forward (F) and the reverse complement (RC) orientation were tested. In each construct the GAL4 activation domain was replaced with the 4xGS linker sequence for a non-activating reference. In addition, the uninduced expression levels for the CYC1 promoter were obtained in each of these constructs by incubation in glucose media (indicated with 'No TAL'). Bars represent the average median green fluorescence of the cells from 8 individual transformed colonies, given in arbitrary units (AU). Error bars represent the standard deviation of the median of eight clones. Controls (hatched); 'medium sfGFP': sfGFP expressed from the YRA1s medium strength promoter, 'parental strain/WT': wild type strain BY4741. Double asterisks indicate significant differences ($p<0.001$) as determined by a two-sided t-test. Cloning and measurements by RC.

### 5.3.4 Characterising induction of TATA-box promoters

The results for the PFY1 promoter showed that individual promoters can have very erratic properties. In order to increase the chances of finding a promoter with favourable properties, we decided to test a variety. The minimal CYC1 promoter, which has previously shown favourable induction properties, is a TATA-box containing promoter, while the PFY1 promoter is not. We selected a variety of different minimal TATA-box containing promoters for an initial characterisation experiment.

Not all tested promoters will turn out to be suitable and it is important to find the most promising ones with as little effort as possible. A series of experiments of increasing complexity was designed to select the promoters most suitable to our needs. The first and most straightforward experiment was the determination of the basal expression of the promoters. A low basal expression is favoured to allow strong induction by the TAL-effector.

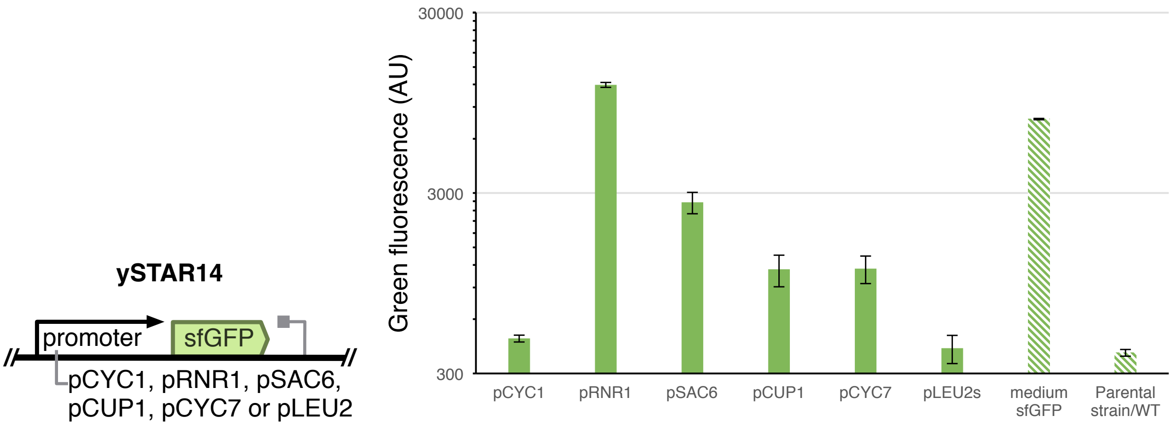
### Testing basal expression of a selection of TATA-box core promoters

To find the basal expression level, the core promoters were cloned to drive expression of sfGFP on a plasmid that could be integrated into the URA3 locus of the yeast genome. The minimal CYC1, RNR1, SAC6, CUP1, CYC7 and LEU2 promoters were selected for this experiment, and the constructs were named ySTAR14-pCYC1, ySTAR14-pRNR1, ySTAR14-CUP1, etc.

These six strains were tested for basal sfGFP expression by flow cytometry after overnight growth in minimal media containing glucose (for method see Materials and Methods **subsection 2.2.3** on page 67). Measurements were performed on the BD Fortessa X-20 flow cytometer. A strain with the medium strength YRA1s promoter driving expression of sfGFP and the parental strain BY4741 were included as controls. The measured data were analysed using FlowJo and are presented in **Figure 5.16**.

The results show a wide range in the basal expression strength of the tested promoters. The CYC1 and LEU2 promoters are the most favourable, with low basal expression which is only marginally higher than autofluorescence. pCUP1 and pCYC7 show low expression, while pSAC6 shows medium-low levels. pRNR1 shows the highest levels of all, comparable to the medium strength YRA1s promoter tested as a control.

Following these results, the RNR1 promoter was eliminated from further analysis as a result of its high basal expression levels. pSAC6, CUP1 and CYC7 were carried forward, since there was enough potential for further activation. LEU2 was so close to autofluorescence levels there was no guarantee it would still function as a promoter. However, it was taken forward just in case. In summary, the CYC1, SAC6, CUP1, CYC7 and LEU2 promoters were selected for further analysis.



**(a)** Circuit diagram

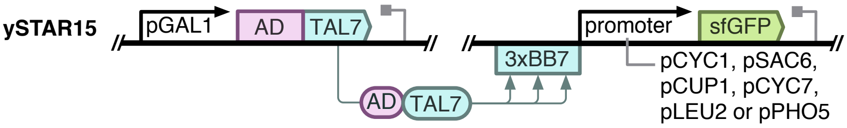
**Figure 5.16:** Characterisation of basal expression levels for a variety of minimal TATA-box promoters. Bars represent the average median green fluorescence of cells from 12 individual transformants in arbitrary units (AU). Error bars represent the standard deviation of the median. Controls (hatched): 'medium sfGFP': sfGFP expressed from the YRA1s medium strength promoter, 'parental strain/WT': wild type strain BY4741. Cloning and measurements by RC.

### Induction characteristics of TATA-box promoters

With a selection of TATA-box core promoters and a new optimised configuration of upstream binding sites, we were next able to measure the induction characteristics of the promoters that had passed the first assay. Originally we intended to also test the PHO5 promoter in the previous screen, but it could not be cloned in time for the first assay. Given the high success rate of the first assay, we decided to also include this potential promoter in the second assay without prior testing.

The constructs for this assay closely resemble those in the previous experiment. Rather than vary the binding site configuration, this was now fixed and instead the core promoter driving sfGFP was exchanged for one of the 6 selected for the assay. As before, all strains contained TAL7 fused to the GAL4 activation domain. Expression of the TALE was driven by the inducible GAL1 promoter and this assembly was integrated into the genome at the URA3 locus.

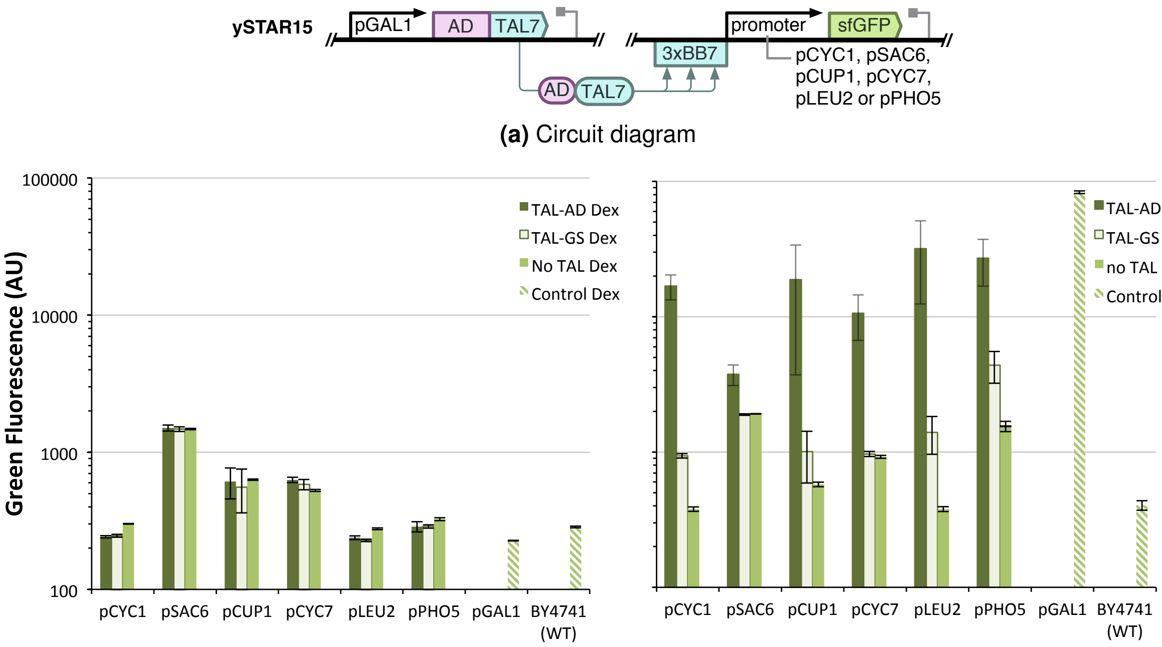
The reporter construct consisted of sfGFP driven by a TATA-box promoter with the inducing TALE binding site array of three BB7 binding sites in reverse orientation cloned upstream, integrated into the LEU2 locus of BY4741. The tested promoters were pCYC1, pSAC6, pCUP1, pCYC7, pLEU2 and pPHO5. These constructs were named ySTAR15-GAL4-pCYC1, ySTAR15-GAL4-pSAC6, etc. and a diagram is shown below.



For each of the six strains, a control strain was created with the GAL4 activation domain replaced by a non-activating 4xGS linker: ySTAR15-4xGS-pCYC1, ySTAR15-4xGS-pSAC6, etc. In addition, for each promoter a control strain was created lacking the TAL-effector construct altogether: ySTAR15-noTAL-pCYC1, ySTAR15-noTAL-pSAC6, etc. This amounts to three strains of yeast being created per screened promoter. Each of these strains was observed in both glucose and galactose media, for a total of 6 data points per promoter. These extensive controls were necessary because these promoters may be affected by the binding of a TAL-effector irrespective of an activation domain and they may be affected by galactose irrespective of TALE presence. To be able to differentiate between each of these effects, the controls detailed above were essential.

These strains were assessed for sfGFP expression by flow cytometry after overnight growth in minimal media with the relevant carbon source (for method see Materials and Methods **subsection 2.2.3** on page 67). Measurements were performed on the Attune NxT flow cytometer. A strain with the strong GAL1 promoter driving expression of yeGFP and the parental strain BY4741 were included as positive and negative controls. The measured data were analysed using FlowJo and are presented in **Figure 5.17** on the following page.

The outcome of this experiment was almost entirely as expected. All core promoters showed activation with GAL4-TAL7 expression. Each of the three strains created per promoter were tested both in glucose and galactose. For each promoter, the three data points corresponding to the glucose measurements were close to identical. This is expected, because the differences in these 3 strains are in the TALE construct, which is not expressed in glucose media. In



**(a)** Circuit diagram

**(b)** Repressed TAL expression (Dextrose media)

**(c)** Induced TAL expression (Galactose media)

**Figure 5.17:** Characterisation of minimal TATA-box core promoter activation through upstream TAL-effector binding. Gal4 activation domain fused to TAL7 was targeted to 3 repeats of the corresponding BB7 binding site oriented in the reverse direction and located upstream of 6 TATA-box promoters. Expression levels were recorded by monitoring sfGFP levels after activation of the TAL-effector (indicated with 'TAL-AD Gal'). For comparison, fluorescent output was also monitored in conditions when TAL-effector expression is repressed: 'Dex', in complete absence of TAL-effectors: 'No TAL' or with TAL-effectors lacking an activation domain: 'TAL-GS'. Bars represent the average median green fluorescence of cells from 8 individual transformants in arbitrary units (AU). Error bars represent the standard deviation of the median. Controls (hatched); 'pGAL1': yeGFP expressed from the strong GAL1 promoter, 'BY4741/WT': Parental/wild type strain BY4741. Cloning partially by RC.

addition, the obtained basal expression levels corresponded very well to the results obtained in **Figure 5.16** on page 186. Basal expression level for the core PHO5 promoter was comparable to the most promising candidates i.e. the CYC1 and LEU2 promoters.

We expected the measurements for the ySTAR15-noTAL constructs in galactose to mirror the measurements in glucose. For most promoters this was indeed the case, however pCYC7 and pPHO5 showed elevated expression compared to their basal expression levels. In pCYC7, expression was 1.75 fold of basal levels, while in pPHO5 expression was 4.8 times higher. This was unexpected and means that the promoters, PHO5 in particular, are induced by galactose irrespective of TAL-effector presence.

Naively, we expected the measurements for the ySTAR15-4xGS strains in galactose to match the ySTAR15-noTAL measurements in galactose. However, it was already observed in **Figure 5.15** on page 185 that binding of a TALE with no activation domain can lead to low levels of activation. We observed this in the majority of the screened promoters: CYC1 - 2.5 fold, CUP1 - 1.7 fold, LEU2 - 3.7 fold and PHO5 - 2.8 fold activation, all due to TALE binding with no activation domain. In the SAC6 and CYC7 promoters this effect was not observed.

Finally, all promoters were activated by the binding of the TAL-AD fusion. Compared to expression in glucose, pCYC1 was induced 70-fold. Activation strengths for the remaining promoters were as follows: pSAC6 - 2.5 fold, pCUP1 - 31 fold, pCYC7 - 17 fold, pLEU2 - 134 fold and pPHO5 - 94 fold. Based on the gathered data we could select which promoters to carry forward for further engineering. The SAC6 promoter was the weakest candidate, due to its high basal expression and poor activation. PHO5 had shown some unexpected characteristics in its ability to be partly induced by galactose. This could potentially be used in future applications, and so it was selected for the next assay. Of the remaining 4, all promoters qualified for further experiments. However, our experimental capacity was not sufficient to accommodate all 5 and thus we dropped two. pCUP1 was chosen over pCYC7 for its slightly lower basal expression in galactose and higher induction level and pCYC1 was chosen over pLEU2 because it is a more widely used and well-characterised promoter.

### 5.3.5 TATA-box based STAR promoters

The final step in the creation of promoters with STAR regulation is the introduction of a TAL-effector binding site into the core promoter region to allow for binding to exert repression. With the BB7 binding site being used for activation, the BB21 site was inserted between the TATA-box and the Transcription Start Site (TSS). Because this change in the DNA sequence of the core promoter could affect the function of the promoter, we first verified whether the promoters could still be induced by the upstream binding of the GAL4-TAL7 activator. The three promoters selected for this assay were CYC1, CUP1 and PHO5 and further details on the assembly of the constructs is given below.

To determine the insertion/replacement site for placing the BB21 binding sequence into each of the promoters, we first identified the TATA-box and TSS. The former was determined by a search for its consensus sequence in *S. cerevisiae*. This sequence is TATAWAWR, where W = A or T and R = A or G[28]. The TSS was found by searching for the consensus $A(A_{rich})_5NYAWNN(A_{rich})_6$, where Y is C or T and W = A or T[35].

Frequently this would yield multiple hits within the same promoter region. To find the correct sites, these hits were cross-referenced with other sources of information. For example, genome wide information from large scale experiments is available on TATA-box elements[279] and TSSs[35,247,254,280] through the Saccharomyces Genome Database (SGD)[255]. This may not always yield results for promoters that are naturally inducible, such as the CUP1 promoter which is induced by copper ions. Where genome-wide data was not sufficient we cross-referenced this with data from studies about the individual promoters.

With the locations of the core regulatory elements identified, we could select a location for the insertion of the TALE binding site. Typically there was a 30-40 bp distance between the end of the TATA-box and the beginning of the TSS, although for CUP1 the distance was 79 bp. This sequence was scanned for the location where the placement of the BB7 or BB21 binding site would have the lowest number of base changes. This strategy resulted in only needing to make 10 (CYC1), 11 (CUP1) or 12 (PHO5) base pairs changes to incorporate a binding site length of total length of 21 bp.

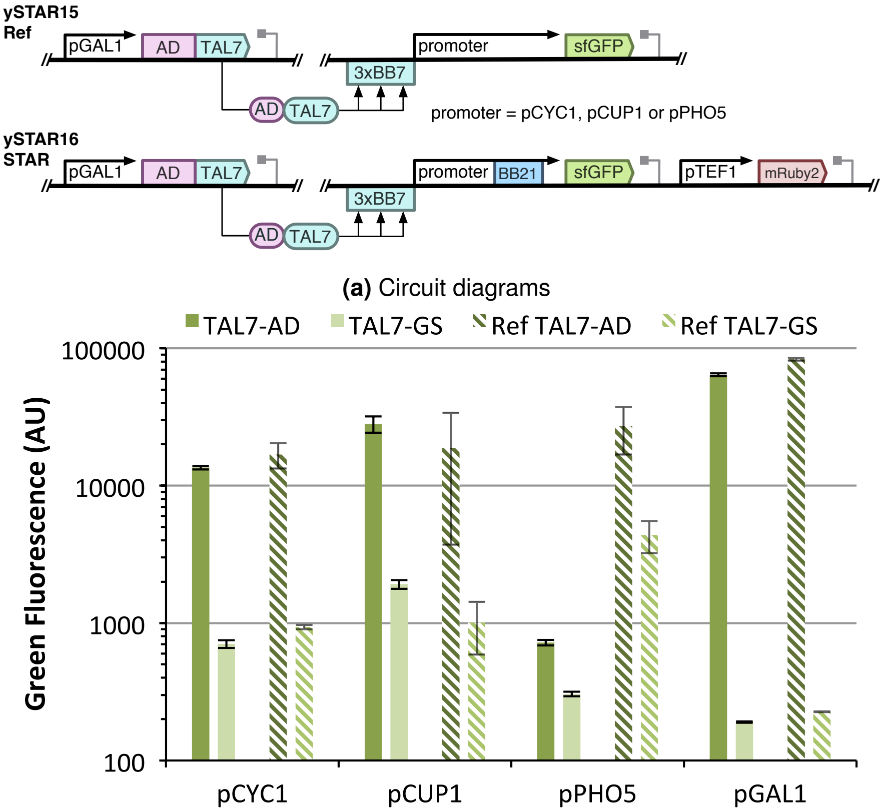
The constructs for this assay closely resemble those in the previous experiment. All strains contained TAL7 fused to the GAL4 activation domain. Expression of the TALE was driven by the inducible GAL1 promoter and this assembly was integrated into the genome at the URA3 locus. The reporter construct consisted of sfGFP driven by a TATA-box promoter with the inducing TALE binding site array of three upstream BB7 binding in reverse orientation, integrated into the HO locus of BY4741 using the HIS3 marker. Unlike the ySTAR15 constructs, the reporter unit was also joined to an mRuby2 transcription unit in a multi-gene assembly step and this red fluorescent protein was expressed from the constitutive TEF1 promoter as an expression control. The tested promoters were pCYC1, pCUP1 and pPHO5, each with the BB21 binding site integrated into the core promoter region and 3 upstream BB7 binding sites. These constructs were named ySTAR16-GAL4-pCYC1, ySTAR16-GAL4-pCUP1 and ySTAR16-GAL4-pPHO5.

For each of the three strains, a control strain was created with the GAL4 activation domain replaced by a non-activating 4xGS linker: ySTAR16-4xGS-pCYC1, ySTAR16-4xGS-pCUP1 and ySTAR16-4xGS-pPHO5. The strains were tested by flow cytometry after overnight growth in minimal media with galactose (for method see Materials and Methods **subsection 2.2.3** on page 67). Measurements were performed on the Attune NxT flow cytometer. A strain with the strong pGAL1 promoter driving expression of yeGFP and the parental strain BY4741 were included as controls after overnight incubation in both galactose and glucose media. The measured data were analysed using FlowJo and are presented in **Figure 5.18** on the following page.

The purpose of this experiment was to verify that the BB21 binding site could be inserted into the core promoter region of the three selected promoters and these would retain their ability to be induced. To compare characteristics of the modified promoters to the originals, data from the relevant promoters from **Figure 5.17** on page 188 was reproduced here. Although these data were not collected on the same run as the data for ySTAR16, they should provide a suitable reference.

When analysing the induction characteristics for the modified promoters, it is immediately obvious that the results for the PHO5 promoter do not resemble the results for the unmodified promoter. Introducing the BB21 binding site in this promoter resulted in severely reduced expression levels, both in conditions with and without the GAL4 activation domain fused to the TAL-effector. This indicated that critical components of the promoter had been compromised by the sequence changes. The PHO5 promoter could therefore not be put forward for further characterisation.

The CYC1 and CUP1 promoters did show expression levels comparable to their unmodified counterparts. This indicated the essential promoter elements of the promoter were still functioning as before. These promoters could thus be tested for their ability to be repressed by a TAL-effector targeted to the BB21 binding site in the core promoter region. The following experiment was designed to provide a full characterisation of simultaneous transcription activation and repression in these promoters.



**(a)** Circuit diagrams

**Figure 5.18:** Verification of promoter activity of TATA-box promoters designed for STAR regulation. The CYC1, CUP1 and PHO5 promoters were modified with the addition of the BB21 binding site in the core promoter. This figure shows how the newly created versions (solid bars) compare to the originals (hatched). Expression levels for the original promoters are hatched diagonally and indicated with 'Ref' in the legend. The expressed TAL-effector was fused to the GAL4 activation domain (TAL7-AD) or the control 4xGlycine-Serine linker (TAL7-GS). Bars represent the average median green fluorescence of cells from 8 individual transformants in arbitrary units (AU). Error bars represent the standard deviation of the median. 'pGAL1': sfGFP expressed from the strong GAL1 promoter in galactose conditions (high bar) or glucose conditions (low bar).

### Full characterisation of two promoters for STAR regulation

The above series of increasingly stringent assays led us to a final set of two engineered promoters that show good potential for simultaneous activation and repression. In this final experiment we perform a full characterisation of the modified CYC1 and CUP1 core promoters with the aim of assessing their performance parameters for both repression and activation.

At this point, the main property that had not been characterised was the ability for these promoters to be repressed by a TAL-effector bound to the core promoter region. The strains constructed for this assay closely resemble those in the previous experiment. In essence, the ySTAR16-pCYC1 and ySTAR16-pCUP1 strains were simply transformed to also incorporate expression of GAL4-TAL21 from the genome. Below we describe this in more detail.

Two sets of strains were created for this experiment. The first set, ySTAR17, based around the core CYC1 promoter and the second, ySTAR18, based around the core CUP1 promoter. These strains contained two TAL-effectors: TAL7 and TAL21. The TALEs were fused to the GAL4 activation domain and in both cases controlled by the inducible GAL1 promoter. The

TAL7 transcription unit was integrated into the genome at the URA3 locus, while the TAL21 transcription unit was integrated at the LEU2 locus.

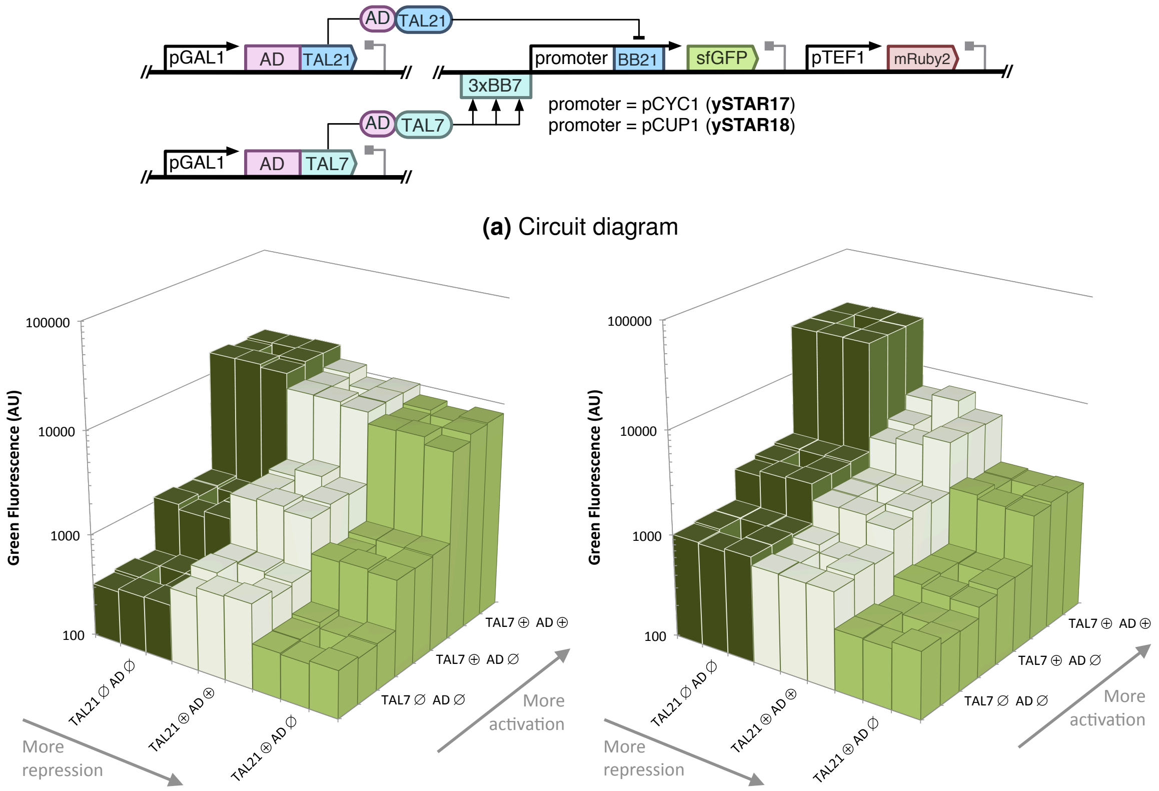
After switching to the GAL4 activation domain for TALE fusions we observed no more strong indications of evolutionary instability of the TAL-effectors. However, we recognised that the introduction of a second TALE into the same strain might exacerbate the issue. To reduce the probability of recombination within the TALE or with the other TALE construct, we chose to codon-optimise TAL21. We used the Evolutionary Failure Mode (EFM) calculator as a measure for the DNA sequence repetitiveness of the construct[16]. This calculator computes the occurrence of homologous regions and simple sequence repeats, both of which are known to have a higher probability of causing mutations. The output of the computation is the likelihood that a certain construct will acquire mutations relative to the background mutation rate. A score of 100 means that a certain sequence is 100 times more likely to acquire mutations compared to a typical section of DNA of the same length on the genome.

Before optimisation of TAL21, the EFM score was 93784 (very high), indicating that this sequence is exceptionally susceptible to mutations. After computational optimisation virtually all homology within the TAL-effector was eliminated and the EFM score was now 2.7 (very low). This optimisation was performed by Robert Chen (RC). The lack of repetitive sequences does not completely protect the construct against any type of mutation, but the change was considered to be a significant improvement. For this reason we chose to incorporate the improved version of TAL21 into the ySTAR17 and ySTAR18 strains.

The reporter construct consisted of sfGFP driven by the TATA-box promoters with the inducing TALE binding site array of three BB7 binding sites in reverse orientation placed upstream. The reporter unit was joined to a control mRuby2 transcription unit in a multi-gene assembly step and subsequently integrated into the HO locus of BY4741 using the HIS3 marker. The red fluorescent protein was expressed from the constitutive TEF1 promoter. The tested core promoters were pCYC1 and pCUP1, each with the BB21 binding site integrated into the core region.

As in some previous experiments we found it crucial to determine the effects imparted by the binding of the TAL-effector and the activation domain separately. To do this, three versions of each TAL-effector were tested. Firstly, the regular version of the TALE fused to the GAL4 activation domain (TAL$\oplus$ AD$\oplus$). Secondly, a version where the activation domain was replaced by the control 4xGS linker (TAL$\oplus$ AD$\oslash$). Finally, a strain was included where the construct expressing the TAL-effector was completely absent (TAL$\oslash$ AD$\oslash$). Testing all possible combinations of these variations for the two individual TALEs resulted in a total of (3 x 3 =) 9 strains. With two promoters to characterise, this amounted to a total of 18 measured strains for this experiment. 8 colonies for each strain were measured after overnight growth in minimal media with galactose (for method see Materials and Methods **subsection 2.2.3** on page 67). Measurements were performed on the Attune NxT flow cytometer. The measured data were analysed using FlowJo and are presented in **Figure 5.19** on the next page.

The results are presented in a three-dimensional bar chart, where each of the two horizontal axes represents one of the two TAL-effectors. The samples are ordered based on the expected size of the effect of the TALEs. Based on previous results, we expect TAL21$\oplus$ AD$\oslash$ to have the strongest repressive effect. TAL21$\oplus$ AD$\oplus$ is expected to show less repression, since the



**(a)** Circuit diagram

**(b)** ySTAR17: pCYC1

**(c)** ySTAR18: pCUP1

**Figure 5.19:** Characterisation of promoters engineered to be dual-regulated. The two tested promoters - CYC1 and CUP1 - have three upstream BB7 binding sites for activation and a BB21 binding site in the core promoter region for repression. The TAL-effectors were expressed fused to the GAL4 activation domain (TAL⊕ AD⊕), with the control 4xGS linker (TAL⊕ AD⊘) or not expressed at all (TAL⊘ AD⊘). The measurements for the resulting 9 combinations of TAL-effectors are shown in this figure. Each bar represents the median measurement of a single transformant (10,000 events). For every one of the 9 TAL-combinations, median measurements for eight individual transformant populations are shown to give an indication of variance.

activation domain appears to interfere with repression. Finally, TAL21⊘ AD⊘ is the no repression control. The results are arranged in this order, starting with the least repression (TAL21⊘ AD⊘) on the left.

The TAL7 samples are arranged from no activation to high activation. No activation is expected for TAL7⊘ AD⊘ and from previous experiments we expect slight activation from TAL7⊕ AD⊘. Finally, high induction is expected for the complete TALE with the GAL4 activation domain: TAL7⊕ AD⊕. These strains are ordered from low to high activation starting with low activation on the left of the second axis in **Figure 5.19**.

It is immediately clear from the results that the response to the presence of the various versions of the TALEs is different in some critical ways between the two promoters. Before discussing this, we will first discuss the specific behaviour of the two promoters individually.

For ySTAR17, activation characteristics are in perfect agreement with those obtained previously for pCYC1. Compare for example the three measurements for ySTAR17-TAL21⊘ AD⊘ (set of dark bars) with the results obtained for pCYC1 in galactose in **Figure 5.17** on page 188. However, expression of the strongest repressor, TAL21⊕ AD⊘, in these 3 strains leads to no significant decrease in expression levels, leaving the expression only 2.1 to 1.2 times lower than in unrepressed conditions. This indicates that expression of TAL21 has no meaningful repressive effect on the CYC1 promoter.

In fact, the weaker repressor TAL21$\oplus$ AD$\oplus$ shows a net activating effect when expressed in a strain that is not already highly induced by TAL7$\oplus$ AD$\oplus$. Compare for example the lightest set of columns row-wise to the darker green sets of columns and note that expression is 1.8 to 2.0-fold higher for the light set than their neighbours in two of the three conditions. This effect had not been found before, was unexpected and shows that the modified CYC1 promoter made here is not suited for its intended purpose.

A final observation for ySTAR17 is that binding of a TAL with no activation domain (ySTAR17-TAL7$\oplus$ AD$\oslash$) led to 3.0-3.4 fold induction in all tested conditions, when compared to conditions without any upstream binding TALE (ySTAR17-TAL7$\oslash$ AD$\oslash$). This confirmed observations made in previous experiments. It remains unknown why this induction is so strong, but it offers potential benefits that we consider further in the discussion.

We now look at the results obtained for ySTAR18. Like the CYC1 promoter, induction characteristics for CUP1 (dark green bars) match earlier results. In addition, expression of the repressing TALEs shows reduction of expression levels in most circumstances. The strongest repressor, TAL21$\oplus$ AD$\oslash$, shows significant repression in each of the three TAL7 variants (2.75, 4.7 and 17 fold, going from ySTAR18-TAL7$\oslash$ AD$\oslash$ to TAL7$\oplus$ AD$\oplus$). Encouragingly, the TAL21$\oplus$ AD$\oslash$ repressor is capable of reducing CUP1 expression below basal expression levels in all strains that are not strongly induced by TAL7$\oplus$ AD$\oplus$. Repressed expression levels in these cases even approach yeast autofluorescence levels. In the most activated context, repression by TAL21$\oplus$ AD$\oslash$ results in expression levels comparable to basal expression level of CUP1, showing that the binding of this repressor can essentially fully negate activation.

Unfortunately, the weaker repressor TAL21$\oplus$ AD$\oplus$ was not able to repress the promoter significantly below its unactivated basal expression levels in any context. However, it was able to reduce expression considerably when bound to the promoter when activated fully by TAL7$\oplus$ AD$\oplus$, a 5.8 fold repression. This means the CUP1 promoter qualitatively exhibits many of the features required for the intended application - *i.e.* it is capable of simultaneous activation and repression (STAR). The implications, limitations and possible further improvements are discussed in the following section.

## 5.4 Discussion

In this chapter, we aimed to create modular transcription factors that are capable of both activation and repression, depending on their binding location on engineered promoters. We planned to use a set of orthogonally-repressible promoters based on the constitutive PFY1 promoter. In our characterisation experiments we found that the PFY1 promoter is highly context dependent. This trait is unfavourable for engineering purposes, which is the reason we instead selected a set of minimal TATA-box promoters to characterise. Of these, the CYC1 and CUP1 promoters showed the greatest potential. Interestingly, they showed very different behaviour when binding sites in the core promoter and in the upstream sequence were targeted by TAL-effectors that could repress and activate, respectively. The CUP1 promoter was shown to be repressible and inducible by TALEs tethered to an activation domain, depending on the bound location. This offers potential for these promoters and TAL-effectors to simplify existing circuits and allow more complex regulation through the application of transcription factors capable of both activation and repression. Below we discuss the details of the experiments that led to this result and look into elements that could be improved upon.

### 5.4.1 Initial characterisation experiments

The initial characterisation experiments verified the orthogonality of the TAL-effectors in their ability to repress of PFY1-based promoters. These experiments confirmed that the TAL-effectors selectively bound their respective promoters and did not interfere with non-cognate promoters. Another important result from this work was the observation that the addition of an activation domain to the TAL-effector did have an impact on the repression strength of the TALE. TALEs with an AD were not as capable repressors as those without an activation domain. Despite the sub-optimal positioning of the activation domain at the core promoter region rather than at an upstream activation site, it was capable of exerting activating effect compared to a TALE with no activation domain. Evidently, the AD could still form meaningful interactions with the pre-initiation complex in order to stabilise its assembly and enhance its function. Despite this, the net effect of TAL-AD binding was still repressive, fulfilling one of the basic requirements for the intended application of the synthetic transcription factors.

In order to confirm that the TAL-AD fusions were capable of giving a net positive effect on transcription when bound to sites upstream of the core promoter we performed a set of control experiments activating the CYC1 core promoter. These showed that the GAL4 and VP64 activation domains were both capable of significantly increasing the expression levels when targeted upstream of the CYC1 core promoter. VP64 showed 2-fold higher expression compared to the GAL4-AD and was initially selected for further experiments.

### 5.4.2 Context dependency in the PFY1 promoter

With both activation and repression by TAL-effector fused to activation domains shown to work, we next attempted to combine these two forms of regulation into a single promoter. Initially this was tried with PFY1-based promoters but multiple problems arose in these experiments. While some of these problems may have been experimental or cloning issues (e.g. positive controls not behaving as expected, large measured deviations between replicate transformants), one major issue stood out: the PFY1 promoter is highly context dependent and unpredictable. The experiments and controls unequivocally showed that as the DNA immediately upstream of the PFY1 promoter was changed, the expression from this was severely and unexpectedly changed.

This effect was best illustrated when different lengths spacers were placed between the upstream mRuby2 control sequence and the PFY1-based promoters, shown in **Figure 5.12** on page 177. With random DNA sequences of different lengths placed in the ySTAR10 strains, the spacer plus the BB21 binding site immediately upstream of the PFY1 promoter together appear to insulate expression from context changes. However, in the ySTAR11 strains the opposite is true and changing the spacer sequences has a dramatic impact on expression levels, despite there are only being 12 bp of difference between the corresponding ySTAR10 and ySTAR11 strains at the TALE binding sites. What causes such dramatic sequence context at this promoter is currently unknown. It may be related to the topology of the DNA near the upstream TALE binding site, where a natural bend in the double helix is know to play an important role in constitutive induction[58]. Sequence changes around this area could potentially disrupt this bend or alternatively could alter nucleosome positioning around the promoter.

Given all the difficulties and unexpected results from using the PFY1-based promoters, we decided to move away from these to instead base our engineering on TATA-box containing promoters. While much more is known about regulation engineering with this class of promoter, it meant that sadly we could no longer make use of the already existing set of orthogonal repressible PFY1 promoters for our study.

### 5.4.3 Optimisation experiments

Results obtained in our initial experiments with TALE activators indicated that although these were able to activate TATA-box containing core promoters they may suffer from sub-optimal design. We therefore concentrated on two features to improve: (a) lowering the burden and improving the genetic stability of strains expressing the TALE activators, and (b) improving the magnitude of the activation.

Three strategies were used in this section to optimise the genetic stability and lower the burden and stress on cells expressing TAL-effectors. Firstly, moving all TAL-AD constructs to the genome ensured lower expression by placing them at single-copy only per cell. Their expression was still from the very high strength GAL1 promoter, and so in future work if there is a further need to lower burden we suggest using less strong regulated promoters (such as pGAL10) for their expression.

Secondly, for the final experiment we recoded the RVD-sequence in one of the two TALEs that were used simultaneously in the cell. By using synonymous codon changes and DNA synthesis we were able to move away from a gene sequence predicted to be highly unstable due to repetitive DNA elements to one with a much lower likelihood of recombination and mutation. The effect of this change was not directly assayed in this work, but would hopefully allow us to build TALE-based constructs more robust in long-term use. Ideally, in future work all TALE-encoding genes would be synthesised like this, using codon changes to reduce repetitive sequences. However, this may be prohibitively expensive, especially when many different TALEs with different binding-specificities need to be made and tested. Probably the most cost-effective strategy is to stick to using the Golden Gate kit to build and test TALE-encoding genes, and then later when these are verified to work in the circuits as anticipated, have these resynthesised to reduce as much DNA repetition as possible. Alternatively, new kits may also come to market, that are designed to significantly reduce repetitiveness by using recoded parts in the assembly.

Thirdly, to further limit the stress to the host cell during TAL-AD expression, we assessed the evolutionary stability of designs using either the VP64 activation domain or the GAL4 activation domain using a functional assay that looked at transcription factor expression performance over 100 generations. Despite the superior performance by VP64 in terms of activation strength and reduced interference at the core promoter, the functional assay suggested that this domain was particularly stressful for the yeast, and thus we had no choice but to use the GAL4-AD for the remainder of the project.

However, it later became clear that the design of the experiment allowed chance to have a disproportionally large impact on the outcome. Since mutations arise stochastically, they cannot be relied on to appear with the same frequency in every instance. By only sampling a single culture for each condition of the experiment, it is possible that the obtained results are not consistent with the true impact of expression of the activation domains. For example, it is theoretically possible that the GAL4 domain has a more dramatic impact on the fitness of the cells than the VP64 domain. This means that in our experiment, the mutation(s) leading to reduced expression of functional GAL4 activation domain arose later than those leading to reduced functional VP64 expression, purely by chance. We currently have no way of knowing whether this was indeed the case, which undermines the conclusions we drew from this experiment.

To correct for this shortcoming, we could adapt methods that are commonly used in a related technique, called fluctuation analysis. This technique is used to determine mutation rates in non-selective conditions. There are many ways of interpreting the results mathematically[281,282], however there is a strong unifying feature in the experimental procedure, which is the use of multiple cultures for each experimental condition. This allows the incidence of mutations over different cultures to be averaged, smoothing the impact of stochasticity and better approaching the true value.

The major difference between fluctuation analysis and our experiment is that we explicitly assume a difference in growth rate between mutants and the original population, while in fluctuation analysis the assumption is that mutations have no impact on growth rate. Some attempts have been made to adapt the analyses to allow a mild negative impact of mutations on growth rate, as is sometimes the case for antibiotic resistance[283,284]. Presumably these adaptations could be applied to the case where mutations have a strong positive impact on growth rate. However, whether this is indeed the case remains unclear.

Irrespective of the precise mathematical implementation, it is clear that the inclusion of multiple cultures in parallel is essential to reducing the effect of chance on the outcome of the experiment and increasing the validity of the conclusions. Scientifically, we can not exclude the possibility that the conclusion we drew was incorrect. From an engineering perspective we can note that in subsequent experiments we saw no more issues with growth defects, so in that respect the change to the GAL4 activation domain had been beneficial. If the VP64 domain does need to be explored further for its lower interference with repression, then its impact on cell fitness can be reviewed with the method suggested above. If it is found to be problematic, TALE expression levels could be lowered to potentially reduce off-target effects of the VP64 domain which are thought to be causing the growth defect. Additionally, the use of alternative activation domains such as VP16 could be explored. VP16 is effectively one quarter of the VP64 domain and would therefore be unlikely to produce as much cell stress. For the same reason, it would likely also lead to less activation.

In terms of improving the magnitude of activation from the TAL-AD regulators, we looked at changing the linker between the TALE and activation domain, its orientation on the DNA and the number of sites for it to bind on its intended promoter. Linker optimisation did not yield any substantially improved activators, suggesting that the current design is already close to optimal. In contrast, binding the TALE to the DNA in the orientation that places the activation domain towards the core promoter did show a significant effect, improving the level of activation in all designs except in one case with only a single binding site. We are currently unsure why this one design did not see an improvement and speculate that the binding site may have acquired a mutation in the cloning process.

The most effective strategy to increase activation was the use of multiple TALE binding sites upstream of the minimal promoter. By moving to 3 or 9 upstream binding sites, we were able to dramatically increase the output of expression, presumably by providing more activation domains in the local DNA area to quicker recruit the PIC to the core promoter. Although we only tested 1, 3 and 9 site designs, the induction strength did not appear to linearly correlate with the number of binding sites. As every copy of the binding site introduces more repetitive DNA into the construct, we resisted the opportunity to go for maximum expression from 9 sites and instead moved forward with a 3 binding site design.

For greater activation in future designs more elaborate binding site configurations could be explored. DNA looping, for example, is used naturally by both prokaryotes and eukaryotes for enhancing expression from promoters with long-range interactions and a similar strategy could theoretically be taken with TAL-ADs. As different activation domains are considered in future work, it may also become apparent that some are better suited for long-range control, e.g. via chromatin remodelling, as opposed to the short-range interactions used here.

#### Characterisation of various TATA-box promoters and modifications to make them both inducible and repressible

With an optimal design selected for TAL-ADs and their binding site configuration we next looked at different minimal TATA-box promoters with which to pair these two. Although the minimal CYC1 promoter has been extensively used in past studies, we wanted to assess other alternatives in case it was complicated to recode the core promoter region to enable TALE-based repression. The ideal promoter for our study should be capable of strong activation from 3 upstream TALE binding sites and repressible from 1 or 2 TALE binding sites in the core promoter region. Ideally, both the basal (OFF) and activated (ON) expression levels should not change when bases are mutated in the core promoter region to recode it for repression by a different TALE regulator.

Our initial screening showed that all of the minimal core promoters except pRNR1 and pSAC6 had low basal expression that could be significantly increased by activation by upstream TAL-AD binding. From these, the promoters with the least expression when unactivated (pLEU2 and pCYC1) would instinctively be the best choices to take forward. However, there is also an argument that promoters with higher basal output would be better suited as the basis for STAR regulation. If the unactivated promoter drives very low levels of expression then repression of the unactivated promoter does not offer a different regulation. If a promoter with low-to-medium strength basal expression is instead used then this offers the ability for down-regulation as well as up-regulation, depending on which TAL-AD binds. For this reason we pursued pCUP1 and pPHO5 beyond the initial screening and theoretically could have done the same with pCYC7 too.

The PHO5 core promoter was an especially interesting case, as this shows low basal output in glucose media, which is significantly activated in galactose media even without any TALE proteins present. We were not aware of the galactose-reponsiveness of this promoter but it could have potentially offered an extra layer of regulation for our system and it would be interesting to explore further in future studies. Unfortunately, as we moved forward with this promoter, it became clear that its core region is not tolerable to sequence changes in the region we selected. In fact it was very peculiar that the PHO5 minimal promoter lost activity after insertion of a BB21 binding site, as previous work had shown that deletion of the region where we created the site can actually increase transcriptional output by 50%[285].

### Characterisation of two STAR promoters

The work of this chapter resulted in two engineered TATA-box containing promoters with upstream activation sites and core promoter repression sites designed for the binding of different optimised TAL-AD proteins. The two engineered promoters were based on the CYC1 and CUP1 minimal promoters, of which the latter appeared to be better suited for STAR regulation.

The STAR promoter based on pCYC1 showed lower basal expression when uninduced and was capable of good activation from upstream binding of the TAL-AD, resulting in an increase in gene expression of close to two orders of magnitude. Unfortunately, repression by binding of the TAL-AD only shows a decreased expression for the highly-induced state, and even then, it is a very minor effect. Sadly, this means that this promoter is not well suited for STAR regulation as only activation is seen, and repression is not really evident. Interestingly, even binding of the TALE absent of any AD to the core promoter has almost no effect on the resulting expression levels. One would have assumed that steric repression from TALE binding to the core would have repressed expression, but this is not the case. Perhaps in future work, the placement of the TAL21 binding site within the pCYC1 core region could be explored to determine an optimal location to ensure strong repression.

The STAR promoter based on pCUP1 compares much more favourably to pCYC1 with respect to the desired characteristics of both activation and repression. In induced conditions, the upstream binding of the TAL-AD gives a strong increase in expression to levels, comparable to those seen with activated pCYC1 that will be perfect for the ON state of a switch circuit. Unfortunately, in uninduced conditions the binding of the TAL-AD to the repression site within the core promoter region does not lead to measurable decreased gene expression. As suggested for pCYC1 above, this could potentially be improved in future work by optimising the placement and orientation of the TAL binding site within the core promoter region.

Significant (17 fold) repression of expression was seen when the core promoter was bound by the TALE without an activation domain. And interestingly enough, some activation of expression was also seen when the TALE without an activation domain bound to the promoter upstream sites, especially in the CYC1 promoter (3 fold induction). Because of this we can imagine a STAR device being built using this promoter and just plain TAL proteins without any fused activation domains. From the results shown here, both the induced and uninduced promoter would be efficiently repressed to near autofluorescence levels by TAL binding at the core promoter. Unfortunately, while possible, activation would not be especially strong. For this system to work for our needs, we would probably need to further increase activation from the TAL. However, as we have seen, this would not be easy to achieve without concurrently negating the ability to repress the promoter.

Ultimately the most likely route to improving the pCUP1-based STAR promoters would be to move to a system with weaker activation domains (e.g. VP16) in combination with more upstream binding sites, while optimising the position and orientation of the repression binding sites within the core promoter. It would be especially useful if a phenotypic screen could be devised for the second part of this optimisation, so that a large number of sequence position variants could be assessed at once and the most effective designs could quickly be selected, e.g. by fluorescence-activated cell sorting (FACS) cytometry.

### Implications for the future creation of a bistable switch using STAR regulation

The ultimate goal of this chapter was to generate a new paradigam for modular regulation for yeast synthetic biology in order to simplify the design of a genetic circuit for a robust bistable switch. Unforunately in the time afforded for this work, we were never able to attempt to realise the desired genetic circuit, although with the engineered CUP1-based promoter, we are now close to having the tools available to do this.

As it stands, it is unclear whether the characteristics of the engineered CUP1-based promoters are good enough for the creation of a bistable circuit. In the implementation with TALs fused to activation domains, the promoter is excellently activated but poorly repressed (except in the activated state). Conversely, when no activation domains are used with the TALs, the activation is weak, but the repression is good. Using our CUP1-based promoter as the basis for a second promoter with the binding sites swapped around would not be expected to pose major problems. Theoretically, hundreds of orthogonally-regulated STAR promoters could be built using the pCUP1 core, so long as enough usable and orthogonal TAL proteins existed.

Perhaps for further work we should turn to gene expression modelling before undertaking any experimental improvements or implementations. Mathematical modelling may provide an answer as to whether a bistable switch built with our current genetic parts would stand a realistic chance of success. To produce a useful model, we may need to first perform more characterisation of the performance of selected promoters with varying concentrations of the TAL proteins. This will be needed as the model will require information on the induction and repression characteristics of the promoter at various concentrations of the two TALs or two TAL-ADs.

## 5.5 Conclusion

In this chapter, we used the synthetic biology design, build, test-cycle to produce pairs of engineered transcription factors and promoters that are capable of simultaneous activation and repression. With a view to incorporating these into the design of future synthetic genetic circuits in yeast, we intentionally followed a modular framework so that once we had a single successful design, this could quickly be used to produce many further orthogonal regulators. We determined that adding both activation and repression into a short constitutive promoter was not feasible and instead we moved to engineering these features into a selection of minimal core regions from natural TATA-box containing promoters. An optimal configuration was determined for TAL-effector based transcription factors capable of activating these promoters, but adequate repression was not seen when these were targeted to the promoter core regions in order to have repressive effects. A final engineered promoter based on the core sequence of pCUP shows promise for simultaneous activation and repression regulation and with further improvements could form the basis for regulated promoters within an optimised genetic circuit encoding a robust bistable switch.

# 6. Discussion

In this thesis, we have presented foundational work towards the expansion of the regulatory repertoire available for engineering synthetic genetic circuits in *S. cerevisiae*. We aimed to develop novel parts and tools that will allow richer functionality and increased predictability while also showing greater robustness and having reduced burden on the host. In this chapter we discuss in what way and to what extent these aims were achieved.

Firstly, we discuss the outcomes of Chapter 3, where we devised a strategy where designed hairpins are placed in the 5'UTR of mRNA transcripts to modulate their translation efficiency. In this chapter we showed that translation levels are reduced with increasing strength of the hairpin and that the magnitude of this effect is predictable, based on the calculated folding energy of the local RNA. In addition, we showed that this approach appears to be context independent and as such is applicable to a variety of promoters and ORFs. Thus this work realises our aim of creating a generalised method to predictably adjust the expression strengths from any promoter in yeast and offers a new tool for yeast synthetic biology.

In Chapter 4, we tested approaches to utilise transcriptional interference to apply additional regulation layers, ideally reinforcing a model bistable switch genetic circuit. We showed that transcriptional interference does indeed reduce transcriptional activity in a way we anticipated. Unexpectedly, however, translation in the model system was completely abolished. Based on the results in Chapter 3 we hypothesised that significant secondary structure in the 5'UTRs of the involved genes could cause the lack of expression. In this section we discuss the variety of potential solutions that were tested and what could be tried next to fulfil the potential of exploiting transcriptional interference in genetic circuits designs.

Finally, we also discuss Chapter 5, where we worked towards synthetic transcription factors that are capable of both transcriptional activation and repression, depending on the binding location at the promoter. This type of behaviour has not been applied generally in synthetic transcription factors, but is seen relatively widely in nature. It has also been predicted in theoretical work to have a positive impact on circuit robustness, response times and simplicity, as it eliminates the need for inverter motifs in genetic circuits. We discuss our efforts towards the creation of a promoter that can both be activated and repressed by a synthetic transcription factor coupled to an activation domain. We also speculate how repression in this system can be improved to make it more functional in practical applications.

## 6.1 Tuning of expression at the 5'UTR

We consider the results presented in Chapter 3 to be the most successful body of work presented in this thesis and this will form a publication in its own right that describes our method for predictable tuning of protein expression levels in yeast. Here, we look at how this was achieved and how we envision this being applied in practice. Furthermore, we consider what can be done to improve this method further, as well as opportunities for applications in other organisms.

To achieve predictable tuning of protein levels, we engineered the 5'UTR of mRNA transcripts to contain a hairpin structure that interferes with translation initiation. Using knowledge about the base pairing properties of RNA and existing software tools that predict the secondary structures of known mRNA sequences, we were able to establish a relationship between the folding energy of the mRNA 5'UTR and the translation efficiency. This relationship could then be used to predict the distribution of expression levels from various designed libraries of RNA structures synthesised to contain degenerate bases. By doing this, we established that a library with an average folding energy of -32.2 kcal/mol and a Gaussian distribution between -18 and -46 kcal/mol resulted in a near uniform sampling of measured yeGFP outputs in the entire spectrum of expression strengths that exist between the autofluorescence levels of wild-type cells and the full strength of the original very strong promoter.

We further investigated how inclusion of tetraloops affected the predictability of the method. Tetraloops are unusually stable loop sequences in the hairpin structure, that can act as anchors for RNA binding proteins. These proteins offer exciting possibilities for additional regulation on the targeted mRNAs. However, we found that the inclusion of tetraloops severely interfered with the predictability of translation efficiency. For this reason we excluded tetraloops from further experiments.

To confirm our assertion that the hairpin structures mainly affected protein expression at the translational level, not the transcriptional level, we performed qPCR on a set of constructs to determine their RNA levels. We tested transcript levels for mRNAs with each of the following structures in the 5'UTR: a strong hairpin with a tetraloop, a strong hairpin without a tetraloop, a weak hairpin and the wild-type sequence. Transcript levels for all tested constructs were found to be comparable, indicating that changes in protein output for these transcripts are unlikely to be caused by events at the level of transcription.

To test the robustness and context dependency of the method, we tested whether changes in upstream and downstream parts affected the predictability of the output protein levels. Six different constitutive promoters of various strengths were tested with a library with an average folding energy of -28.9 kcal/mol. As expected, this resulted in near uniform distributions of expression levels, especially in medium strength promoters and stronger. For downstream sequence robustness we tested a variety of libraries with a yeGFP and mRuby2 reporter with near identical results. These results together show that predictability of the method is not affected by changes in upstream or downstream sequence and that our method provides a robust framework for protein level tuning.

Finally, we tested the method in a real-world application to compare it to existing methods for promoter library generation. Using our newly developed method, we re-created a library based on a LacI repressible, GAL1-based promoter. This library had previously been created using targeted mutagenesis of the core promoter region but our results showed that our method was superior. We determined that we required ten times fewer colonies to be screened in order to create a library of similar size and expression distribution. Additionally, the repression characteristics of the original promoter were considerably less affected by the library sequence in our method. Lastly, the lead time for library creation and analysis from start to finish was only one week, which is decidedly shorter than other methods currently in use.

Taken together, the results in Chapter 3 demonstrate the creation of a robust and predictable method for expression library creation. It is an excellent example how the design-build-test cycle can lead to predictable engineering in biology. Moreover, we expect this method will be applied in other design cycles in order to improve the predictability of biological parts. We expect this method to be especially effective in applications that require the use of a set promoter at unknown expression strength. The method can be applied to tune the output of the chosen promoter, without affecting the function of any regulatory sequences present in the promoter. The hairpin sequence can even be designed to preserve any existing regulatory sequences, should they be downstream in the promoter and as such present in the 5'UTR. Using our method a 5'UTR library can rapidly be generated and tested within the intended circuit. Because promoter library generation has previously been unpredictable and time consuming, the focus has been on the testing of a small set of individual promoters sequentially to first determine the suitable part in a circuit. With the increased predictability and faster generation of libraries afforded by our new method, we envision a move away from sequential testing, towards massive parallel testing in order to find the optimal part. This will lead to more rapid progression of the design cycle and a more optimal selection of expression strengths, ultimately leading to better circuit function in synthetic biology.

An aspect that has been left unexplored in our work is whether the method can also be applied to control or tune noise in gene expression. It is well established in prokaryotes that changes in transcription versus translation rates have different impacts on the cell-to-cell variability of protein expression[136–139,286]. Specifically, increases in transcriptional efficiency have a much lower impact on the increase of expression noise compared to increases in translational efficiency. In other words, a construct with a strong promoter and weak RBS shows substantially less noise in gene expression than a construct with a weak promoter and strong RBS. This stems from the fact that transcription initiation occurs in bursts, while translation is typically seen to be a continuous process. A strong RBS with a weak promoter amplifies these transcriptional bursts, while a weak RBS with a strong promoter dampens them, with corresponding effects in gene expression noise.

In eukaryotes this concept has been largely unexplored, because the Kozak-sequence, which is the eukaryotic equivalent of RBS sequences, does not confer the same dynamic range in terms of expression levels compared to the RBS in prokaryotes. Since our method of expression tuning using hairpins acts at the level of translation initiation, it now opens up the opportunity for similar studies into gene expression noise in eukaryotes. Transcriptional bursting is a well-established phenomenon in eukaryotic organisms, as it is in prokaryotes. It is therefore not unreasonable to expect that a reduction in translational efficiency will have similar noise-dampening effects in eukaryotes. Since stochasticity in gene expression has been shown to have a concrete impact on the behaviour of genetic circuits, e.g. by increasing the uninduced switching rate in bistable switches[287], the introduced method has the potential to contribute to robustness in synthetic circuits in yet another way.

This leads to the question of whether this method could be applied in other eukaryotes as well. With the existence of well established tools such as the RBS Calculator and riboswitches, the development of a method for protein tuning using mRNA hairpins would be superfluous in prokaryotes[288]. In fact, a comparable system already exists[289]. However, these kinds of tools do not exist in eukaryotes and it would therefore be pertinent to investigate how transferable our method is to other relevant eukaryotic model organisms, such as *Arabidopsis thaliana* and mammalian systems including commonly used human cell lines.

The fact that many of the components of the yeast cellular machinery have close homologs in higher eukaryotes indicates that it is likely that the underlying mechanism of translation inhibition by hairpins is shared between all eukaryotes. We therefore expect that our method could be adapted to work in higher eukaryotes. To do this one will likely have to re-establish the relationship between the folding energy of the hairpin and the expression level, since subtle differences in cellular machinery can change the shape of the transfer function. Reports also disagree on the folding strengths where repression starts to play a role in mammalian systems. According to some, translation is not negatively affected until folding strengths of -50 kcal/mol are reached[97,290]. This is considerably stronger than what we have reported for yeast. On the other hand, some report that translation efficiency decreases abruptly as hairpin stabilities increase from -25 to -35 kcal/mol, which is in line with our findings[196].

Additionally, it will also need to be investigated whether the location of the hairpin plays a more prominent role in mammalian systems than in yeast. In mammalian cells there have been reports that repression from a hairpin gets stronger when it is situated near the 5'cap of the mRNA. The effect of a shift of as little as 9 base-pairs can result in a change in translation rate of over 50-fold[196]. In yeast, on the other hand, it is reported that the hairpin has a stronger effect when situated near the start codon, although it is unclear whether the effect is of similar magnitude.

This brings us back to the original system in yeast. Several aspects could be more fully explored to increase the accuracy and predictive strength of the method. Folding energy alone does not fully predict expression strength and the current equation under- and overestimates expression strengths by several-fold in certain cases. As mentioned above, repression by hairpins in yeast is reported to have a stronger effect when placed near the start codon. However, we have not characterised this effect in our system. Our current libraries contain between 6 and 10 base pairs of spacing between the stem of the hairpin and the start codon. It would be interesting to see what effect changing of this distance has on the strength of repression. Likewise, it would be interesting to similarly investigate the distance of the hairpin to the 5' cap of the mRNA. This could lead to the incorporation of new terms into the logistic relationship we established between the folding energy and expression strength, and improve the prediction accuracy.

Another way prediction accuracy can be improved is to consider more explicitly the G/C content of the hairpin. There are indications that G/C pairs contribute disproportionally to the repression, relative to their thermodynamic contribution[204]. We could investigate whether the number of G/C pairs, or the length of G/C pair stretches within the hairpin offers any further predictive power towards the expression strength.

Continuing in this frame of mind, it would be relevant to investigate the precise thermodynamic contribution of tetraloops in this system. The indirect estimation that we derived from our experiments does not correspond to values found in literature. We could look at whether the direct RNA context of the tetraloop plays a role, or whether the contribution towards repression is disproportionate relative to the folding energy, as we suspect is the case for G/C pairs.

In addition, a variety of alternative RNA structure prediction algorithms can be tested. We did not optimise the choice of software and simply chose a commonly used and well maintained software package. It is possible that other folding algorithms may offer a tighter fit between predictions and experimental measurements. For example, the RNAfold algorithm that was used here does not include special parameters for tetraloop contributions, while it is beyond doubt that they have a significant impact on hairpin stability. The current folding algorithm also does not take into account the possibility for pseudoknot folding either. Although not an intentional design feature, pseudoknots may form spontaneously within a subset of any library, reducing its predictability. Using an algorithm that takes these into account may to some extent limit issues of unpredictability. Finally, we could also investigate the merit of a new type of secondary structure prediction that takes into account the dynamics of RNA folding. In a concept called ribosome drafting, it has been postulated that the speed of refolding of RNA structures in the wake of a translating ribosome has an impact on the translational inhibition that a second ribosome experiences[291]. If the speed of refolding is low, a second ribosome will pass along more easily. Algorithms now exist that estimate the kinetics of (re)folding of the mRNA and this may inform whether a hairpin has a stronger or weaker effect on translation inhibition than could be expected from the folding energy alone[292,293].

One final consideration is the challenge of repeatability of measurements between labs. In order for the characterised relationship between folding and expression strength to carry any predictive power for third parties, it is a requirement that flow cytometry measurements between labs are comparable. It is known that absolute expression values vary wildly between labs, between users and even between experiments, because the settings and type of flow cytometry system have a large impact on the obtained results. We have attempted to overcome this problem through the definition of expression as a relative unit to autofluorescence levels. However, we have not managed to test how reproducible the results are for different settings and different flow cytometers. This could be an interesting avenue to pursue in future work. It also touches on the wider problem of making flow cytometry reproducible between labs. This is a problem actively being researched in labs around the world, but it is also surprisingly technical. It is certainly possible that some hurdles will need to be taken to allow precise expression level predictions to be made universally for all flow cytometry systems. However, at the very least our system can be used for relative predictions, based on some initial characterisation on the particular hardware at hand.

## 6.2 Engineered regulation using transcriptional interference

The second results chapter was the longest running project presented in this thesis, starting before the projects in the other chapters began. In this work we explored the possibility of using transcriptional interference as a supplemental mode of gene regulation and here we look at what was undertaken to achieve this and what more can be done to realise the vision that inspired this idea. Finally we reflect on the merit and potential of this project as a whole.

This project started with the observation that a robust bistable switch genetic circuit had never been created in yeast, despite several prior attempts. The bistable switch represents a core type of functionality in synthetic biological circuits that is essential for the progression of the field towards circuits with richer and more complex functions. We hypothesised that implementation of transcriptional interference could recover bistable behaviour in a poorly functioning bistable circuit that had been previously created and shown to be leaky. Furthermore, the extra layer of regulation that transcriptional interference would generate may make it possible to build bistable switches using the new generation of engineerable DNA-binding proteins that do not have cooperative binding properties (i.e. TAL-Effectors and dCas9).

Transcriptional interference was implemented in this circuit in the form of two convergent promoters placed directly opposite each other, in a head-to-head configuration. Interference was expected to arise from the displacement of the pre-initiation complex and Pol II at the opposite promoter by the transiting polymerases originating from the first promoter. Behind the opposing promoters lay on each side the coding sequence for mutually inhibiting repressors. This formed the core functionality of the bistable switch. Transcriptional interference and regular repression were expected to work in concert with this to achieve robust bistable behaviour.

In the first set of experiments, we established via qPCR that transcriptional interference was affecting transcript levels in ways that were consistent with expectations. We found that reducing the strength of one promoter in a system that originally contained head-to-head promoters of equal strength not only decreased transcript levels of the corresponding transcript. It also increased transcript levels arising from the opposing promoter. It is important to note that the promoter sequence of the opposing promoter was not changed and that elevated transcript levels from this promoter could be attributed exclusively to the effects of transcriptional interference.

Despite these encouraging results, it was also found that the transcripts generated by the circuit did not result in the expression of any protein. We hypothesised that the unusually long 5'UTR introduced by the head-to-head setup caused this issue. This notion is supported by the results from the first results chapter, where we found that strong hairpin structures in the 5'UTR of mRNA can completely abolish translation of the transcript. In the experiments following on from this premise, we attempted to solve this issue by eliminating or bypassing the long 5'UTR.

First, we tested a short promoter that shortened the 5'UTR from 500 bases to 160 bases. Next, we assessed if cap-independent translation through the incorporation of an internal ribosome entry site (IRES) could bypass the structure in the 5'UTR. Finally, we tested whether inclusion of the sequence of the reverse promoter within an intron could eliminate it from the transcript altogether, thereby restoring translation to levels matching the observed transcript levels. Disappointingly, none of the tested solutions showed elevated levels of translation. An attempt to functionalise the generated transcripts using dCas9 was also unsuccessful. We concluded that despite all of this work transcriptional interference fell short of our vision of a convenient additional mode of regulation that could readily be applied in synthetic circuits.

As we mentioned, a substantial amount of time and effort had been invested towards realising the goals of this project. With the conclusion that none of the aims that we set out with have been fully realised, we must ask ourselves the question of whether the concept of this project is fundamentally flawed and any attempt at making it work is essentially futile.

When addressing such fundamental questions, we turn to nature for insights. Evolution has resulted in an exceedingly rich and diverse set of solutions to challenges that are very similar to the ones that we choose to engage in. It is therefore frequently the case that gene circuit engineering problems can be informed from what is seen in nature. For this reason we chose a design that has equivalents in coliphages and bacteria. We were also very glad to see that during the time our project ran others working in bacteria produced synthetic biology papers also exploiting transcriptional interference[227]. This combined with our findings that transcript level changes were consistent with the expected effect in our head-to-head system leads us to conclude that the basic concept of TI is not fundamentally flawed.

However, we must conclude that the porting of this design from its natural and engineered instances in prokaryotes into a eukaryotic host presented a much greater challenge than we anticipated. The difficulties in making this system effective in yeast illustrates that gaps remain in our knowledge into the crucial differences between the two eukaryotic and prokaryotic domains. With hindsight, it may have been a better choice to construct a design based on the 'duelling' type of transcriptional interference, where inhibition arises from the collision of RNA polymerases as they transcribe through an open reading frame and into the opposing ORF. Several examples of this type of TI are known to occur naturally in *S. cerevisiae* and the effect is also observed on a genome wide scale[222,228,229]. This likely would have led to fewer challenges in the realisation of a functional circuit expressing protein.

The challenge of achieving protein outputs led to attempts at various creative solutions. However, the work raised the question at what point the implementation of TI stops being a tool for augmenting regulation and at what point it becomes a goal in and of itself. For example, should the implementation of the iPFY1 promoters have been the only promoters for which the system was found to work, it would have placed such tight restrictions on the design of the system that it lost its relevance to wider implementation. On the other hand, most of the possible solutions that were tested would have relieved restrictions present in the original design, such as the requirement for absence of CAT sequences within the promoter. In this sense, these would have further contributed to the implementation of TI as a tool, if successful.

Still, the question remains whether the continued efforts to get this system to work had been a good allocation of time and resources. The decision to cut one's losses and move on to other endeavours is a difficult one to make when a breakthrough could happen with every new experiment. Indeed, we felt this way for a significant part of the project. But possibly this decision should not have been made in a binary manner. Perhaps it would have been best to put this work on hold while new technologies emerged that could inform or aid our design. Looking back now, such technologies have indeed emerged and below we discuss what these are and how they could take this project forward.

The prime example of an emergent high-impact technology is the discovery of Cas9 as a gene-editing and programmable DNA binding tool. Towards the finish of the project, we applied this new technique with moderate success. We showed that repression could be achieved through targeting of a dCas9 protein to the core promoter. However, our aim of creating guide RNAs from the convergent promoter system was unsuccessful. This was related to the method of gRNA creation, which at the time had only been shown in yeast to occur from a special RNA polymerase III type promoter. This was incompatible with the more commonly used Pol II promoters that formed the basis of our system. However, new insights have since led to robust methods for gRNA synthesis from Pol II promoters using self-cleaving ribozymes[262]. This could be used to functionalise the mRNA that is produced in our current implementation of the head-to-head transcriptional interference system.

However, this approach does not solve the fundamental problem of lack of protein expression. We envisioned TI augmenting current regulation, not replacing it entirely, as it may not offer sufficient regulatory strength to implement the required functionality without the addition of protein-based regulation. As such, it would be more promising to return to a more fundamental solution. For example, our efforts to find functional IRES sequences to promote 5'cap-independent translation could be aided substantially by recent work by the Segal lab, that has systematically identified functional IRES sequences[120]. Although the work was done in mammalian cells, due to the homologies of the cellular machinery with yeast, the large number of strong cap-independent translation initiation elements that were identified are promising candidates for application in our work.

Alternatively, we could harness the recent advances that have been made in promoter design. New insights have led to the ability to design minimal synthetic promoters in yeast[29,31,53]. These promoters are substantially shorter than the 500 base-pairs that seem to have become the standard for natural promoters. Because forward engineering principles are used in the design of these promoters, they can be designed to the required specifications. In this case, that will include parameters that will optimise the length of the 5'UTR when the promoter is transcribed by the RNA polymerase from the opposing promoter. With the results from Chapter 3 in mind, this may also include optimisations with respect to folding energy of the 5'UTR, to avoid sequences that will form strong structures when the reverse complement of the promoter is transcribed.

These new promoters can be tested in the basic head-to-head design, in conjunction with previously tested optimisations such as inclusion in introns, or with new optimisations such as a newly characterised IRES. In particular, the improvements in cloning technology offered by the Yeast ToolKit cloning system now make this a realistic endeavour. Without this important innovation, constructing the suggested improvements would likely have presented a prohibitively resource-demanding amount of work. However, the increased throughput and lead-times offered by the YTK will now allow this suggested work to be completed in a fraction of the time that would have been needed using traditional restriction enzyme cloning.

In conclusion, the various technological advancements both separately and in combination offer substantial potential for the future realisation of transcriptional interference in genetic circuits. We hope that one day this can be reliably used to augment gene regulation by transcription factors in synthetic circuits.

## 6.3 Transcription factors capable of activation and repression

In Chapter 5, the final results chapter presented in this thesis, we describe the creation of synthetic transcription factors that can activate and repress, depending on their binding location on the promoter. Here we summarise how this was achieved and reflect on some of the design decisions that were made. We then discuss some of the fundamental weaknesses of the system and conclude with the introduction of an improved design that may address these weaknesses going forward.

The motivation for the creation of these transcription factors is that they obviate the need for inverter modules in synthetic circuits by allowing activation and repression from the same protein. In theoretical work, circuits incorporating such TFs have been shown to be more robust. In addition, by reducing the total number of components in a circuit, the circuit will be more amenable to modelling and simulation and therefore more predictable.

This project relied on a tight integration between promoter engineering and protein engineering. For the ideal results both the transcription factor and the promoter architecture needed to be optimised and tuned, as separate entities but also in relation to one another. This process started with the introduction of an upstream binding site into PFY1-based promoters to enable these to be inducibly activated. Over the course of a number of experiments it became apparent that this promoter is severely context dependent and does not allow for predictable application in synthetic circuits.

Next, in a series of optimisation experiments, we found that the choice of activation domain was important. Although the viral VP64-domain offered a higher activation strength than the native GAL4, it was shown to be less stable and more disruptive to the host cell than the GAL4 activation domain. We then optimised the activation strength of the TF by targeting TALE-GAL4 fusions to the minimal CYC1 promoter. We found that increasing the number of upstream TF binding sites led to an increase in maximum expression strength of the promoter. The effect did not scale linearly with the number of binding sites, suggesting that a point exists where inclusion of additional binding sites will not lead to a further increase in expression. We also found the activation strength to be higher when the TAL-effector is bound to the promoter in such a way that the side fused to activation domain is proximal to the core promoter.

With these optimisations in place, we proceeded with the construction of promoters with upstream binding sites for activation and binding sites within the core promoter for steric repression. To increase the chances of finding a suitable promoter, we characterised a set of core TATA-box promoter regions selected from the host genome. The most promising candidates, the CYC1 and CUP1 promoters, were fully characterised in a system with activating and repressing transcription factors binding to their respective binding sites upstream, or within the core of the promoter.

The promoters reacted similarly to activation by binding of the TAL-effector to the upstream binding sites. Both were efficiently induced by binding of the TAL fused to an activation domain and partially induced by the binding of a TAL with no AD. Repression by binding to the core promoter did not give similar results in the two promoters, however. In pCYC1 repression was generally poor and under certain circumstances it was induced, rather than repressed, through the binding of a TAL-AD fusion to the core sites. In pCUP1, on the other hand, binding of a TAL with or without an AD led to repression in all tested conditions. Repression was especially strong in the fully induced state, which is encouraging for application of this promoter in our intended model-system: the bistable switch. Construction of the bistable switch was not achieved due to time constraints, and below we will reflect on the merit of attempting this in the future.

The early determination that the PFY1 promoter is too context dependent for application in the intended circuit is both a victory for the rigorous approach prescribed by synthetic biology and a testimony to the fact that context dependency is a critical factor in the design of biological circuits. The PFY1 promoter is TATA-less and it is known that TATA-less promoters are generally constitutive promoters that are not heavily regulated. It can therefore be said that the effort to implement both positive and negative regulation in this promoter was somewhat naive. In hindsight we can only concur with this notion, although it is not surprising that our judgement was somewhat clouded by the fact that we had readily achieved repression in this promoter in earlier work[42]. Given the large number of TATA-less promoters in the yeast genome and the number of essential genes they express, the lack of understanding of these promoters by the community is remarkable. It is therefore not surprising that virtually all promoter engineering studies have been performed with TATA-box promoters.

Despite the achievement of creating a promoter that can both be activated and repressed by a synthetic transcription factor, some fundamental problems with the current design remain that cannot be overlooked. Primarily, the necessity for the transcription factor to be fused to an activation domain interferes with its ability to repress. In all tested cases, targeting a TF with an activation domain to the core promoter resulted in less repression compared to the same TF without an activation domain. We have suggested previously that the strength of the activation domain and precise binding location can be optimised to improve the performance. However, it is unlikely that this will completely solve the issue. Unintended activation at the repression site is therefore considered a fundamental issue with this design.

Other problems stem from the use of TAL-effectors. Despite the potential that is afforded by the programmable nature of TALEs, their application carries some practical issues. TALEs are particularly repetitive sequences in nature and as we have highlighted before, this has a negative impact on evolutionary stability. We have solved this here in a bespoke manner by manually codon-optimising the TALE sequence, but this approach is not scalable. It is conceivable that new kits with pre-codon optimised parts will become available in the future, resulting in non-repetitive DNA sequences. However, the fact remains that changing the binding specificity of a TAL-effector requires a substantial cloning effort, typically on the order of several days of lab work. In this context, we cannot avoid the comparison to the dCas9 protein, whose binding specificity can be changed simply by expressing a different gRNA.

In this thesis we applied dCas9 in Chapter 4, for the repression of TATA-box promoters through steric hindrance. We showed that dCas9 can act as a potent repressor when targeted to the core promoter with a complementary gRNA. At the inception of this project, designing regulation based on dCas9 was challenging because guide RNAs could only be expressed from particular Pol-III transcribed promoters. However, this problem has since been solved with the production of gRNA from Pol-II promoters using of self-cleaving ribozyme sequences[262].

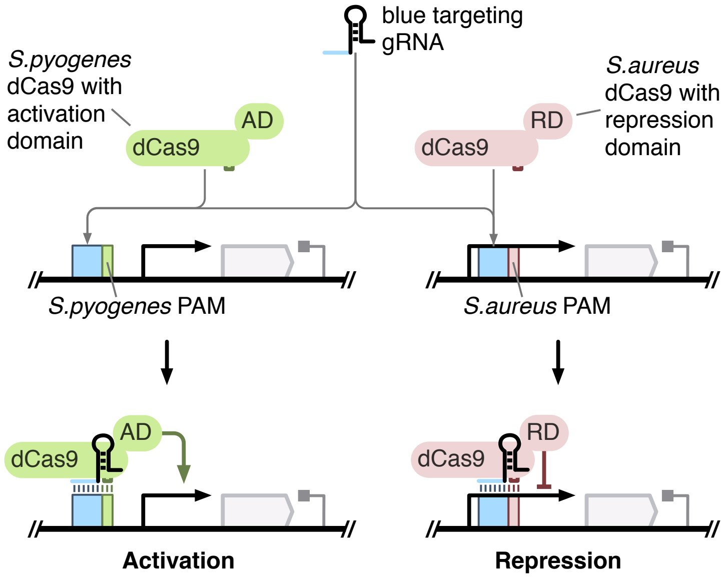


**Figure 6.1:** New design for the Simultaneous Transcription Activation and Repression (STAR) system. In this design, binding specificity mediated by the gRNA and the transcriptional effect mediated by dCas9 are separated. Two species of dCas9 with unique PAM recognition sequences are each fused to a different effector domain. dCas9 from *S.pyogenes* is fused to an activator domain, while the *S.aureus* version is fused to a repression domain. Both can be directed to a 'blue' target by the 'blue targeting gRNA', but only the dCas9 whose corresponding PAM is downstream of the target will bind that location. This allows the 'blue targeting gRNA' to activate one promoter and repress another.

One feature of the Cas9 family of proteins that has not been widely applied is the fact that proteins from different bacterial species recognise different PAM sequences. This allows the sequence specificity encoded by the gRNA to be separated from the protein that is recruited to that specific site through the use of different PAM sequences. Combining this with the fact that the different versions of dCas9 can be fused to different effector domains makes the dCas9 family ideally suited for our purposes. We now introduce an updated design of the STAR system, incorporating these beneficial features and solving the fundamental issue of unintended activation that affected the previous design.

The improved design of the STAR system is shown in **Figure 6.1**. This figure shows how the same guide RNA can lead to activation in one promoter and repression in the other. Rather than relying on steric hindrance, repression is realised by the inclusion of a repression domain fused to the dCas9 protein. The inclusion of a separate TF that mediates repression is made possible by the separation of binding specificity and transcriptional effect that is afforded by the use of different versions of the dCas9 protein. As described in the introduction of chapter 4 (**subsection 4.1.5** on page 113), the Cas9 protein can only bind its target sequence if it is accompanied by the correct PAM. Since the exact PAM sequences are species specific, we can attract different Cas9 proteins to different target sites by changing the PAM. By fusing an activation domain to dCas9 from *S. pyogenes* and a repression domain to dCas9 from *S. aureus*, we can target an activating transcription factor to a promoter using the *S. pyogenes* PAM while targeting the same recognition site with the same gRNA at a different promoter with a repressor by including the *S. aureus* PAM at that site.

Some conditions need to be met, in order for this approach to succeed. Both versions of the Cas9 protein need to be catalytically inactive. In the case of Cas9 from *S. aureus*, a 'dead' version does indeed exist[294]. In cases where an inactive version does not exist, the sequence homology between the various species may guide the effort of creating one. Additionally, the tracrRNA component of the gRNA needs to be recognised by the different dCas9s. There is some variation in tracrRNA sequence between the *S. aureus* and *S. pyogenes* versions. It is unclear to what extent Cas9 requires a specific sequence in the tracrRNA rather than the correct secondary structure. However, tracerRNAs of the *S. thermophilus* and *S. pasteurianus* CRISPR systems are highly similar and would likely be interchangeable[294]. The most elegant way to avoid interoperability problems, is to use two versions of the same dCas9 that have been engineered to target different PAM sites. These have been reported by Kleinstiver *et al* in 2015, demonstrating the flexibility and utility of *S. pyogenes* Cas9[295].

The implications of this approach are profound. It realises the potential of scalable transcription factor generation that is promised by technologies such as CRISPR/Cas9 and TAL-effectors. In this system the protein component of regulation can be expressed constitutively, while regulation essentially happens at the RNA level. This enhances the scalability of the system, because the burden of protein expression does not scale with the number of regulatory connections, but instead remains fixed. In addition, it will result in fast response times for the created circuits, since the slow step of protein expression is not required. Finally, for increased functionality additional dCas9 proteins from other species may be introduced that are fused to other types of effector domains. This will extend the functionality beyond straightforward activation and repression.

Considering the advantages of this new STAR design, it would seem that future resources are better spent developing this, rather than to continue developing TALE-based transcription factors. Although implementation of a bistable switch can be attempted with the STAR promoter developed in this work, this new design seems more likely to yield a functional circuit. In both the current and the new design, it will be beneficial to optimise the repression binding site. Optimisation of the activation site(s) had a large impact on the maximum activation. There is no reason not to expect optimisations of the repressor binding site to potentially have a similar effect on repression strength.

Elements from the current work can also be applied to this new design. The optimisations for the orientation and number of activator binding sites can be transferred and the most promising minimal promoters that were characterised in this work can also be tested in the new system. In addition to pCYC1 and pCUP1, other promoters that were discarded due to experimental capacity constraints such as pCYC7, pLEU2 and pPHO5 could also be tested. In principle only a single promoter is required, since once the system works, the binding sites can be changed at will to create a large number of alternatively regulated and orthogonal promoters. However, as we observed with the PFY1, CYC1 and CUP1 promoters, context effects and promoter intrinsic properties can greatly affect the behaviour in response to stimuli. For this reason we recommend a more thorough characterisation of the remaining promoters to find the one with the most favourable characteristics.

To decide what the most favourable characteristics are, we will likely need to turn to modelling in the future. Simulations can give meaningful answers to questions about the tolerable level of leakiness and the desired activation and repression strengths. These answers can be generated quickly through computational work. Although it is not guaranteed that the resulting answers are always valid, this can be verified in experiments in the wet lab. This strategy will likely be able to give more focus to the characterisation experiments than the wet lab work would do on its own.

In conclusion, we have successfully implemented simultaneous transcription activation and repression in an implementation based on TAL-effectors. A fundamental limitation of this design is the requirement for the repressing transcription factor to carry an activation domain, thereby reducing repression efficiency. To solve this we have suggested an improved design that is based on the dCas9 technology. We are optimistic that in the future this new design can improve upon the weaknesses of our current system, while retaining its strengths, such as characterised promoters, activation domain and activation site optimisations. The new design has the potential to be scalable, low burden and fast. Realising this potential would mean a significant advancement in the field of synthetic biology and beyond.

## 6.4 Reflections on future directions

Having already discussed the specific future work for each chapter, here we focus on future directions of the overall work of this thesis and the relative importance of the constituent projects.

Since the three projects presented in this thesis have limited overlap, any further work will need to be done in prioritised manner and this should follow-on logically from the results attained. We consider the project for protein level tuning using hairpins in the 5'UTR of mRNAs as the most successful body of work. With some minor modifications, such as the inclusion of a Python script developed to allow third parties to easily generate hairpin libraries, this work will form a manuscript for immediate publication. We envision this system being used in yeast similarly to how the RBS Calculator is used in *E. coli* and we are looking forward to seeing how others apply it. As this is almost ready-to-go, this work will have the highest priority going forward.

Of the remaining two chapters, we have invested the most time and effort into developing the work related to transcriptional interference. Sadly, this work has not yielded a system that is readily applicable in the design of synthetic circuits. Recent advances in the field of synthetic biology have yielded new knowledge and tools that offer promising leads for future work. However, so far, every new solution that we have tested has initially looked promising, yet no solution has been successful. We therefore conclude that we must prioritise further development of the STAR project in favour of the TI work.

The dCas9-based system proposed for the future development of synthetic transcription factors for simultaneous activation and repression in biological circuits is highly-promising and builds on successful work. The promise of massively-parallelised creation of orthogonal transcription factors with low burden and fast regulation speed is phenomenal, especially if they allow both activation and repression. For this reason the decision to prioritise this work over TI based regulation is not a complicated one to make. Regardless of whether it is developed by us or by others, dCas9-based regulation will have a great impact on the field going forward. We expect to see a boom in regulatory circuits of unprecedented complexity in the very near future. It will further invigorate our field which has been gaining momentum from other technological advances in the recent past.

Another example of such a technological advancement is the Yeast ToolKit, which we have used extensively in this thesis. The increased throughput and efficiency of this system over traditional cloning methods has brought about a step change in the assembly efficiency of DNA constructs. In addition, through changes in the mechanism for chromosomal integration, this method results in dramatically lower incidence of multiple integration events. This facilitates the fast and efficient characterisation of constructs, further speeding up the development process.

In summary we are optimistic about the field of synthetic biology going forward, and especially the construction of more complex, yet predictable genetic circuits in yeast. The field thrives on technological advancements within the field, such as the YTK and dCas9, but also needs advances from related fields to result in decreasing costs and increasing throughput for DNA synthesis and sequencing. With these developments, we hope soon to be working towards an equivalent to Moore's law, where we see the number of regulatory connections used in synthetic biological circuits routinely double every few years in the near future.

# Bibliography

[1] T. S. Gardner, C. R. Cantor, and J. J. Collins, “Construction of a genetic toggle switch in escherichia coli.,” *Nature*, vol. 403, pp. 339–342, Jan 2000.

[2] M. B. Elowitz and S. Leibler, “A synthetic oscillatory network of transcriptional regulators,” *Nature*, vol. 403, pp. 335–338, 01 2000.

[3] D. E. Cameron, C. J. Bashor, and J. J. Collins, “A brief history of synthetic biology,” *Nat Rev Microbiol*, vol. 12, pp. 381–90, May 2014.

[4] E. J. Olson and J. J. Tabor, “Optogenetic characterization methods overcome key challenges in synthetic and systems biology,” *Nat Chem Biol*, vol. 10, pp. 502–11, Jul 2014.

[5] A. A. K. Nielsen, B. S. Der, J. Shin, P. Vaidyanathan, V. Paralanov, E. A. Strychalski, D. Ross, D. Densmore, and C. A. Voigt, “Genetic circuit design automation,” *Science*, vol. 352, p. aac7341, Apr 2016.

[6] P. E. M. Purnick and R. Weiss, “The second wave of synthetic biology: from modules to systems,” *Nat Rev Mol Cell Biol*, vol. 10, pp. 410–22, Jun 2009.

[7] B. Canton, A. Labno, and D. Endy, “Refinement and standardization of synthetic biological parts and devices,” *Nat Biotechnol*, vol. 26, pp. 787–93, Jul 2008.

[8] M. E. Lee, W. C. DeLoache, B. Cervantes, and J. E. Dueber, “A highly characterized yeast toolkit for modular, multipart assembly,” *ACS Synth Biol*, vol. 4, pp. 975–86, Sep 2015.

[9] N. Agmon, L. A. Mitchell, Y. Cai, S. Ikushima, J. Chuang, A. Zheng, W.-J. Choi, J. A. Martin, K. Caravelli, G. Stracquadanio, and J. D. Boeke, “Yeast golden gate (ygg) for the efficient assembly of s. cerevisiae transcription units,” *ACS Synth Biol*, vol. 4, pp. 853–9, Jul 2015.

[10] B. C. Stanton, A. A. K. Nielsen, A. Tamsir, K. Clancy, T. Peterson, and C. A. Voigt, “Genomic mining of prokaryotic repressors for orthogonal logic gates,” *Nat Chem Biol*, vol. 10, pp. 99–105, Feb 2014.

[11] V. K. Mutalik, J. C. Guimaraes, G. Cambray, C. Lam, M. J. Christoffersen, Q.-A. Mai, A. B. Tran, M. Paull, J. D. Keasling, A. P. Arkin, and D. Endy, “Precise and reliable gene expression via standard transcription and translation initiation elements,” *Nat Methods*, vol. 10, pp. 354–60, Apr 2013.

[12] S. Cardinale and A. P. Arkin, “Contextualizing context for synthetic biology–identifying causes of failure of synthetic biological systems,” *Biotechnol J*, vol. 7, pp. 856–66, Jul 2012.

[13] D. Del Vecchio, A. J. Ninfa, and E. D. Sontag, “Modular cell biology: retroactivity and insulation,” *Mol Syst Biol*, vol. 4, p. 161, 2008.

[14] D. Mishra, P. M. Rivera, A. Lin, D. Del Vecchio, and R. Weiss, “A load driver device for engineering modularity in biological networks,” *Nat Biotechnol*, vol. 32, pp. 1268–75, Dec 2014.

[15] J. A. N. Brophy and C. A. Voigt, “Principles of genetic circuit design,” *Nat Methods*, vol. 11, pp. 508–20, May 2014.

[16] B. R. Jack, S. P. Leonard, D. M. Mishler, B. A. Renda, D. Leon, G. A. Suárez, and J. E. Barrick, “Predicting the genetic stability of engineered dna sequences with the efm calculator,” *ACS Synth Biol*, vol. 4, pp. 939–43, Aug 2015.

[17] A. S. Khalil, T. K. Lu, C. J. Bashor, C. L. Ramirez, N. C. Pyenson, J. K. Joung, and J. J. Collins, “A synthetic biology framework for programming eukaryotic transcription functions.,” *Cell*, vol. 150, pp. 647–658, Aug 2012.

[18] T. Cermak, E. L. Doyle, M. Christian, L. Wang, Y. Zhang, C. Schmidt, J. A. Baller, N. V. Somia, A. J. Bogdanove, and D. F. Voytas, “Efficient design and assembly of custom talen and other tal effector-based constructs for dna targeting,” *Nucleic Acids Research*, 04 2011.

[19] L. A. Gilbert, M. H. Larson, L. Morsut, Z. Liu, G. A. Brar, S. E. Torres, N. Stern-Ginossar, O. Brandman, E. H. Whitehead, J. A. Doudna, W. A. Lim, J. S. Weissman, and L. S. Qi, “Crispr-mediated modular rna-guided regulation of transcription in eukaryotes,” *Cell*, vol. 154, pp. 442–51, Jul 2013.

[20] J. Chappell, K. E. Watters, M. K. Takahashi, and J. B. Lucks, “A renaissance in rna synthetic biology: new mechanisms, applications and tools for the future,” *Curr Opin Chem Biol*, vol. 28, pp. 47–56, Oct 2015.

[21] B. A. Blount, T. Weenink, and T. Ellis, “Construction of synthetic regulatory networks in yeast.,” *FEBS Lett*, vol. 586, pp. 2112–2121, Jul 2012.

[22] H. Goetze, M. Wittner, S. Hamperl, M. Hondele, K. Merz, U. Stoeckl, and J. Griesenbeck, “Alternative chromatin structures of the 35s rrna genes in saccharomyces cerevisiae provide a molecular basis for the selective recruitment of rna polymerases i and ii,” *Mol Cell Biol*, vol. 30, pp. 2028–45, Apr 2010.

[23] O. Harismendy, C.-G. Gendrel, P. Soularue, X. Gidrol, A. Sentenac, M. Werner, and O. Lefebvre, “Genome-wide location of yeast rna polymerase iii transcription machinery,” *EMBO J*, vol. 22, pp. 4738–47, Sep 2003.

[24] I. M. Willis, “Rna polymerase iii. genes, factors and transcriptional specificity,” *Eur J Biochem*, vol. 212, pp. 1–11, Feb 1993.

[25] L. Schramm and N. Hernandez, “Recruitment of rna polymerase iii to its target promoters,” *Genes Dev*, vol. 16, pp. 2593–620, Oct 2002.

[26] S. Sainsbury, C. Bernecky, and P. Cramer, “Structural basis of transcription initiation by rna polymerase ii,” *Nat Rev Mol Cell Biol*, vol. 16, pp. 129–43, Mar 2015.

[27] C. Yang, E. Bolotin, T. Jiang, F. M. Sladek, and E. Martinez, “Prevalence of the initiator over the tata box in human and yeast genes and identification of dna motifs enriched in human tata-less core promoters,” *Gene*, vol. 389, pp. 52–65, Mar 2007.

[28] A. D. Basehoar, S. J. Zanton, and B. F. Pugh, “Identification and distinct regulation of yeast tata box-containing genes,” *Cell*, vol. 116, pp. 699–709, Mar 2004.

[29] S. Lubliner, I. Regev, M. Lotan-Pompan, S. Edelheit, A. Weinberger, and E. Segal, “Core promoter sequence in yeast is a major determinant of expression level,” *Genome Res*, vol. 25, pp. 1008–17, Jul 2015.

[30] K. A. Curran, N. C. Crook, A. S. Karim, A. Gupta, A. M. Wagman, and H. S. Alper, “Design of synthetic yeast promoters via tuning of nucleosome architecture,” *Nat Commun*, vol. 5, p. 4002, 2014.

[31] S. Lubliner, L. Keren, and E. Segal, “Sequence features of yeast and human core promoters that are predictive of maximal promoter activity,” *Nucleic Acids Res*, vol. 41, pp. 5569–81, Jun 2013.

[32] J. E. F. Butler and J. T. Kadonaga, “The rna polymerase ii core promoter: a key component in the regulation of gene expression,” *Genes Dev*, vol. 16, pp. 2583–92, Oct 2002.

[33] S. Hahn, E. T. Hoar, and L. Guarente, “Each of three "tata elements" specifies a subset of the transcription initiation sites at the cyc-1 promoter of saccharomyces cerevisiae,” *PNAS*, vol. 82, pp. 8562–6, Dec 1985.

[34] E. M. Furter-Graves and B. D. Hall, “Dna sequence elements required for transcription initiation of the schizosaccharomyces pombe adh gene in saccharomyces cerevisiae,” *Mol Gen Genet*, vol. 223, pp. 407–16, Sep 1990.

[35] Z. Zhang and F. S. Dietrich, “Mapping of transcription start sites in saccharomyces cerevisiae using 5’ sage,” *Nucleic Acids Res*, vol. 33, no. 9, pp. 2838–51, 2005.

[36] J. N. Kuehner and D. A. Brow, “Quantitative analysis of in vivo initiator selection by yeast rna polymerase ii supports a scanning model,” *J Biol Chem*, vol. 281, pp. 14119–28, May 2006.

[37] `https://mutagenetix.utsouthwestern.edu/phenotypic/phenotypic_rec.cfm?pk=399`, “Gene expression activation facilitated by the mediator complex..”

[38] P. Khosravi, V. H. Gazestani, L. Pirhaji, B. Law, M. Sadeghi, B. Goliaei, and G. D. Bader, “Inferring interaction type in gene regulatory networks using co-expression data,” *Algorithms Mol Biol*, vol. 10, p. 23, 2015.

[39] Q. Huang, C. Gong, J. Li, Z. Zhuo, Y. Chen, J. Wang, and Z.-C. Hua, “Distance and helical phase dependence of synergistic transcription activation in cis-regulatory module,” *PLoS One*, vol. 7, no. 1, p. e31198, 2012.

[40] Z. Zhang and J. C. Reese, “Redundant mechanisms are used by ssn6-tup1 in repressing chromosomal gene transcription in saccharomyces cerevisiae,” *J Biol Chem*, vol. 279, pp. 39240–50, Sep 2004.

[41] M. Levine and J. L. Manley, “Transcriptional repression of eukaryotic promoters,” *Cell*, vol. 59, pp. 405–8, Nov 1989.

[42] B. A. Blount, T. Weenink, S. Vasylechko, and T. Ellis, “Rational diversification of a promoter providing fine-tuned expression and orthogonal regulation for synthetic biology,” *PLoS One*, vol. 7, pp. e33279 EP –, 03 2012.

[43] T. Ellis, X. Wang, and J. J. Collins, “Diversity-based, model-guided construction of synthetic gene networks with predicted functions,” *Nat Biotech*, vol. 27, pp. 465–471, 05 2009.

[44] Z. Zaman, A. Z. Ansari, S. S. Koh, R. Young, and M. Ptashne, “Interaction of a transcriptional repressor with the rna polymerase ii holoenzyme plays a crucial role in repression,” *PNAS*, vol. 98, pp. 2550–4, Feb 2001.

[45] E. Maldonado, M. Hampsey, and D. Reinberg, “Repression: targeting the heart of the matter,” *Cell*, vol. 99, pp. 455–8, Nov 1999.

[46] W. Hanna-Rose and U. Hansen, “Active repression mechanisms of eukaryotic transcription repressors,” *Trends Genet*, vol. 12, pp. 229–34, Jun 1996.

[47] D. Kadosh and K. Struhl, “Repression by ume6 involves recruitment of a complex containing sin3 corepressor and rpd3 histone deacetylase to target promoters,” *Cell*, vol. 89, pp. 365–71, May 1997.

[48] H. Wang and D. J. Stillman, “Transcriptional repression in saccharomyces cerevisiae by a sin3-lexa fusion protein,” *Mol Cell Biol*, vol. 13, pp. 1805–14, Mar 1993.

[49] N. Schreiber-Agus, L. Chin, K. Chen, R. Torres, G. Rao, P. Guida, A. I. Skoultchi, and R. A. DePinho, “An amino-terminal domain of mxi1 mediates anti-myc oncogenic activity and interacts with a homolog of the yeast transcriptional repressor sin3,” *Cell*, vol. 80, pp. 777–86, Mar 1995.

[50] S. K. Kurdistani and M. Grunstein, “Histone acetylation and deacetylation in yeast,” *Nat Rev Mol Cell Biol*, vol. 4, pp. 276–84, Apr 2003.

[51] P. Prochasson, L. Florens, S. K. Swanson, M. P. Washburn, and J. L. Workman, “The hir corepressor complex binds to nucleosomes generating a distinct protein/dna complex resistant to remodeling by swi/snf,” *Genes Dev*, vol. 19, pp. 2534–9, Nov 2005.

[52] T. Kodadek, “How does the gal4 transcription factor recognize the appropriate dna binding sites in vivo?,” *Cell Mol Biol Res*, vol. 39, no. 4, pp. 355–60, 1993.

[53] H. Redden and H. S. Alper, “The development and characterization of synthetic minimal yeast promoters,” *Nat Commun*, vol. 6, p. 7810, 2015.

[54] M. Hampsey, “Molecular genetics of the rna polymerase ii general transcriptional machinery,” *Microbiol Mol*

*Biol Rev*, vol. 62, pp. 465–503, Jun 1998.

[55] S. T. Smale and J. T. Kadonaga, “The rna polymerase ii core promoter,” *Annu Rev Biochem*, vol. 72, pp. 449–79, 2003.

[56] I. Kamenova, L. Warfield, and S. Hahn, “Mutations on the dna binding surface of tbp discriminate between yeast tata and tata-less gene transcription,” *Mol Cell Biol*, vol. 34, pp. 2929–43, Aug 2014.

[57] M. Seizl, H. Hartmann, F. Hoeg, F. Kurth, D. E. Martin, J. Söding, and P. Cramer, “A conserved ga element in tata-less rna polymerase ii promoters,” *PLoS One*, vol. 6, no. 11, p. e27595, 2011.

[58] M. Angermayr, U. Oechsner, and W. Bandlow, “Reb1p-dependent dna bending effects nucleosome positioning and constitutive transcription at the yeast profilin promoter,” *J Biol Chem*, vol. 278, pp. 17918–26, May 2003.

[59] P. Richard and J. L. Manley, “Transcription termination by nuclear rna polymerases,” *Genes Dev*, vol. 23, pp. 1247–69, Jun 2009.

[60] W. Luo, A. W. Johnson, and D. L. Bentley, “The role of rat1 in coupling mrna 3’-end processing to transcription termination: implications for a unified allosteric-torpedo model,” *Genes Dev*, vol. 20, pp. 954–65, Apr 2006.

[61] Z. Guo and F. Sherman, “3âĂš-end-forming signals of yeast mrna,” *Trends in Biochemical Sciences*, vol. 21, pp. 477–481, 12 1996.

[62] J. H. Graber, G. D. McAllister, and T. F. Smith, “Probabilistic prediction of saccharomyces cerevisiae mrna 3’-processing sites,” *Nucleic Acids Res*, vol. 30, pp. 1851–8, Apr 2002.

[63] K. A. Curran, A. S. Karim, A. Gupta, and H. S. Alper, “Use of expression-enhancing terminators in saccharomyces cerevisiae to increase mrna half-life and improve gene expression control for metabolic engineering applications,” *Metab Eng*, vol. 19, pp. 88–97, Sep 2013.

[64] K. A. Curran, N. J. Morse, K. A. Markham, A. M. Wagman, A. Gupta, and H. S. Alper, “Short synthetic terminators for improved heterologous gene expression in yeast,” *ACS Synth Biol*, Feb 2015.

[65] B. I. Osborne and L. Guarente, “Mutational analysis of a yeast transcriptional terminator,” *PNAS*, vol. 86, pp. 4097–101, Jun 1989.

[66] S. Heidmann, B. Obermaier, K. Vogel, and H. Domdey, “Identification of pre-mrna polyadenylation sites in saccharomyces cerevisiae,” *Mol Cell Biol*, vol. 12, pp. 4215–29, Sep 1992.

[67] S. Irniger, C. M. Egli, and G. H. Braus, “Different classes of polyadenylation sites in the yeast saccharomyces cerevisiae,” *Mol Cell Biol*, vol. 11, pp. 3060–9, Jun 1991.

[68] S. Hocine, R. H. Singer, and D. Grünwald, “Rna processing and export,” *Cold Spring Harb Perspect Biol*, vol. 2, p. a000752, Dec 2010.

[69] J.-P. Hsin and J. L. Manley, “The rna polymerase ii ctd coordinates transcription and rna processing,” *Genes Dev*, vol. 26, pp. 2119–37, Oct 2012.

[70] S. F. de Almeida and M. Carmo-Fonseca, “The ctd role in cotranscriptional rna processing and surveillance,” *FEBS Lett*, vol. 582, pp. 1971–6, Jun 2008.

[71] D. L. Bentley, “Coupling mrna processing with transcription in time and space,” *Nat Rev Genet*, vol. 15, pp. 163–75, Mar 2014.

[72] A. Ramanathan, G. B. Robb, and S.-H. Chan, “mrna capping: biological functions and applications,” *Nucleic Acids Res*, vol. 44, pp. 7511–26, Sep 2016.

[73] C. L. Hsu and A. Stevens, “Yeast cells lacking 5’–>3’ exoribonuclease 1 contain mrna species that are poly(a) deficient and partially lack the 5’ cap structure,” *Mol Cell Biol*, vol. 13, pp. 4826–35, Aug 1993.

[74] J. D. Lewis and E. Izaurralde, “The role of the cap structure in rna processing and nuclear export,” *Eur J Biochem*, vol. 247, pp. 461–9, Jul 1997.

[75] E. C. Merkhofer, P. Hu, and T. L. Johnson, “Introduction to cotranscriptional rna splicing,” *Methods Mol Biol*, vol. 1126, pp. 83–96, 2014.

[76] J. Parenteau, M. Durand, S. Véronneau, A.-A. Lacombe, G. Morin, V. Guérin, B. Cecez, J. Gervais-Bird, C.-S. Koh, D. Brunelle, R. J. Wellinger, B. Chabot, and S. Abou Elela, “Deletion of many yeast introns reveals a minority of genes that require splicing for function,” *Mol Biol Cell*, vol. 19, pp. 1932–41, May 2008.

[77] M. Spingola, L. Grate, D. Haussler, and M. Ares, Jr, “Genome-wide bioinformatic and molecular analysis of introns in saccharomyces cerevisiae,” *RNA*, vol. 5, pp. 221–34, Feb 1999.

[78] D. R. Semlow and J. P. Staley, “Staying on message: ensuring fidelity in pre-mrna splicing,” *Trends Biochem Sci*, vol. 37, pp. 263–73, Jul 2012.

[79] T. A. Clark, C. W. Sugnet, and M. Ares, Jr, “Genomewide analysis of mrna processing in yeast using splicing-specific microarrays,” *Science*, vol. 296, pp. 907–910, 05 2002.

[80] K. Juneau, C. Palm, M. Miranda, and R. W. Davis, “High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing,” *PNAS*, vol. 104, pp. 1522–7, Jan 2007.

[81] J. A. Pleiss, G. B. Whitworth, M. Bergkessel, and C. Guthrie, “Transcript specificity in yeast pre-mrna splicing revealed by mutations in core spliceosomal components,” *PLoS Biol*, vol. 5, p. e90, Apr 2007.

[82] F. C. Oesterreich, L. Herzel, K. Straube, K. Hujer, J. Howard, and K. M. Neugebauer, “Splicing of nascent rna coincides with intron exit from rna polymerase ii,” *Cell*, vol. 165, pp. 372–381, April 2016.

[83] P. Ma and X. Xia, “Factors affecting splicing strength of yeast genes,” *Comparative and Functional Genomics*, p. 212146, 2011.

[84] L. B. Crotti and D. S. Horowitz, “Exon sequences at the splice junctions affect splicing fidelity and alternative splicing,” *PNAS*, vol. 106, pp. 18954–9, Nov 2009.

[85] M. Ares, Jr, L. Grate, and M. H. Pauling, “A handful of intron-containing genes produces the lion’s share of yeast mrna,” *RNA*, vol. 5, pp. 1138–9, Sep 1999.

[86] K. L. Fox-Walsh, Y. Dou, B. J. Lam, S.-p. Hung, P. F. Baldi, and K. J. Hertel, “The architecture of pre-mrnas affects mechanisms of splice-site pairing,” *PNAS*, vol. 102, pp. 16176–16181, 11 2005.

[87] D. A. Sterner, T. Carlo, and S. M. Berget, “Architectural limits on split genes,” *PNAS*, vol. 93, pp. 15081–5, Dec 1996.

[88] A. Mougin, A. Grégoire, J. Banroques, V. Ségault, R. Fournier, F. Brulé, M. Chevrier-Miller, and C. Branlant, “Secondary structure of the yeast saccharomyces cerevisiae pre-u3a snorna and its implication for splicing efficiency.,” *RNA*, vol. 2, pp. 1079–1093, 11 1996.

[89] O. Gahura, C. Hammann, A. Valentová, F. Půta, and P. Folk, “Secondary structure is required for 3’ splice site recognition in yeast,” *Nucleic Acids Res*, vol. 39, pp. 9759–67, Dec 2011.

[90] K. J. Howe and M. Ares, Jr, “Intron self-complementarity enforces exon inclusion in a yeast pre-mrna,” *PNAS*, vol. 94, pp. 12467–72, Nov 1997.

[91] R. E. Hector, K. R. Nykamp, S. Dheur, J. T. Anderson, P. J. Non, C. R. Urbinati, S. M. Wilson, L. Minvielle-Sebastia, and M. S. Swanson, “Dual requirement for yeast hnrnp nab2p in mrna poly(a) tail length control and nuclear export,” *EMBO J*, vol. 21, pp. 1800–10, Apr 2002.

[92] T. Preiss, M. Muckenthaler, and M. W. Hentze, “Poly(a)-tail-promoted translation in yeast: implications for translational control,” *RNA*, vol. 4, pp. 1321–31, Nov 1998.

[93] C. E. Brown and A. B. Sachs, “Poly(a) tail length control in saccharomyces cerevisiae occurs by message-specific deadenylation,” *Mol Cell Biol*, vol. 18, pp. 6548–59, Nov 1998.

[94] T. Preiss and M. W. Hentze, “Dual function of the messenger rna cap structure in poly(a)-tail-promoted translation in yeast,” *Nature*, vol. 392, pp. 516–20, Apr 1998.

[95] S. K. Archer, N. E. Shirokikh, T. H. Beilharz, and T. Preiss, “Dynamics of ribosome scanning and recycling revealed by translation complex profiling,” *Nature*, vol. 535, pp. 570–4, Jul 2016.

[96] S. E. O’Leary, A. Petrov, J. Chen, and J. D. Puglisi, “Dynamic recognition of the mrna cap by saccharomyces cerevisiae eif4e,” *Structure*, vol. 21, pp. 2197–207, Dec 2013.

[97] Y. V. Svitkin, A. Pause, A. Haghighat, S. Pyronnet, G. Witherell, G. J. Belsham, and N. Sonenberg, “The requirement for eukaryotic initiation factor 4a (elf4a) in translation is in direct proportion to the degree of mrna 5’ secondary structure,” *RNA*, vol. 7, pp. 382–94, Mar 2001.

[98] S. Rocak and P. Linder, “Dead-box proteins: the driving forces behind rna metabolism,” *Nat Rev Mol Cell Biol*, vol. 5, pp. 232–41, Mar 2004.

[99] U. Harms, A. Z. Andreou, A. Gubaev, and D. Klostermeier, “eif4b, eif4g and rna regulate eif4a activity in translation initiation by modulating the eif4a conformational cycle,” *Nucleic Acids Res*, vol. 42, pp. 7911–22, Jul 2014.

[100] A. Z. Andreou and D. Klostermeier, “eif4b and eif4g jointly stimulate eif4a atpase and unwinding activities by modulation of the eif4a conformational cycle,” *J Mol Biol*, vol. 426, pp. 51–61, Jan 2014.

[101] V. Rajagopal, E.-H. Park, A. G. Hinnebusch, and J. R. Lorsch, “Specific domains in yeast translation initiation factor eif4g strongly bias rna unwinding activity of the eif4f complex toward duplexes with 5’-overhangs,” *J Biol Chem*, vol. 287, pp. 20301–12, Jun 2012.

[102] L. David, W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz, “A high-resolution map of transcription in the yeast genome,” *PNAS*, vol. 103, pp. 5320–5, Apr 2006.

[103] Z. Lin and W.-H. Li, “Evolution of 5’ untranslated region length and gene expression reprogramming in yeasts,” *Mol Biol Evol*, vol. 29, pp. 81–9, Jan 2012.

[104] M. R. Vega Laso, D. Zhu, F. Sagliocco, A. J. Brown, M. F. Tuite, and J. E. McCarthy, “Inhibition of translational initiation in the yeast saccharomyces cerevisiae as a function of the stability and position of hairpin structures in the mrna leader,” *J Biol Chem*, vol. 268, pp. 6453–62, Mar 1993.

[105] A. Kochetov, D. Vorobiev, O. Sirnik, L. Kisselev, and N. Kolchanov, “Contextual features of yeast mrna 5’utrs potentially important for their translational activity,” in *Bioinformatics Of Regulatory Genomic Sequences*, 2000.

[106] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal, “Genome-wide measurement of rna secondary structure in yeast,” *Nature*, vol. 467, pp. 103–7, Sep 2010.

[107] A. E. Koromilas, A. Lazaris-Karatzas, and N. Sonenberg, “mrnas containing extensive secondary structure in their 5’ non-coding region translate efficiently in cells overexpressing initiation factor eif-4e,” *EMBO J*, vol. 11, pp. 4153–8, Nov 1992.

[108] M. Senissar, A. Le Saux, N. Belgareh-Touzé, C. Adam, J. Banroques, and N. K. Tanner, “The dead-box helicase ded1 from yeast is an mrnp cap-associated protein that shuttles between the cytoplasm and nucleus,” *Nucleic Acids Res*, vol. 42, pp. 10005–22, Sep 2014.

[109] A. Hilliker, Z. Gao, E. Jankowsky, and R. Parker, “The dead-box protein ded1 modulates translation by the formation and resolution of an eif4f-mrna complex,” *Molecular Cell*, vol. 43, pp. 962–72, Sep 2011.

[110] N. D. Sen, F. Zhou, N. T. Ingolia, and A. G. Hinnebusch, “Genome-wide analysis of translational efficiency reveals distinct but overlapping functions of yeast dead-box rna helicases ded1 and eif4a,” *Genome Res*, vol. 25, pp. 1196–205, Aug 2015.

[111] A. Robbins-Pianka, M. D. Rice, and M. P. Weir, “The mrna landscape at yeast translation initiation sites,” *Bioinformatics*, vol. 26, pp. 2651–5, Nov 2010.

[112] T. Ben-Yehezkel, H. Zur, T. Marx, E. Shapiro, and T. Tuller, “Mapping the translation initiation landscape of an s. cerevisiae gene using fluorescent proteins,” *Genomics*, vol. 102, pp. 419–29, Oct 2013.

[113] R. Hamilton, C. K. Watanabe, and H. A. de Boer, “Compilation and comparison of the sequence context around

the aug startcodons in saccharomyces cerevisiae mrnas," *Nucleic Acids Research*, vol. 15, pp. 3581–3593, 04 1987.

[114] S. Dvir, L. Velten, E. Sharon, D. Zeevi, L. B. Carey, A. Weinberger, and E. Segal, "Deciphering the rules by which 5'-utr sequences affect protein expression in yeast," *PNAS*, vol. 110, pp. E2792–801, Jul 2013.

[115] M. Liss, D. Daubert, K. Brunner, K. Kliche, U. Hammes, A. Leiherer, and R. Wagner, "Embedding permanent watermarks in synthetic genes," *PLoS One*, vol. 7, no. 8, p. e42465, 2012.

[116] B. K.-S. Chung and D.-Y. Lee, "Computational codon optimization of synthetic gene for protein expression," *BMC Syst Biol*, vol. 6, p. 134, Oct 2012.

[117] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel, "An evolutionarily conserved mechanism for controlling the efficiency of protein translation," *Cell*, vol. 141, pp. 344–54, Apr 2010.

[118] T.-D. M. Plank and J. S. Kieft, "The structures of nonprotein-coding rnas that drive internal ribosome entry site function," *Wiley Interdiscip Rev RNA*, vol. 3, no. 2, pp. 195–212, 2012.

[119] T. Masek, V. Vopalensky, O. Horvath, L. Vortelova, Z. Feketova, and M. Pospisek, "Hepatitis c virus internal ribosome entry site initiates protein synthesis at the authentic initiation codon in yeast," *J Gen Virol*, vol. 88, pp. 1992–2002, Jul 2007.

[120] S. Weingarten-Gabbay, S. Elias-Kirma, R. Nir, A. A. Gritsenko, N. Stern-Ginossar, Z. Yakhini, A. Weinberger, and E. Segal, "Systematic discovery of cap-independent translation sequences in human and viral genomes," *Science Comparative genetics.*, vol. 351, Jan 2016.

[121] R. Shalgi, M. Lapidot, R. Shamir, and Y. Pilpel, "A catalog of stability-associated sequence elements in 3' utrs of yeast mrnas," *Genome Biol*, vol. 6, no. 10, p. R86, 2005.

[122] M. Rabani, M. Kertesz, and E. Segal, "Computational prediction of rna structural motifs involved in posttranscriptional regulatory processes," *PNAS*, vol. 105, pp. 14885–90, Sep 2008.

[123] R. J. Ulbricht and W. M. Olivas, "Puf1p acts in combination with other yeast puf proteins to control mrna stability," *RNA*, vol. 14, pp. 246–62, Feb 2008.

[124] S. E. Wells, P. E. Hillner, R. D. Vale, and A. B. Sachs, "Circularization of mrna by eukaryotic translation initiation factors," *Mol Cell*, vol. 2, pp. 135–40, Jul 1998.

[125] B. Mazumder, V. Seshadri, and P. L. Fox, "Translational control by the 3'-utr: the ends specify the means," *Trends Biochem Sci*, vol. 28, pp. 91–8, Feb 2003.

[126] S. M. Tan-Wong, J. B. Zaugg, J. Camblong, Z. Xu, D. W. Zhang, H. E. Mischo, A. Z. Ansari, N. M. Luscombe, L. M. Steinmetz, and N. J. Proudfoot, "Gene loops enhance transcriptional directionality," *Science*, vol. 338, pp. 671–5, Nov 2012.

[127] L. Weill, E. Belloc, F.-A. Bava, and R. Méndez, "Translational control by changes in poly(a) tail length: recycling mrnas," *Nat Struct Mol Biol*, vol. 19, pp. 577–85, Jun 2012.

[128] S. Meyer, C. Temme, and E. Wahle, "Messenger rna turnover in eukaryotes: Pathways and enzymes," *Crit. Rev. Biochem. Mol. Biol.*, vol. 39, pp. 197–216, Jul-Aug 2004.

[129] W. Hu, T. J. Sweet, S. Chamnongpol, K. E. Baker, and J. Coller, "Co-translational mrna decay in saccharomyces cerevisiae," *Nature*, vol. 461, pp. 225–9, Sep 2009.

[130] R. Parker, "Rna degradation in saccharomyces cerevisae," *Genetics*, vol. 191, pp. 671–702, July 2012.

[131] B. Linz, N. Koloteva, S. Vasilescu, and J. E. McCarthy, "Disruption of ribosomal scanning on the 5'-untranslated region, and not restriction of translational initiation per se, modulates the stability of nonaberrant mrnas in the yeast saccharomyces cerevisiae," *J Biol Chem*, vol. 272, pp. 9131–40, Apr 1997.

[132] D. O. Passos, M. K. Doma, C. J. Shoemaker, D. Muhlrad, R. Green, J. Weissman, J. Hollien, and R. Parker, "Analysis of dom34 and its function in no-go decay," *Mol Biol Cell*, vol. 20, pp. 3025–32, Jul 2009.

[133] T. Becker, J.-P. Armache, A. Jarasch, A. M. Anger, E. Villa, H. Sieber, B. A. Motaal, T. Mielke, O. Berninghausen, and R. Beckmann, "Structure of the no-go mrna decay complex dom34-hbs1 bound to a stalled 80s ribosome," *Nat Struct Mol Biol*, vol. 18, pp. 715–20, Jun 2011.

[134] W. J. Blake, M. KAErn, C. R. Cantor, and J. J. Collins, "Noise in eukaryotic gene expression," *Nature*, vol. 422, pp. 633–7, Apr 2003.

[135] J. M. Raser and E. K. O'Shea, "Control of stochasticity in eukaryotic gene expression," *Science*, vol. 304, pp. 1811–4, Jun 2004.

[136] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden, "Regulation of noise in the expression of a single gene," *Nat Genet*, vol. 31, pp. 69–73, May 2002.

[137] P. S. Swain, M. B. Elowitz, and E. D. Siggia, "Intrinsic and extrinsic contributions to stochasticity in gene expression," *PNAS*, vol. 99, pp. 12795–800, Oct 2002.

[138] A. M. Kierzek, J. Zaim, and P. Zielenkiewicz, "The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression," *J Biol Chem*, vol. 276, pp. 8165–72, Mar 2001.

[139] E. Dacheux, H. Firczuk, and J. E. G. McCarthy, "Rate control in yeast protein synthesis at the population and single-cell levels," *Biochem Soc Trans*, vol. 43, pp. 1266–70, Dec 2015.

[140] L. Potvin-Trottier, N. D. Lord, G. Vinnicombe, and J. Paulsson, "Synchronous long-term oscillations in a synthetic gene circuit," *Nature*, Oct 2016.

[141] R. Kwok, "Five hard truths for synthetic biology," *Nature*, vol. 463, pp. 288–90, Jan 2010.

[142] O. Borkowski, F. Ceroni, G.-B. Stan, and T. Ellis, "Overloaded and stressed: whole-cell considerations for bacterial synthetic biology," *Curr Opin Microbiol*, vol. 33, pp. 123–130, Oct 2016.

[143] M. B. Kopniczky, S. J. Moore, and P. S. Freemont, "Multilevel regulation and translational switches in synthetic

biology," *IEEE Trans Biomed Circuits Syst*, vol. 9, pp. 485–96, Aug 2015.

[144] M. L. Woods, M. Leon, R. Perez-Carrasco, and C. P. Barnes, "A statistical approach reveals designs for the most robust stochastic gene oscillators," *ACS Synth Biol*, Feb 2016.

[145] S.-H. Park, A. Zarrinpar, and W. A. Lim, "Rewiring map kinase pathways using alternative scaffold assembly mechanisms," *Science*, vol. 299, pp. 1061–4, Feb 2003.

[146] R. S. Sikorski and P. Hieter, "A system of shuttle vectors and yeast host strains designed for efficient manipulation of dna in saccharomyces cerevisiae," *Genetics*, vol. 122, pp. 19–27, May 1989.

[147] T. Durfee, R. Nelson, S. Baldwin, G. Plunkett, 3rd, V. Burland, B. Mau, J. F. Petrosino, X. Qin, D. M. Muzny, M. Ayele, R. A. Gibbs, B. Csörgo, G. Pósfai, G. M. Weinstock, and F. R. Blattner, "The complete genome sequence of escherichia coli dh10b: insights into the biology of a laboratory workhorse," *J Bacteriol*, vol. 190, pp. 2597–606, Apr 2008.

[148] D. C. Amberg, J. N. Strathern, and D. J. Burke, *Methods in yeast genetics*. Cold Spring Harbor Laboratory Press, 2005.

[149] R. D. Gietz and R. H. Schiestl, "Frozen competent yeast cells that can be transformed with high efficiency using the liac/ss carrier dna/peg method," *Nature Protocols*, vol. 2, no. 1, pp. 1–4, 2007.

[150] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, pp. 764–766, 7 2013.

[151] K. K. YUEN, "The two-sample trimmed t for unequal population variances," *Biometrika*, vol. 61, pp. 165–170, 04 1974.

[152] R. R. Wilcox, "Rallfun statistics package."

[153] R. R. Wilcox and T. S. Tian, "Measuring effect size: a robust heteroscedastic approach for two or more groups," *Journal of Applied Statistics*, vol. 38, pp. 1359–1368, 07 2011.

[154] R. Wilcox, *Chapter 5 - Comparing Two Groups*, pp. 137–213. Boston: Academic Press, 2012.

[155] `http://openwetware.org/wiki/Qiagen_Buffers`, "Open wetware generic miniprep buffer recipes."

[156] C. B. Brachmann, A. Davies, G. J. Cost, E. Caputo, J. Li, P. Hieter, and J. D. Boeke, "Designer deletion strains derived from saccharomyces cerevisiae s288c: a useful set of strains and plasmids for pcr-mediated gene disruption and other applications," *Yeast*, vol. 14, pp. 115–32, Jan 1998.

[157] `www.neb.com/nebecomm/ManualFiles/manualF-530.pdf`, "Neb phusion high-fidelity dna polymerase manual."

[158] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative pcr and the 2(-delta delta c(t)) method," *Methods*, vol. 25, pp. 402–8, Dec 2001.

[159] H. M. Salis, E. A. Mirsky, and C. A. Voigt, "Automated design of synthetic ribosome binding sites to control protein expression," *Nat Biotechnol*, vol. 27, pp. 946–50, Oct 2009.

[160] B. Reeve, T. Hargest, C. Gilbert, and T. Ellis, "Predicting translation initiation rates for designing synthetic biology," *Front Bioeng Biotechnol*, vol. 2, p. 1, 2014.

[161] S. R. Eddy, "How do rna folding algorithms work?," *Nat Biotechnol*, vol. 22, pp. 1457–8, Nov 2004.

[162] I. Aviram, I. Veltman, A. Churkin, and D. Barash, "Efficient procedures for the numerical simulation of mid-size rna kinetics," *Algorithms Mol Biol*, vol. 7, no. 1, p. 24, 2012.

[163] M. Mann, M. Kucharík, C. Flamm, and M. T. Wolfinger, "Memory-efficient rna energy landscape exploration," *Bioinformatics*, vol. 30, pp. 2584–91, Sep 2014.

[164] A. Krokhotin and N. V. Dokholyan, "Computational methods toward accurate rna structure prediction using coarse-grained and all-atom models," *Methods Enzymol*, vol. 553, pp. 65–89, 2015.

[165] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Res*, vol. 31, pp. 3406–15, Jul 2003.

[166] N. R. Markham and M. Zuker, "Unafold: software for nucleic acid folding and hybridization," *Methods Mol Biol*, vol. 453, pp. 3–31, 2008.

[167] J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, and N. A. Pierce, "Nupack: Analysis and design of nucleic acid systems," *J Comput Chem*, vol. 32, pp. 170–3, Jan 2011.

[168] Z. Z. Xu and D. H. Mathews, "Experiment-assisted secondary structure prediction with rnastructure," *Methods Mol Biol*, vol. 1490, pp. 163–76, 2016.

[169] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "Viennarna package 2.0," *Algorithms Mol Biol*, vol. 6, p. 26, 2011.

[170] A. R. Gruber, S. H. Bernhart, and R. Lorenz, "The viennarna web services," *Methods Mol Biol*, vol. 1269, pp. 307–26, 2015.

[171] J. Wolters, "The nature of preferred hairpin structures in 16s-like rrna variable regions," *Nucleic Acids Res*, vol. 20, pp. 1843–50, Apr 1992.

[172] C. Tuerk, P. Gauss, C. Thermes, D. R. Groebe, M. Gayle, N. Guild, G. Stormo, Y. d'Aubenton Carafa, O. C. Uhlenbeck, and I. Tinoco, Jr, "Cuucgg hairpins: extraordinarily stable rna secondary structures associated with various biochemical processes," *PNAS*, vol. 85, pp. 1364–8, Mar 1988.

[173] I. Tinoco, Jr and C. Bustamante, "How rna folds," *J Mol Biol*, vol. 293, pp. 271–81, Oct 1999.

[174] V. P. Antao, S. Y. Lai, and I. Tinoco, Jr, "A thermodynamic study of unusually stable rna and dna hairpins," *Nucleic Acids Res*, vol. 19, pp. 5901–5, Nov 1991.

[175] J. P. Sheehy, A. R. Davis, and B. M. Znosko, "Thermodynamic characterization of naturally occurring rna tetraloops," *RNA*, vol. 16, pp. 417–29, Feb 2010.

[176] P. L. Vanegas, T. S. Horwitz, and B. M. Znosko, "Effects of non-nearest neighbors on the thermodynamic

stability of rna gnra hairpin tetraloops," *Biochemistry*, vol. 51, pp. 2192–8, Mar 2012.

[177] D. Chakraborty, R. Collepardo-Guevara, and D. J. Wales, "Energy landscapes, folding mechanisms, and kinetics of rna tetraloop hairpins," *J Am Chem Soc*, vol. 136, pp. 18052–61, Dec 2014.

[178] M. Molinaro and I. Tinoco, Jr, "Use of ultra stable uncg tetraloop hairpins to fold rna structures: thermodynamic and spectroscopic applications," *Nucleic Acids Res*, vol. 23, pp. 3056–63, Aug 1995.

[179] C. C. Correll and K. Swinger, "Common and distinctive features of gnra tetraloops based on a guaa tetraloop structure at 1.4 a resolution," *RNA*, vol. 9, pp. 355–63, Mar 2003.

[180] F. M. Jucker and A. Pardi, "Solution structure of the cuug hairpin loop: a novel rna tetraloop motif," *Biochemistry*, vol. 34, pp. 14416–27, Nov 1995.

[181] Q. Zhao, H.-C. Huang, U. Nagaswamy, Y. Xia, X. Gao, and G. E. Fox, "Unac tetraloops: to what extent do they mimic gnra tetraloops?," *Biopolymers*, vol. 97, pp. 617–28, Aug 2012.

[182] H.-C. Huang, U. Nagaswamy, and G. E. Fox, "The application of cluster analysis in the intercomparison of loop structures in rna," *RNA*, vol. 11, pp. 412–23, Apr 2005.

[183] C. R. Woese, S. Winker, and R. R. Gutell, "Architecture of ribosomal rna: constraints on the sequence of "tetra-loops"," *PNAS*, vol. 87, pp. 8467–71, Nov 1990.

[184] P. Kührová, P. Banáš, R. B. Best, J. Šponer, and M. Otyepka, "Computer folding of rna tetraloops? are we there yet?," *J Chem Theory Comput*, vol. 9, pp. 2115–25, Apr 2013.

[185] V. P. Antao and I. Tinoco, Jr, "Thermodynamic parameters for loop formation in rna and dna hairpin tetraloops," *Nucleic Acids Res*, vol. 20, pp. 819–24, Feb 1992.

[186] H.-K. Cheong, N.-K. Kim, and C. Cheong, *RNA Structure: Tetraloops*. John Wiley & Sons, Ltd, 2001.

[187] L. Jaeger, F. Michel, and E. Westhof, "Involvement of a gnra tetraloop in long-range rna tertiary interactions," *J Mol Biol*, vol. 236, pp. 1271–6, Mar 1994.

[188] J. H. Cate, A. R. Gooding, E. Podell, K. Zhou, B. L. Golden, C. E. Kundrot, T. R. Cech, and J. A. Doudna, "Crystal structure of a group i ribozyme domain: principles of rna packing," *Science*, vol. 273, pp. 1678–85, Sep 1996.

[189] R. Thapar, A. P. Denmon, and E. P. Nikonowicz, "Recognition modes of rna tetraloops and tetraloop-like motifs by rna-binding proteins," *Wiley Interdiscip Rev RNA*, vol. 5, no. 1, pp. 49–67, 2014.

[190] S. Konermann, M. D. Brigham, A. E. Trevino, J. Joung, O. O. Abudayyeh, C. Barcena, P. D. Hsu, N. Habib, J. S. Gootenberg, H. Nishimasu, O. Nureki, and F. Zhang, "Genome-scale transcriptional activation by an engineered crispr-cas9 complex," *Nature*, vol. 517, pp. 583–8, Jan 2015.

[191] J. G. Zalatan, M. E. Lee, R. Almeida, L. A. Gilbert, E. H. Whitehead, M. La Russa, J. C. Tsai, J. S. Weissman, J. E. Dueber, L. S. Qi, and W. A. Lim, "Engineering complex synthetic transcriptional programs with crispr rna scaffolds," *Cell*, vol. 160, pp. 339–50, Jan 2015.

[192] S. Hocine, P. Raymond, D. Zenklusen, J. A. Chao, and R. H. Singer, "Single-molecule analysis of gene expression using two-color rna labeling in live yeast," *Nat Methods*, vol. 10, pp. 119–21, Feb 2013.

[193] S. B. Baim and F. Sherman, "mrna structures influencing translation in the yeast saccharomyces cerevisiae," *Mol Cell Biol*, vol. 8, pp. 1591–601, Apr 1988.

[194] F. A. Sagliocco, M. R. Vega Laso, D. Zhu, M. F. Tuite, J. E. McCarthy, and A. J. Brown, "The influence of 5'-secondary structures upon ribosome binding to mrna during translation in yeast," *J Biol Chem*, vol. 268, pp. 26522–30, Dec 1993.

[195] M. Ringnér and M. Krogh, "Folding free energies of 5'-utrs impact post-transcriptional regulation on a genomic scale in yeast," *PLoS Comput Biol*, vol. 1, p. e72, Dec 2005.

[196] J. R. Babendure, J. L. Babendure, J.-H. Ding, and R. Y. Tsien, "Control of mammalian translation by mrna structure near caps," *RNA*, vol. 12, pp. 851–61, May 2006.

[197] K. Endo, J. A. Stapleton, K. Hayashi, H. Saito, and T. Inoue, "Quantitative and simultaneous translational control of distinct mammalian mrnas," *Nucleic Acids Res*, vol. 41, p. e135, Jul 2013.

[198] R. E. Cerny, Y. Qi, C. M. Aydt, S. Huang, J. J. Listello, B. J. Fabbri, T. W. Conner, L. Crossland, and J. Huang, "Rna-binding protein-mediated translational repression of transgene expression in plants," *Plant Mol Biol*, vol. 52, pp. 357–69, May 2003.

[199] E. Lamping, M. Niimi, and R. D. Cannon, "Small, synthetic, gc-rich mrna stem-loop modules 5' proximal to the aug start-codon predictably tune gene expression in yeast," *Microb Cell Fact*, vol. 12, p. 74, Jul 2013.

[200] A. V. Anzalone, A. J. Lin, S. Zairis, R. Rabadan, and V. W. Cornish, "Reprogramming eukaryotic translation with ligand-responsive synthetic rna switches," *Nat Methods*, vol. 13, pp. 453–8, May 2016.

[201] E. Paraskeva, A. Atzberger, and M. W. Hentze, "A translational repression assay procedure (trap) for rna-protein interactions in vivo," *Proc Natl Acad Sci U S A*, vol. 95, pp. 951–6, Feb 1998.

[202] M. Nie and H. Htun, "Different modes and potencies of translational repression by sequence-specific rna-protein interaction at the 5'-utr," *Nucleic Acids Res*, vol. 34, no. 19, pp. 5528–40, 2006.

[203] P. Kötter, J. E. Weigand, B. Meyer, K.-D. Entian, and B. Suess, "A fast and efficient translational control system for conditional expression of yeast genes," *Nucleic Acids Res*, vol. 37, p. e120, Oct 2009.

[204] C. Hsu, V. Jaquet, M. Gencoglu, and A. Becskei, "Protein dimerization generates bistability in positive feedback loops," *Cell Rep*, vol. 16, pp. 1204–10, Aug 2016.

[205] A. H. Babiskin and C. D. Smolke, "A synthetic library of rna control modules for predictable tuning of gene expression in yeast," *Mol Syst Biol*, vol. 7, p. 471, Mar 2011.

[206] R. D. Gietz and R. H. Schiestl, "High-efficiency yeast transformation using the liac/ss carrier dna/peg method," *Nature Protocols*, vol. 2, no. 1, pp. 31–34, 2007.

[207] S. M. Castillo-Hair, J. T. Sexton, B. P. Landry, E. J. Olson, O. A. Igoshin, and J. J. Tabor, “Flowcal: A user-friendly, open source software tool for automatically converting flow cytometry data from arbitrary to calibrated units,” *ACS Synth Biol*, vol. 5, pp. 774–80, Jul 2016.

[208] R. M. Dirks, M. Lin, E. Winfree, and N. A. Pierce, “Paradigms for computational nucleic acid design,” *Nucleic Acids Res*, vol. 32, no. 4, pp. 1392–403, 2004.

[209] J. A. Garcia-Martin, P. Clote, and I. Dotu, “Rnaifold: a web server for rna inverse folding and molecular design,” *Nucleic Acids Res*, vol. 41, pp. W465–70, Jul 2013.

[210] S. Lee, W. A. Lim, and K. S. Thorn, “Improved blue, green, and red fluorescent protein tagging vectors for s. cerevisiae,” *PLoS One*, vol. 8, no. 7, p. e67902, 2013.

[211] A. Cankorur-Cetinkaya, E. Dereli, S. Eraslan, E. Karabekmez, D. Dikicioglu, and B. Kirdar, “A novel strategy for selection and validation of reference genes in dynamic multidimensional experimental design in yeast,” *PLoS One*, vol. 7, no. 6, p. e38351, 2012.

[212] D. Zenklusen, D. R. Larson, and R. H. Singer, “Single-rna counting reveals alternative modes of gene expression in yeast,” *Nat Struct Mol Biol*, vol. 15, pp. 1263–71, Dec 2008.

[213] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, “Blast+: architecture and applications,” *BMC Bioinformatics*, vol. 10, p. 421, 2009.

[214] T. Weenink and T. Ellis, “Creation and characterization of component libraries for synthetic biology,” *Methods Mol Biol*, vol. 1073, pp. 51–60, 2013.

[215] N. V. Bhagavan, *Medical Biochemistry, page 70.* Harcourt/Academic Press, 2002.

[216] M. K. Doma and R. Parker, “Endonucleolytic cleavage of eukaryotic mrnas with stalls in translation elongation,” *Nature*, vol. 440, pp. 561–4, Mar 2006.

[217] A. J. Ninfa and A. E. Mayo, “Hysteresis vs. graded responses: the connections make all the difference,” *Sci STKE*, vol. 2004, p. pe20, May 2004.

[218] J. E. Pérez-Ortín, P. M. Alepuz, and J. Moreno, “Genomics and gene transcription kinetics in yeast,” *Trends Genet*, vol. 23, pp. 250–7, May 2007.

[219] A. C. Palmer, A. Ahlgren-Berg, J. B. Egan, I. B. Dodd, and K. E. Shearwin, “Potent transcriptional interference by pausing of rna polymerases over a downstream promoter,” *Mol Cell*, vol. 34, pp. 545–55, Jun 2009.

[220] S. Hahn, “Structure and mechanism of the rna polymerase ii transcription machinery,” *Nat Struct Mol Biol*, vol. 11, pp. 394–403, May 2004.

[221] B. P. Callen, K. E. Shearwin, and J. B. Egan, “Transcriptional interference between convergent promoters caused by elongation over the promoter,” *Mol Cell*, vol. 14, pp. 647–56, Jun 2004.

[222] E. M. Prescott and N. J. Proudfoot, “Transcriptional collision between convergent genes in budding yeast,” *PNAS*, vol. 99, pp. 8796–801, Jun 2002.

[223] K. Sneppen, I. B. Dodd, K. E. Shearwin, A. C. Palmer, R. A. Schubert, B. P. Callen, and J. B. Egan, “A mathematical model for transcriptional interference by rna polymerase traffic in escherichia coli,” *J Mol Biol*, vol. 346, pp. 399–409, Feb 2005.

[224] A. C. Palmer, J. B. Egan, and K. E. Shearwin, “Transcriptional interference by rna polymerase pausing and dislodgement of transcription factors,” *Transcription*, vol. 2, no. 1, pp. 9–14, 2011.

[225] I. B. Dodd, B. Kalionis, and J. B. Egan, “Control of gene expression in the temperate coliphage 186. viii. control of lysis and lysogeny by a transcriptional switch involving face-to-face promoters,” *J Mol Biol*, vol. 214, pp. 27–37, Jul 1990.

[226] A. Chatterjee, C. M. Johnson, C.-C. Shu, Y. N. Kaznessis, D. Ramkrishna, G. M. Dunny, and W.-S. Hu, “Convergent transcription confers a bistable switch in enterococcus faecalis conjugation,” *PNAS*, vol. 108, pp. 9721–6, Jun 2011.

[227] A. E. Bordoy, U. S. Varanasi, C. M. Courtney, and A. Chatterjee, “Transcriptional interference in convergent promoters as a means for tunable gene expression,” *ACS Synth Biol*, Jul 2016.

[228] S. Puig, J. E. Pérez-Ortín, and E. Matallana, “Transcriptional and structural study of a region of two convergent overlapping yeast genes,” *Curr Microbiol*, vol. 39, pp. 369–0373, Dec 1999.

[229] L. Wang, N. Jiang, L. Wang, O. Fang, L. J. Leach, X. Hu, and Z. Luo, “3’ untranslated regions mediate transcriptional interference between convergent genes both locally and ectopically in saccharomyces cerevisiae,” *PLoS Genet*, vol. 10, p. e1004021, Jan 2014.

[230] S. A. Hoffmann, S. M. Kruse, and K. M. Arndt, “Long-range transcriptional interference in e. coli used to construct a dual positive selection system for genetic switches,” *Nucleic Acids Res*, vol. 44, p. e95, Jun 2016.

[231] J. A. Brophy and C. A. Voigt, “Antisense transcription as a tool to tune gene expression,” *Mol Syst Biol*, vol. 12, no. 1, p. 854, 2016.

[232] A. Korde, J. M. Rosselot, and D. Donze, “Intergenic transcriptional interference is blocked by rna polymerase iii transcription factor tfiiib in saccharomyces cerevisiae,” *Genetics*, vol. 196, pp. 427–38, Feb 2014.

[233] S. Irniger, C. M. Egli, M. Kuenzler, and G. H. Braus, “The yeast actin intron contains a cryptic promoter that can be switched on by preventing transcriptional interference,” *Nucleic Acids Res*, vol. 20, pp. 4733–9, Sep 1992.

[234] S. C. Murray, A. Serra Barros, D. A. Brown, P. Dudek, J. Ayling, and J. Mellor, “A pre-initiation complex at the 3’-end of genes drives antisense transcription independent of divergent sense transcription,” *Nucleic Acids Res*, vol. 40, pp. 2432–44, Mar 2012.

[235] J. Houseley, L. Rubbi, M. Grunstein, D. Tollervey, and M. Vogelauer, “A ncrna modulates histone modification and mrna induction in the yeast gal gene cluster,” *Mol Cell*, vol. 32, pp. 685–95, Dec 2008.

[236] M. Pinskaya, S. Gourvennec, and A. Morillon, “H3 lysine 4 di- and tri-methylation deposited by cryptic

transcription attenuates promoter activation," *EMBO J*, vol. 28, pp. 1697–707, Jun 2009.

[237] C. F. Hongay, P. L. Grisafi, T. Galitski, and G. R. Fink, "Antisense transcription controls cell fate in saccharomyces cerevisiae," *Cell*, vol. 127, pp. 735–45, Nov 2006.

[238] T. Nguyen, H. Fischl, F. S. Howe, R. Woloszczuk, A. Serra Barros, Z. Xu, D. Brown, S. C. Murray, S. Haenni, J. M. Halstead, L. O'Connor, G. Shipkovenska, L. M. Steinmetz, and J. Mellor, "Transcription mediated insulation and interference direct gene cluster expression switches," *Elife*, vol. 3, p. e03635, Nov 2014.

[239] R. Barrangou, "Diversity of crispr-cas immune systems and molecular machines," *Genome Biol*, vol. 16, p. 247, Nov 2015.

[240] A. Agrotis and R. Ketteler, "A new age in functional genomics using crispr/cas9 in arrayed library screening," *Front Genet*, vol. 6, p. 300, 2015.

[241] http://www.ersgenomics.com/crispr-cas-9-technology.php, "Crispr cas9 gene regulation."

[242] A. Didovyk, B. Borek, L. Tsimring, and J. Hasty, "Transcriptional regulation with crispr-cas9: principles, advances, and applications," *Curr Opin Biotechnol*, vol. 40, pp. 177–84, Aug 2016.

[243] C. M. Ajo-Franklin, D. A. Drubin, J. A. Eskin, E. P. S. Gee, D. Landgraf, I. Phillips, and P. A. Silver, "Rational design of memory in eukaryotic cells," *Genes Dev*, vol. 21, pp. 2271–6, Sep 2007.

[244] D. Siegal-Gaskins, M. K. Mejia-Guerra, G. D. Smith, and E. Grotewold, "Emergence of switch-like behavior in a large family of simple biochemical networks," *PLoS Comput Biol*, vol. 7, p. e1002039, May 2011.

[245] F. He and A. Jacobson, "Nonsense-mediated mrna decay: Degradation of defective transcripts is only part of the story," *Annu Rev Genet*, vol. 49, pp. 339–66, 2015.

[246] E. Kristiansson, M. Thorsen, M. J. Tamás, and O. Nerman, "Evolutionary forces act on promoter length: identification of enriched cis-regulatory elements," *Mol Biol Evol*, vol. 26, pp. 1299–307, Jun 2009.

[247] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The transcriptional landscape of the yeast genome defined by rna sequencing," *Science*, vol. 320, pp. 1344–9, Jun 2008.

[248] W. Zhou, G. M. Edelman, and V. P. Mauro, "Isolation and identification of short nucleotide sequences that affect translation initiation in saccharomyces cerevisiae," *PNAS*, vol. 100, pp. 4457–62, Apr 2003.

[249] W. V. Gilbert, K. Zhou, T. K. Butler, and J. A. Doudna, "Cap-independent translation is required for starvation-induced differentiation in yeast," *Science*, vol. 317, pp. 1224–7, Aug 2007.

[250] L. C. Reineke and W. C. Merrick, "Characterization of the functional role of nucleotides within the ure2 ires element and the requirements for eif2a-mediated repression," *RNA*, vol. 15, pp. 2264–77, Dec 2009.

[251] A. A. Komar, B. Mazumder, and W. C. Merrick, "A new framework for understanding ires-mediated translation," *Gene*, vol. 502, pp. 75–86, Jul 2012.

[252] W. V. Gilbert, "Alternative ways to think about cellular internal ribosome entry," *J Biol Chem*, vol. 285, pp. 29033–8, Sep 2010.

[253] M. Kozak, "A second look at cellular mrna sequences said to function as internal ribosome entry sites," *Nucleic Acids Res*, vol. 33, no. 20, pp. 6593–602, 2005.

[254] F. Miura, N. Kawaguchi, J. Sese, A. Toyoda, M. Hattori, S. Morishita, and T. Ito, "A large-scale full-length cdna analysis to explore the budding yeast transcriptome," *PNAS*, vol. 103, pp. 17846–51, Nov 2006.

[255] J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, K. Karra, C. J. Krieger, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, M. Simison, S. Weng, and E. D. Wong, "Saccharomyces genome database: the genomics resource of budding yeast," *Nucleic Acids Res*, vol. 40, pp. D700–5, Jan 2012.

[256] S. Kiani, J. Beal, M. R. Ebrahimkhani, J. Huh, R. N. Hall, Z. Xie, Y. Li, and R. Weiss, "Crispr transcriptional repression devices and layered circuits in mammalian cells," *Nat Methods*, vol. 11, pp. 723–6, Jul 2014.

[257] A. D. Riggs, H. Suzuki, and S. Bourgeois, "Lac repressor-operator interaction. i. equilibrium studies," *J Mol Biol*, vol. 48, pp. 67–83, Feb 1970.

[258] M. Lewis, "The lac repressor.," *C R Biol*, vol. 328, pp. 521–548, Jun 2005.

[259] C. M. Falcon and K. S. Matthews, "Operator dna sequence variation enhances high affinity binding by hinge helix mutants of lactose repressor protein," *Biochemistry*, vol. 39, pp. 11074–83, Sep 2000.

[260] A. Kamionka, J. Bogdanska-Urbaniak, O. Scholz, and W. Hillen, "Two mutations in the tetracycline repressor change the inducer anhydrotetracycline to a corepressor," *Nucleic Acids Res*, vol. 32, no. 2, pp. 842–7, 2004.

[261] C. Berens, D. Schnappinger, and W. Hillen, "The role of the variable region in tet repressor for inducibility by tetracycline," *J Biol Chem*, vol. 272, pp. 6936–42, Mar 1997.

[262] M. W. Gander, J. D. Vrana, W. E. Voje, J. M. Carothers, and E. Klavins, "Robust digital logic circuits in eukaryotic cells with crispr/dcas9 nor gates," *bioRxiv*, 03 2016.

[263] B. Pfeuty and K. Kaneko, "The combination of positive and negative feedback loops confers exquisite flexibility to biochemical switches," *Phys Biol*, vol. 6, p. 046013, Nov 2009.

[264] S. Kurtz and D. Shore, "Rap1 protein activates and silences transcription of mating-type genes in yeast," *Genes Dev*, vol. 5, pp. 616–28, Apr 1991.

[265] A. Goppelt, G. Stelzer, F. Lottspeich, and M. Meisterernst, "A mechanism for repression of class ii gene transcription through specific binding of nc2 to tbp-promoter complexes via heterodimeric histone fold domains," *EMBO J*, vol. 15, pp. 3105–16, Jun 1996.

[266] J. V. Geisberg, Z. Moqtaderi, L. Kuras, and K. Struhl, "Mot1 associates with transcriptionally active promoters and inhibits association of nc2 in saccharomyces cerevisiae," *Mol Cell Biol*, vol. 22, pp. 8122–34, Dec 2002.

[267] D. T. Auble, K. E. Hansen, C. G. Mueller, W. S. Lane, J. Thorner, and S. Hahn, "Mot1, a global repressor of rna polymerase ii transcription, inhibits tbp binding to dna by an atp-dependent mechanism," *Genes Dev*, vol. 8,

pp. 1920–34, Aug 1994.

[268] A. G. Frey, A. J. Bird, M. V. Evans-Galea, E. Blankman, D. R. Winge, and D. J. Eide, "Zinc-regulated dna binding of the yeast zap1 zinc-responsive activator," *PLoS One*, vol. 6, no. 7, p. e22535, 2011.

[269] M. C. Teixeira, P. Monteiro, P. Jain, S. Tenreiro, A. R. Fernandes, N. P. Mira, M. Alenquer, A. T. Freitas, A. L. Oliveira, and I. Sá-Correia, "The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae," *Nucleic Acids Res*, vol. 34, pp. D446–51, Jan 2006.

[270] T. Lebar, U. Bezeljak, A. Golob, M. Jerala, L. Kadunc, B. Pirš, M. Stražar, D. Vučko, U. Zupančič, M. Benčina, V. Forstnerič, R. Gaber, J. Lonzarić, A. Majerle, A. Oblak, A. Smole, and R. Jerala, "A bistable genetic switch based on designable dna-binding domains," *Nat Commun*, vol. 5, p. 5007, Sep 2014.

[271] O. de Lange, A. Binder, and T. Lahaye, "From dead leaf, to new life: Tal effectors as tools for synthetic biology," *Plant J*, vol. 78, pp. 753–71, Jun 2014.

[272] R. Moore, A. Chandrahas, and L. Bleris, "Transcription activator-like effectors: A toolkit for synthetic biology," *ACS Synth Biol*, vol. 3, pp. 708–16, Oct 2014.

[273] S. Fields and O. Song, "A novel genetic system to detect protein-protein interactions," *Nature*, vol. 340, pp. 245–6, Jul 1989.

[274] F. Farzadfard, S. D. Perli, and T. K. Lu, "Tunable and multifunctional eukaryotic transcription factors based on crispr/cas," *ACS Synth Biol*, vol. 2, pp. 604–13, Oct 2013.

[275] J. D. Smith, S. Suresh, U. Schlecht, M. Wu, O. Wagih, G. Peltz, R. W. Davis, L. M. Steinmetz, L. Parts, and R. P. St Onge, "Quantitative crispr interference screens in yeast identify chemical-genetic interactions and new rules for guide rna design," *Genome Biol*, vol. 17, p. 45, Mar 2016.

[276] O. Purcell, J. Peccoud, and T. K. Lu, "Rule-based design of synthetic transcription factors in eukaryotes," *ACS Synth Biol*, vol. 3, pp. 737–44, Oct 2014.

[277] X. Chen, J. L. Zaro, and W.-C. Shen, "Fusion protein linkers: property, design and functionality," *Adv Drug Deliv Rev*, vol. 65, pp. 1357–69, Oct 2013.

[278] J. Renkawitz, C. A. Lademann, and S. Jentsch, "Mechanisms and principles of homology search during recombination," *Nat Rev Mol Cell Biol*, vol. 15, pp. 369–83, Jun 2014.

[279] H. S. Rhee and B. F. Pugh, "Genome-wide structure and organization of eukaryotic pre-initiation complexes," *Nature*, vol. 483, pp. 295–301, Mar 2012.

[280] M. Yassour, T. Kaplan, H. B. Fraser, J. Z. Levin, J. Pfiffner, X. Adiconis, G. Schroth, S. Luo, I. Khrebtukova, A. Gnirke, C. Nusbaum, D.-A. Thompson, N. Friedman, and A. Regev, "Ab initio construction of a eukaryotic transcriptome by massively parallel mrna sequencing," *PNAS*, vol. 106, pp. 3264–9, Mar 2009.

[281] Q. Zheng, "A new practical guide to the luria-delbrück protocol," *Mutat Res*, vol. 781, pp. 7–13, Nov 2015.

[282] W. A. Rosche and P. L. Foster, "Determining mutation rates in bacterial populations," *Methods*, vol. 20, pp. 4–17, Jan 2000.

[283] A. L. Koch, "Mutation and growth rates from luria-delbrück fluctuation tests," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 95, pp. 129–143, 8 1982.

[284] B. Mandelbrot, "A population birth-and-mutation process, i: explicit distributions for the number of mutants in an old culture of bacteria," vol. 11, no. 3, pp. 437–444, 1974.

[285] H. Rudolph and A. Hinnen, "The yeast pho5 promoter: phosphate-control elements and sequences mediating mrna start-site selection," *PNAS*, vol. 84, pp. 1340–4, Mar 1987.

[286] M. Thattai and A. van Oudenaarden, "Intrinsic noise in gene regulatory networks," *PNAS*, vol. 98, pp. 8614–9, Jul 2001.

[287] N. R. Zabet and D. F. Chu, "Stochasticity and robustness in bi-stable systems," *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*, pp. 1–4, Jun 2010.

[288] M. Krishnamurthy, S. P. Hennelly, T. Dale, S. R. Starkenburg, R. Martí-Arbona, D. T. Fox, S. N. Twary, K. Y. Sanbonmatsu, and C. J. Unkefer, "Tunable riboregulator switches for post-transcriptional control of gene expression," *ACS Synth Biol*, vol. 4, pp. 1326–34, Dec 2015.

[289] A. Espah Borujeni, A. S. Channarasappa, and H. M. Salis, "Translation rate is controlled by coupled trade-offs between site accessibility, selective rna unfolding and sliding at upstream standby sites," *Nucleic Acids Res*, vol. 42, pp. 2646–59, Feb 2014.

[290] F. Mignone, C. Gissi, S. Liuni, and G. Pesole, "Untranslated regions of mrnas," *Genome Biol*, vol. 3, no. 3, p. REVIEWS0004, 2002.

[291] A. Espah Borujeni and H. M. Salis, "Translation initiation is controlled by rna folding kinetics via a ribosome drafting mechanism," *J Am Chem Soc*, vol. 138, pp. 7016–23, Jun 2016.

[292] A. Xayaphoummine, T. Bucher, and H. Isambert, "Kinefold web server for rna/dna folding path and structure prediction including pseudoknots and knots," *Nucleic Acids Res*, vol. 33, pp. W605–10, Jul 2005.

[293] E. Senter and P. Clote, "Fast, approximate kinetics of rna folding," *J Comput Biol*, vol. 22, pp. 124–44, Feb 2015.

[294] F. A. Ran, L. Cong, W. X. Yan, D. A. Scott, J. S. Gootenberg, A. J. Kriz, B. Zetsche, O. Shalem, X. Wu, K. S. Makarova, E. V. Koonin, P. A. Sharp, and F. Zhang, "In vivo genome editing using staphylococcus aureus cas9," *Nature*, vol. 520, pp. 186–91, Apr 2015.

[295] B. P. Kleinstiver, M. S. Prew, S. Q. Tsai, V. V. Topkar, N. T. Nguyen, Z. Zheng, A. P. W. Gonzales, Z. Li, R. T. Peterson, J.-R. J. Yeh, M. J. Aryee, and J. K. Joung, "Engineered crispr-cas9 nucleases with altered pam specificities," *Nature*, vol. 523, pp. 481–5, Jul 2015.