Expanding the Design Space of Synthetic Promoters for CHO Cell Engineering

Adrian Bourke

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Department of Chemical and Biological Engineering
University of Sheffield
Sponsored by Merck Serono

September 2022

Declaration

I, Adrian Bourke declare that this work is entirely my own and have knowledge of the University's Guidance on the Use of Unfair Means. This work has not been submitted for any other award or personal qualification.

Dedication

I would like to dedicate this thesis to my loving partner Frances Nolan for her support and my mom for her encouragement to pursue a PhD.

Finally, I would like to dedicate this thesis in loving memory to my father, Austin Bourke (1952-2020), who would have loved to have seen this day and who was more passionate about academia than I.

Acknowledgements

Throughout the four years of this PhD there has been a lot of hard work, late nights and failed experiments. But without the help of people in the lab and outside, none of this would have been possible.

Firstly, I would like to thank Prof. David James for allowing me to do this PhD. I would also like to thank Julien Douet, Katarzyna Sobkowiak, Julie Frentzel and Mylene Talabardon for supporting me throughout this project and helping me plan and execute the RNA-seq experiment. I would also like to thank you for onboarding and testing my final constructs at Merck on my behalf.

To all the C14 crowd, Fergal, Alex, Pam, Molly and Thilo, I would like to say thanks for creating such a friendly atmosphere to do this PhD and you all made it a much more enjoyable experience, through both the highs and lows.

Thanks to all the postdocs who, without their guidance, I would probably have no data at this point in my PhD. Yusuf, Stephen, Yomi, Renata, Theo, Cristina, Alejandro and Yash, thank you for all the help and the interesting chats we have had in the labs over the years.

Finally, I would like to thank Dr. Nolan for all her help throughout my PhD. No matter how stressed or hard it got, you were always there to pick me up. I would also like to say thank you to my mom Christina for her never-ending support.

Acronyms

AAV Adeno-associated virus.

BLA Biologics licence application.

cDNA complementary DNA.

CHO Chinese Hamster Ovary.

CLP Cre-Lox-P.

CMV Cytomegalovirus.

CRISPR clustered regularly interspaced short palindromic repeats.

DRE Dioxin response element.

DTE Difficult to Express Protein.

E.coli Escherichia coli.

ELISA Enzyme-linked immunosorbent assay.

FCMV Full-length CMV.

FPFT FLP-FRT.

GGPPS Geranylgeranyl diphosphate synthase.

GNU Generating New Units.

GOI Gene of interest.

GS glutamine synthetase.

HC Heavy chain.

INR Initiator element.

ITR Inverted terminal repeats.

LC Light chain.

MARS Matrix attachment regions.

MMEJ Microhomology-mediated end joining.

MPR Mis folded protein response.

NHEJ Non-homologous end joining.

PCA Principal component analysis.

PCR Polymerase Chain Reaction.

PIC Pre Initiation Complex.

QbD Quality by Design.

RLA Relative luciferase activity.

RNAPolII RNA polymerase II.

SD Standard deviation.

SEAP Secreted Alkaline Phosphatase.

STAR Stabilising anti repressors.

TALENs Transcription activator-like effector nucleases.

TFRE Transcription Factor Regulatory Element.

TPM Transcripts per million.

UCOE Ubiquitous chromatin opening element.

UPR Unfolded protein response.

Abstract

Chinese hamster ovary cells are one of the most widely used in industry due to their ability to create recombinant proteins and their historical use in the biopharmaceutical industry. The medicines being produced are getting more and more complex. However, the genetic engineering approaches are not advancing at the same rate. Current technologies still rely on random screening approaches and nearly wholly use two CMV promoters on the genetic cassette.

This work aimed to try and apply differing designs of synthetic promoters to try and increase expression over the original system. The thesis first focused on trying to build a pipeline which could find building blocks to create synthetic promoters in a more high throughput way. This worked and a completely new TFRE was found called NFE2l2, which rivalled the activity of NFkB, achieving 80% of the activity of full length CMV(FCMV).

The next section of this thesis looked at testing different designs of synthetic promoters to see if there are any design considerations that should be considered. This was tested transiently using SEAP as a high throughput, cheap method of screening them. The 3' weighting of promoters was found to be beneficial, producing 80% more SEAP then the FCMV control along with the inclusion of new TFREs found in the first section of this work.

For industrial relevance, CHO cells were transfected to produce an antibody. Synthetic promoter constructs mimicking the CMV system were tested and compared to the same antibody being produced by a combination of promotes. It was found using two different synthetic promoters achieved the maximal titre output with a increase of 180% over the control. Bidirectional promoters were also created and tested, with one promoter achieving on average 80% over the control.

Finally, to see how the transiently tested constructs performed in Merck's stable systems some of the constructs were onboarded and tested. In general, what was found is that the transient screens are not a good indicator of stable expression, and the synthetic promoters had a similar median expression compared to the control when tested in a beacon. Although, the dispersion of data is more prominent in some bidirectional constructs and may be beneficial to future CHO cloning strategies.

Contents

List of Figures

List of Tables

Chapter 1

Literature Review

Overview

- This chapter covers the literature surrounding the works of this thesis.
- Sections 1.1 and 1.2 cover a general introduction to the biopharmaceutical space and discuss available host cell expression systems for recombinant protein production.
- The current vector engineering tool kit for CHO cells is discussed in Section 1.3 with a particular focus on alternatives to promoter engineering.
- New theories on the mechanisms of transcription will be discussed in Section 1.4, along with the current works on synthetic promoter design in CHO cells.
- Bidirectional promoter construction and potential avenues for synthetic design are reviewed in Section 1.5.

1.1 Introduction

Bio-pharmaceuticals have gained market share over traditional pharmaceuticals in recent years. Recently, with the response to COVID-19, there has been increased awareness about how important these bio-pharmaceutical products are to the mainstream public. Companies such as Merck, Sharp and Dohm, Eli Lilly, Phizer and Johnson and Johnson have been increasingly focusing on the bio-pharmaceutical side of their businesses for years now (Moorkens et al., 2017). The global market for bio-pharmaceuticals is projected to be $300 billion annually by 2025 (Lu et al., 2020).

CHO cells are currently the most widely utilised cells for producing these therapies. Approximately two-thirds of all recombinant molecule-based treatments are being produced in this system (Jayapal et al., 2007; Kim et al., 2012) and as such many improvements have been made to the platform. These include; vector design, clonal selection strategies, process, transfection and feed improvements which, together, have led CHO cells to achievable titres of up to 9 g/L. The techniques above rely heavily on laborious blind screening platforms, even with CHO being such a pivotal cell line. Chance plays a large role in the current method to produce a clonal derivative capable of producing and secreting a complex biomolecule at an industrially viable standard. This process can take up to 9 months to produce a usable clonal derivative and is not guaranteed to be the best clonal population from the heterogeneous pool. The majority of single-cell selected cells will die during the process.

Biopharmaceutical molecules are increasingly becoming more complicated with the emergence of bi-specific antibodies, trifunctional antibodies and even chemically linked antibodies which can be classified as difficult to express (DTE) molecules due to low protein titre or in some cases cytotoxicity to the cell (Alves and Dobrowsky, 2017). In some cases, a molecule can even fail to be market viable due to the blind screening method being unable to produce a clone viable of high sustained expression of the molecule. In the case of a DTE molecule, much like in the cell itself, the first rate-limiting step is transcription. Mead et al. (2009) discuss how the initial limiting step in the production of DTEs is transcription and translation/translocation bottlenecks that only appear in the highest producers.

Not only are issues appearing now in more complicated versions of recombinant molecules, but transcriptional control will become an even larger issue with Adeno Associated Virus (AAV) therapy, gene therapy and personalised medicine(like CAR-T therapy) becoming commonplace. Due to this, a platform to control all caveats of transcription within the cell is required. Promoters cannot just control the production rate through the promoter; it is the first step in the process. Other variables can also have a large impact on the rate of transcription, such as the gene's location in the genome, the chromatin environment surrounding the gene, and the gene cassette itself. The promoter gene cassette will never reach full transcriptional activity if the two prior variables are unsuitable.

A case in point, how will it be possible to ensure a transgene transfected into a cell is

efficacious? Not only will a promoter of a certain transcriptional power be required but other variables such as; the stability of the transgene, selective insertion to ensure the gene is in a transcriptional hotspot and ensuring their one copy number must be considered. If current methodologies were applied, a random mix would be screened until a suitable combination was found, but this is time-consuming, laborious and costly. Platforms already exist to produce large mammalian gene circuits (Guye et al., 2013) and industrially relevant synthetic promoters (Johari et al., 2019). Still, the fine tuning of transcriptional activity, including all the variables is not dealt with in detail. Brown and James (2016) already discussed the need for vectors with multiple genetic components with designed stoichiometry and provided designs in future works (Brown et al., 2017, 2019). The next step in further reducing the development and cost of these processes is to further push the “design space" into a predictive, fast, efficient and stable design environment that improves synthetic promoters’ design and helps increase titre.

1.2 The Biopharmaceutical Process

The process of taking a new pharmaceutical drug to market is extremely costly, estimated to cost an average of $985 million (Wouters et al., 2020). The statistic exasperates that 54% of new pharmaceutical drugs fail to make it to market due to issues with efficacy, clinical safety or even the inability to scale up the production process. Nowadays, clinical development and process development are usually performed in parallel and although not directly related to the process, it is an important concept to remember. This is known as Quality by Design (QbD) and its function is to try and reduce the time to market and also the cost of getting the product to market. Figure 1.1 explains this process in more detail.

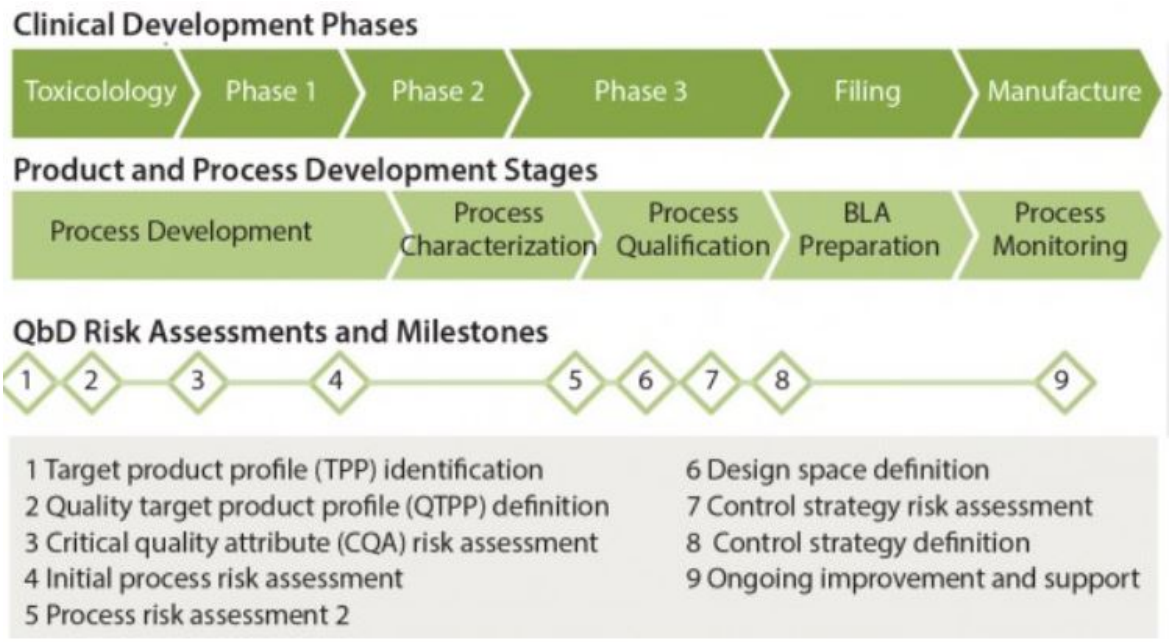


Figure 1.1: Explanation of Quality by Design. The quality by design process tries to parallel drug development's clinical and manufacturing side. Previously drugs have taken an excessive amount of time to reach the market due to the clinical trials being passed and then the design of the manufacturing process being examined. Creating a parallel system can save both time and money. If the drug fails due to production issues, the clinical trials can be halted or vice versa if the new drug appears dangerous. BLA stands for biologics license application. Figure taken from Cooney et al. (2016)

1.2.1 The Expression System Of Choice

As Figure 1.1 depicts, process development can take a substantial amount of time and the very first step in this process is the selection of the host cell that the product will be made. Usually, the harder the product is to produce and the more post-translational modifications it possesses, the more complex the host organism will need to be. Selecting the host type is very important, as it will determine the shape of the whole process, including but not limited to: the gene cassettes, scalability, yield,

efficacy, immunogenicity and cost. Each has many advantages and disadvantages, with bacteria being the most simple and cost-effective solution.

Bacteria are a fast-growing, easy and cheap way to produce protein and are commonly used to replicate DNA plasmids. The most common strain used for protein production tends to be *Escherichia coli* (*E.coli*) (Chen, 2012). Of all the expression systems, they grow the quickest. They are likely the most scalable of all the host systems, reaching a theoretical maximum concentration of 1×10^{13} cells/mL (Rosano and Ceccarelli, 2014). Of course, this would not be achieved without intensifying the process, but in the case of bacteria, this would be much simpler to achieve compared to other expression systems.

The main drawback to bacteria and the reason they are not used in the production of recombinant proteins is their lack of post-translational modifications. When it comes to glycosylation or protein folding, bacteria cannot achieve this. Not only could this lead to a recombinant protein which is functionally inert due to the lack of folding, but it could also be insoluble and require harsh conditions to purify the protein. Interestingly, there is some research into making bacteria viable for producing glycosylated proteins, as if possible it would be the most efficient system available (Du et al., 2019). Although the benefits of this might be minimal, as yeast already has a lot of the benefits of bacteria but can also perform post translational modifications.

Yeast is a very interesting host as they, unlike bacteria, are a eukaryotic host and as such are more suited to producing mammalian proteins. Unlike their mammalian counterparts, yeast also proliferates quickly, have much cheaper media requirements and can produce more than double the amount of antibody produced in terms of volumetric productivity (Gasser and Mattanovich, 2006). The most commonly used strain of yeast used is *S. cerevisiae*, also known as baker's yeast, but other strains have been used, such as *Pichia pastoris*, to varying degrees of success (Lee et al., 2015b). Knowing this, one may wonder why we don't use these systems generally for the production of recombinant proteins? The answer is that because they are quite different from mammalian cells, their glycosylation pattern is different and this can cause increased immunogenicity when used as a medical treatment. They can also possess proteases which further limit their viability in the production of recombinant molecules. Insect cells are the next step along the ladder of complexity and offer similar post-translational modifications to mammalian cells.

Insect cells are the first host system to be discussed, naturally capable of recombinant protein production. They are a transient form of protein production which uses the baculovirus expression vector system to be transduced with DNA (Tripathi and Shrivastava, 2019). This means they require culturing to the required cell density and then being transfected with the gene of interest. Compared to bacteria and yeast alternatives, the media and culturing conditions are much more expensive and insect cells can still not carry out N-glycosylation (Shi and Jarvis, 2007). For this reason, mammalian cells have still stayed the dominant cell line in industry but the other hosts, as aforementioned, do have valuable benefits and in certain circumstances, are better

than mammalian cells.

1.2.1.1 Mammalian Host Cells

Currently, mammalian cells dominate the production of recombinant proteins due to their innate ability to produce large molecules and perform advanced post-translational modifications. From 2015 to July 2018 79% of all biopharmaceutical products approved were produced in mammalian cells and of that 79%, 84% were produced in CHO cell lines, followed by 13% in NS0 and finally, 3% in SP2/0 cells (Walsh, 2018). Figure 2 depicts the general process by which stable cell lines are generated for CHO systems.

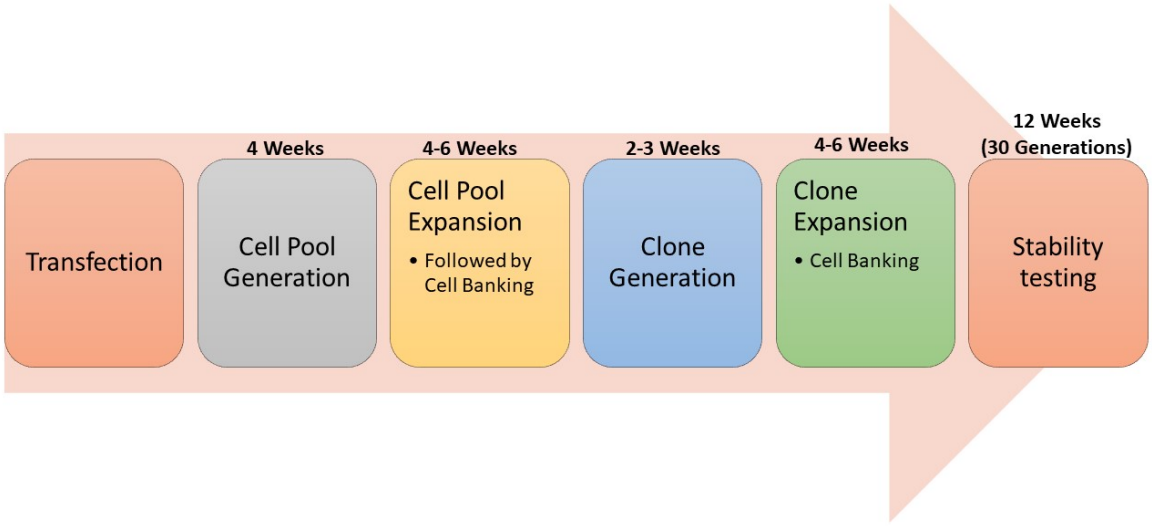


Figure 1.2: Summary of CHO stable cell line generation. CHO stable cell lines are generated using a system known as GS selection. Figure adapted from Noh et al. (2018). A transfection is first performed on CHOK1 or CHOS cells. This is followed by cell pool generation, in which varying concentrations of MSX are added to only select cells which have the GS gene integrated into their genome. The cells are then expanded and banked. Following on from this cells are single cell sorted, usually using FACS and the single cells that grew and showed promising productivity values are taken forward to clonal expansion. The final test is stability testing, in which cells are grown for 30 generations to ensure no productivity loss over time.

Other cell lines more closely related to humans do exist such as HEK293 and HEK293T cells (Hu et al., 2018; Hunter et al., 2019), although these cells have primarily been used for transient culture and not stable, due to their high transfection efficiency and the incorporation of the T-antigen to allow episomal replication (Ho and Pastan, 2009). CHO actually suffer some disadvantages compared to human-derived cell lines, such as differing glycosylation patterns which can cause negative immunogenetic reactions in patients (Dumont et al., 2016) due to increased amounts of galactose-α,3-galactose and N-glycolylneuraminic acid (Dumont et al., 2016). They also lack

certain glycosylation patterns which humans create, such as α[2-6] sialyltransferase and α[1-3/4] fucosyltransferases (Ghaderi et al., 2012).

Knowing this, why is CHO the most dominant cell line in industry? CHO was first used to create a recombinant product in 1986 and can achieve protein titres of up to $10g/L$ (Kim et al., 2012). This means it has had a lot of time for improvements to be made to the systems which support it, such as glutamine synthetase (GS) selection (Noh et al., 2018), vector engineering, site-directed integration and media engineering (Kim et al., 2012). This has allowed it to outshine the other potential hosts despite its drawbacks. The issue with glycosylation can even be reduced by screening stably selected cell lines for variants with much lower levels of unfavourable glycan profiles. One lesser known reason for CHO cell favourability over their human counterparts is their reduced susceptibility to human viruses (Swiech et al., 2012). These advantages, along with their proven record since 1986 make them a favourable cell line to industry and regulatory bodies, as they know what to expect when using them. For this reason, the rest of the review will focus on CHO-specific technologies. Although, a lot of the techniques to be discussed will have applications in not just other mammalian cells but other expression systems also. In the next section, how these DNA engineering techniques can be used to increase CHO cell expression by primarily affecting the transcriptional landscape will be discussed.

1.3 Vector Engineering Elements for CHO Cells

This section intends to discuss the current and emerging methods of improving the gene cassette used for CHO cells to achieve controlled and efficient gene expression, while also providing the basis for a new synthetic biology tool kit, to provide the desired transcriptional environment for any gene of interest. First, methods of introducing a transgene with reduced variability will be outlined and the applications of these technologies in relation to stable CHO cell generation will be presented. From there, ways to increase the stability of the transgene will be examined, considering recent emerging technologies and the pros and cons of all approaches mentioned. Lastly, discussion will be directed at how these technologies complement each other and allow the generation of a CHO clonal population with reduced costs, time and increased productivity.

1.3.1 Transient and Stable Transfection

The most basic way of introducing genes of interest is through transfection. Different transfection methods can lead to a 10-fold difference in the transcriptome of a cell (Jacobsen et al., 2009). Cells, once transfected, can be used transiently or undergo a selection process and be expressed stably. Stable cell lines have the gene of interest integrated into the cell's genome. In transient transfection, the DNA remains episomal and the gene of interest is translated into a protein over the course of a few days or weeks. The main downside to transient transfection is the plasmid dilution, as expression is lost as the cells divide (Kim and Eberwine, 2010). For this reason, transient expression is used for a product that is needed quickly (i.e. for clinical research). In contrast, stable expression is used for large-scale manufacturing due to its reduced variability and the ability to isolate a high-producing clonal population. In light of this, the current review will focus on transcriptional control in CHO for biopharmaceutical purposes, aiming to better control transcription for stable expression platforms.

Although different transfection methods exist, electroporation is the most widely used for the generation of stable clones is electroporation (Kim and Eberwine, 2010), followed by GS selection. GS selection is generally used to cause gene amplification of the transgene using selective pressure to ensure stable clones with high productivity and stability are created. However, electroporation does not guarantee stable integration and even when it does, there can be variability in expression due to the random integration of the non-genomic DNA. This is further worsened by the GS system increasing productivity by gene amplification. Not only does this impact transcriptional control by introducing an unknown environment for the gene of interest, but it also reduces the chances of picking a high-producing clone and increases the amount of random screening required to produce a suitable clone. Technologies such as transposons aim to deal with the issues of low integration rates (Matasci et al., 2011), while technologies such as Cre-Lox-P (CLP) (Kawabe et al., 2015; Bode et al., 2000), FLP-FRT(FPFT) (Zhou et al., 2010), transcription activator-like effector nucleases (TALENs) (Sakuma

et al., 2015, 2016) and finally, the newest technology clustered regularly interspaced short palindromic repeats (CRISPR)/cas9 mediated homologous recombination (Lee et al., 2015b,a; Chi et al., 2019) aim for more thorough site directed integration and even one gene copy cell line platforms.

1.3.2 Targeted Integration Technologies

1.3.2.1 Transposons

Transposons are unique regions of DNA that transfer themselves from one region of the genome to another using the transposase enzyme. There are three types of transposons; (i) excision and relocation using transposase; (ii)Helitrons, which use a mechanism likely related to rolling circle replication; and (iii) Mavericks, whose mechanism of action is currently unknown (Feschotte and Pritham, 2007). Three transposons, piggyBac, Tol2 and Sleeping Beauty, have already been utilised in CHO with successful results (Balasubramanian et al., 2016). These transposons are excision and relocation transposons, which are active in mammalian cells (Wu et al., 2006). The basic mechanism of how they work is similar; Figure 1.3 shows the mechanism for the piggyBac transposon.

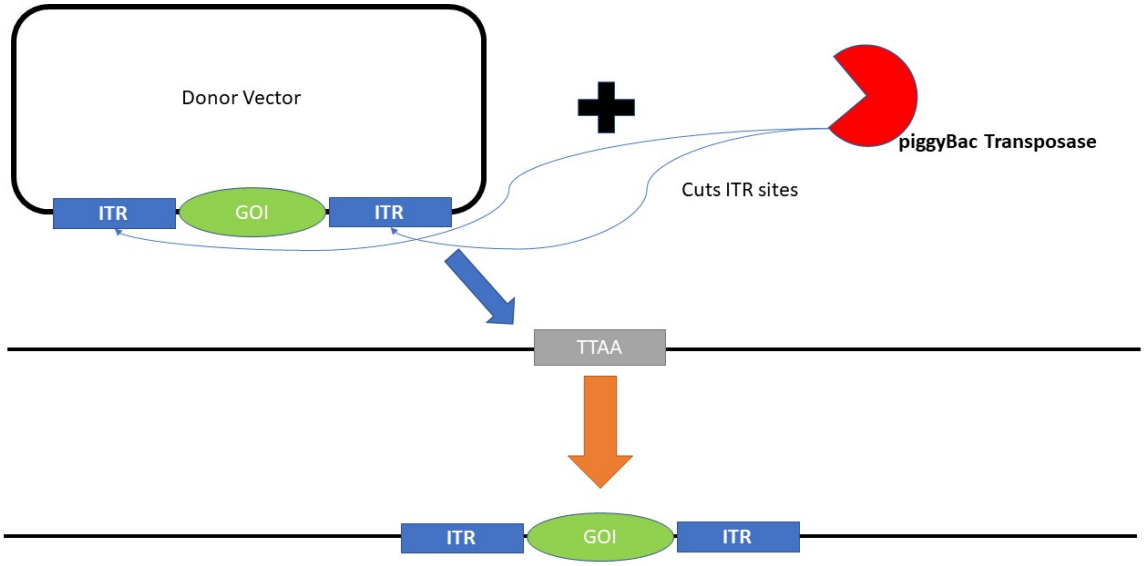


Figure 1.3: Explanation of PiggyBac Transposase. This figure shows the Donor vector with the inverted terminal repeats. The gene of interest will be put between these ITRs. The transposase cuts the ITRs sites and begins the cut and paste mechanism, where the transposase enzyme causes a staggered cut in the plasmid DNA and this will then integrate into certain areas of the genome that contain the TTAA sites. Figure adapted from Mann et al. (2008).

Although they all share similar mechanics, they have different characteristics. Tol2 favours integrating close to TSS start sites and A-T rich sequences (Kondrychyn et al., 2009; Parinov et al., 2004; Grabundzija et al., 2010). Alternatively, piggyBac is predisposed to inserting in transcription units (Wilson et al., 2007; Liang et al., 2009; Galvan et al., 2009) and sleeping beauty relies on a V_{step} structural pattern for integration (Geurts et al., 2006). All transposons differ in their species origin, cargo sizes, targets and properties; Table 1.1 shows the critical characteristics of key transposons applicable to CHO.

Table 1.1: Characteristics of three transposons for use in CHO.

Transposon	Integration Site	Cargo Size	Overproduction Inhibition (OPI)	Copy Number per Cell
Sleeping Beauty	V_{step} structural pattern (TA centre)	$> 10Kb$	Heavy influence	10
piggyBac	Transcription Units (TTAA)	$> 100Kb$	Minor Influence	2
Tol2	Close to gene start sites and AT-rich sequences	$> 10Kb$	Minor Influence	2

The application of transposons may not lead to tight control of transcription, but it does improve the protein production potential of the cell. It may even allow the substitution of transient expression with mini clonal pools. Cells transfected with this transposon technology have been shown to have titres nine-fold higher than cells created by normal clonal generation (Balasubramanian et al., 2016) and recombinant titres four-fold higher (Matasci et al., 2011), along with a stability of over 3 months in both cases. Balasubramanian et al. (2016) showed piggyBac and Sleeping beauty had an advantage over Tol2 in terms of transfection efficiency. This finding is supported by several other studies which have found piggyBac to be highly efficient in the generation of stable expression platforms (Matasci et al., 2011; Galvan et al., 2009; Grabundzija et al., 2010; Meir et al., 2011; Wilson et al., 2007).

Transposon technology will not only allow the generation of rapid bulk pools but will also allow the clonal screening process to be reduced from up to 6 months to obtain a clone to as little as 4 weeks (Balasubramanian et al., 2016). Besides this, Grabundzija et al. (2010) found that Tol2 insertion sites were under represented in areas of transcriptionally repressed heterochromatin, indicating transposons may target more active areas of the genome. In terms of a transcriptional control platform, this would be the perfect integration method in the case of a new cell line where no transcriptional "hotspots" are known or even when getting a product to market quicker. Suppose site-directed integration cannot be performed. In that case, transposon technology will be vital, as although the gene of interest is not going into the same transcriptional hotspot or open chromatin environment, the impact of the surrounding genome will be less restrictive and repressive on the parts of transcription which can be fine tuned (Grabundzija et al., 2010). This transposon technology would then be followed by site-directed integration once more information about the new host cell or recombinant molecule is known.

1.3.2.2 Cre-Lox-P and FLP-FRT

These systems, being the oldest technology, require the most time to use for site directed integration. CLP and FPFT work through site-specific recombinases to insert a gene of interest and have been used in CHO as a method of site-directed integration to produce stable clones (Kameyama et al., 2010; Obayashi et al., 2012; Huang et al., 2007; Kito et al., 2002; Zhou et al., 2010). Both systems work by a similar method and an example of CLP is shown in Figure 1.4. Although the system can be used as site-directed integration, it is even more laborious as a plasmid with the target integration site must be transfected into a host and then the cells must be clonally selected and a single recombination site needs to be checked for by polymerase chain reaction (PCR) (Kameyama et al., 2010). Site-directed integration through this method is very laborious and the reliance on near random or very few targeted integration sites makes this method unsatisfactory. Before the CHO genome was sequenced, this was one of the best methods for site-directed integration but technologies like TALENs make this process much quicker and more affordable (Lee et al., 2015b).

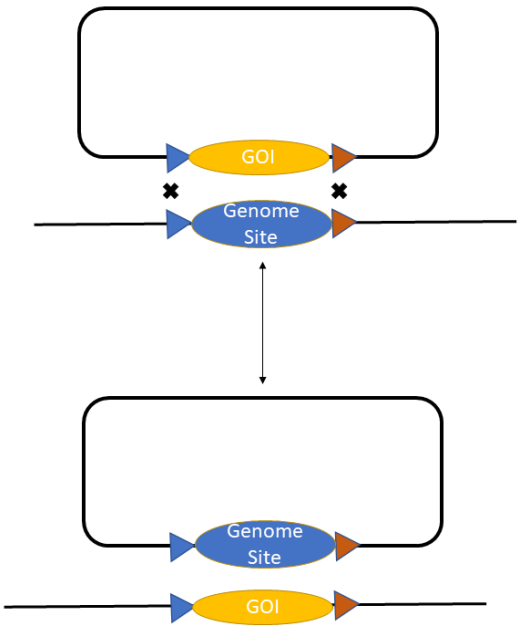


Figure 1.4: Summary of cre-lox-p cassette exchange. This figure gives a general idea of what happens in the cre-lox-P cassette exchange. This process occurs in the presence of cre recombinase.The integration sites, unlike transposase, must be inserted into the clone before site directed integration can occur. After the recombinase reaction, the gene of interest (GOI) is inserted into the genomic DNA.

1.3.2.3 TALENs

TALENs and CRISPR/cas9 work similarly through a guided nuclease, causing a double strand break in the DNA. The cell views this damage as a potentially lethal occurrence and initiates specific pathways to repair the DNA, which can be taken advantage of to introduce a transgene (Carroll, 2014). TALENs are a natural progression from another technology called zinc finger nucleases. They work in a similar way but TALENs are

more cost-effective and easier to modify (Chandrasegaran and Carroll, 2016). TALENs main mechanism of action is recognition through the TALE binding molecule, which was found in plant pathogenic bacteria from the genus *Xanthomonas* (Moscou and Bogdanove, 2009). TALE binds to DNA by single base pair recognition system in which an amino acid residue recognises a single base. For instance, the amino acid couple NI recognises the base A (Chandrasegaran and Carroll, 2016). The ability to join different TALE repeats together allows the targeting of any area of interest in the genome (Reyon et al., 2012). The only caveat is the DNA sequence must be followed by a T to ensure TALE activity (Boch et al., 2009). The guiding ability of the TALE molecule was fused with a *IFokI* nuclease allowing the creation of site-directed DNA double-strand breaks (Bogdanove and Voytas, 2011). *IFokI* only works as a dimer and the TALEN system works much like PCR primers. The double-stranded break target site is designed in pairs, one for the forward and one for the reverse, with a spacer in-between which allows *IFokI* monomers to dimerize and result in a double-stranded break in the DNA (Bogdanove and Voytas, 2011). When this double-stranded break occurs, a donor plasmid can be used to insert a gene of interest using i) homologous recombination ii) non-homologous end joining (NHEJ) and iii) microhomology-mediated end-joining (MMEJ). The most common of these is homologous recombination, as it's less error-prone than NHEJ (Nakade et al., 2014). MMEJ is an upcoming technology that addresses the issue of low homologous recombination in higher eukaryotes. Unlike homologous recombination, which targets the S/G2 phases of cell growth, MMEJ targets G1/early S phase (Nakade et al., 2014; Taleei and Nikjoo, 2013). It has been suggested that the MMEJ pathway may be superior to the other pathways for gene knock-in (Nakade et al., 2014; Sakuma et al., 2016, 2015).

For CHO, things become even more difficult for site-directed integration. Homologous recombination has been shown to be lower in CHO than human cells (Orlando et al., 2010; Cristea et al., 2013; Lee et al., 2015b) and for that reason, NHEJ approaches have primarily been used (Orlando et al., 2010; Cristea et al., 2013). However, this approach is more error-prone (Lee et al., 2015b). MMEJ has been cited as the main mechanism driving genomic integration in CHO (Kostyrko et al., 2017; Kostyrko and Mermod, 2016). Contrastingly, Bosshard et al. (2019) found that homologous recombination and MMEJ may rely on each other. The reason for low homologous recombination in CHO was elevated expression levels of Mre11, Pari and low Rad51 expression. Once these effects were counteracted, site-directed engineering increased by up to 75% (Bosshard et al., 2019). Nevertheless, several studies have shown the use of MMEJ is possible in CHO using TALENs, but these did have issues with off-target integration, partial integration and mutations (Nakade et al., 2014; Sakuma et al., 2015, 2016). Sakuma et al. (2015) suggests that using CRISPR and MMEJ may increase the specificity by allowing the microhomologies to vary in length. Although CRISPR and TALENs are very similar methodologies, TALENs suffer in comparison due to laborious cloning. Still, some methodologies such as the one described in Reyon et al. (2012) should alleviate these issues. The main advantage of TALENs in comparison is their increased fidelity due to the need for *IFokI* monomers to dimerise. Similar to the move from

systems such as Cre-Lox-P to TALENs, CRISPR/Cas9 is arguably a more efficient and cost-effective solution for site-directed integration.

1.3.2.4 CRISPR/Cas9

CRISPR/cas9, since its introduction in mammalian cells (Cong et al., 2013) in 2013, has gained immense popularity for gene editing due to its simplicity and efficiency compared to the other genome editing tools (Doudna and Charpentier, 2014; Hsu et al., 2014). CRISPR/Cas9's mechanism of action is similar to that of TALENs. CRISPR-cas9 works on the basis of transfecting a cell with an endonuclease such as cas9 and a guide RNA. The guide RNA works much like the TALE molecule's ability to recognise single base pairs of DNA. Guide RNA consists of; the target DNA sequence, Protospacer Adjacent Motif (PAM) and a scaffold complex which binds to the Cas9 protein (Hsu et al., 2014). Out of all genome editing techniques, CRISPR has the easiest targeting strategy in which the target can be altered to any 20 nucleotide sequence in the genome as long as it is unique in comparison to the rest of the genome and the target is adjacent to a PAM sequence (Hsu et al., 2014). The need for a PAM sequence could be considered one of CRISPR's weakest areas. Still, different orthologues of the Cas9 proteins have different PAM sequences and newer methods allow the controlled alteration of the recognition to PAM sequences (Nishimasu et al., 2014). When the CRISPR/cas9 and the guide RNA bind to the target sequence, the PAM motif allows cas9 to create a blunt double-stranded break in the DNA (Hsu et al., 2014). Ways to increase the DNA editing specificity do exist, such as using Cas9 (Jinek et al., 2012) which works much like TALENs, in which pairs are required to cause double stranded breaks (Ran et al., 2013).

Most site directed integration in CHO has used homologous recombination (Zhao et al., 2018; Lee et al., 2016, 2015b; Chi et al., 2019), which suffers from the same issues as previously mentioned in TALENs. Although this method may have low efficiency, it has been used to successfully target the C12orf35 hotspot in the CHO genome with a targeting efficiency of 7.4% (Lee et al., 2015b). The low efficiency means clonal selection is still required, but the heterogeneity of the population should be reduced and shorten the number of steps required in clonal selection. Although the efficiency of homologous recombination is low with both methods, technologies are emerging that allow the use of homologous recombination but avoid time-consuming steps such as selection markers. Lee et al. (2015b) use fluorescent enrichment in which the cas9 and target DNA are labelled so that cells can be FACS sorted by the presence or absence of the target DNA and cas9, which led to a three-fold increase in the number of cells that had homologous recombination present. Another novel system that adds to gene insertion using homologous recombination is described in Chi et al. (2019), in which standard homologous recombination is used to insert a “landing pad" with PhiC31 attP sites and PhiC31 integrase. This allowed the creation of a stable CHO-S cell line with a landing pad for any gene of interest. Chi et al. (2019) then transfected a transgene of interest flanked by attB sites. The study found an increase to 97.7% transfection

efficiency and found the population to have less heterogeneity and increased stability for over forty passages. This method has also been performed with BxB1 and FPFT (Inniss et al., 2017). Similar to the TALENs method, several studies have looked at non-homologous end joining for gene insertion using NHEJ (Bachu et al., 2015; He et al., 2016) and the same method of MMEJ as was discussed previously for TALENs, known as CRIS-PITCH (Nakade et al., 2014). CRISPR-cas9 becoming so popular has also led to the development of useful tools to help target specific areas of the CHO genome. CRISPy, described in (Ronda et al., 2014), allows quicker genome engineering, as it reduces workload and variability, allowing the analysis of gRNA off-target effects before experimentation. CRISPR has many advantages over other genome editing techniques; it's easier to create target sequences, more efficient than TALENs (Nemudryi et al., 2014) and it can be used for multiplex genome editing as several guide RNAs can be transfected at once. The disadvantages of CRISPR lie in its potential off-target effects (Zhang et al., 2015) and decreasing recombination efficiency with size (Kung et al., 2013).

1.3.2.5 Which To Choose?

The decision of which technology to use should be based on the desired outcome of the cell line, time constraints and budget allocated. For mini-clonal pools, new cell lines or cells in which a hotspot is currently unknown, transposase would likely guarantee the best environment for the transgene, as they are believed to avoid repressive heterochromatin environments. If the gene is extremely large (greater than 10Kb), piggyBac is not affected by insertion length, unlike other gene insertion techniques that use homologous recombination. As opposed to the above, a cell line for biopharmaceutical production will likely have time for a platform to be set-up. For this reason, CRISPR or TALENs would be beneficial, as a more specific site of the genome can be targeted to ensure minimal impact on the further steps of transcriptional control. CRISPR would likely be used in most cases due to its ease of use and flexibility, but if specificity is required, TALENs may be a better option. A perfect system for the biopharmaceutical industry may be using systems such as those described in Chi et al. (2019); Inniss et al. (2017). In this system, the laborious workflow to ensure the transgene is inserted in the correct place is only carried out once, establishing a base platform. Then recombinant antibody cassettes can be inserted using older technologies such as CLP and FPFT, along with a negative selection marker. Compared to random integration, each method has its benefits with increasing productivity, reducing time expenditure and improving the baseline for further engineering of the gene cassette to obtain optimal transcriptional stoichiometry or transcriptional power.

1.3.3 Stability Elements

CHO stable cell lines can suffer from a loss of productivity over time (Bailey et al., 2012). This is thought to be due to transcriptional silencing, as gene copy number remains the same (Yang et al., 2010; Klose and Bird, 2006; Osterlehner et al., 2011).

This is often thought to be related to CpG dinucleotides promoting DNA methylation (Osterlehner et al., 2011). Still, conflicting evidence suggests CpG island elements can increase stability, even when inserted into a different promoter (Mariati et al., 2014). CpG islands are associated with the majority of extremely stable housekeeping genes (Farré et al., 2007) and the fusion of a partial ACTB promoter with CMV led to four fold higher gene expression (Zúñiga et al., 2019).Stability elements have become a huge area of interest to increase stability in stable CHO cell lines. Several elements exist to enhance stability, including; i) insulators (Maksimenko et al., 2015; Izumi and Gilbert, 2000; Naderi et al., 2018; Chen et al., 2017) ii) matrix attachment regions (MARs) (Harraghy et al., 2011; Majocchi et al., 2014; Girod et al., 2005) iii) Stabilising Anti Repressor (STAR) (Romanova and Noll, 2018; Saunders et al., 2015; Kwaks et al., 2003) iv) ubiquitous chromatin opening elements (UCOEs) (Betts and Dickson, 2016; Majocchi et al., 2014; Saunders et al., 2015; Williams et al., 2005) and v) the novel regulatory element (E77) found in the CHO-K1 genome (Kang et al., 2016). In terms of mechanism, insulators are the most understood of these systems, dividing the genome into segments (Romanova and Noll, 2018).

1.3.3.1 Insulators

Insulators are elements that can block distal enhancers' effects on a promoter and act as barriers to condensed chromatin which may silence the transgene of interest (West et al., 2002). They were found in *Drosphila melangaster* initially but have also been found in eukaryotic organisms and are thought to be more prevalent than previously predicted (Kim et al., 2007). The most researched insulator is HS4, which originates from the 5‘ end of the chicken beta globulin locus (Yusufzai and Felsenfeld, 2004). They are usually large molecular elements, with HS4 being 1.2 kB (Maksimenko et al., 2015). Still, a 250bp core fragment has been used which retains the activity of the complete insulator and for this reason, it is used within vector constructs with limited size constraints (Hanawa et al., 2009; Emery et al., 2000). The mechanism of action has been described in Nakayama et al. (2012); Ghirlando et al. (2012); Vorobyeva et al. (2013); Yajima et al. (2012) but only one insulator binding protein CTCF has been found in mammalian cells (Herold et al., 2012). In essence, insulators act as binding sites for the CTCF protein (Bastiaan Holwerda and de Laat, 2013; Jia et al., 2020), which allows two insulators to interact, resulting in a “gene loop" which is insulated from the rest of the genome (Tokuda et al., 2011).

Very few studies exist which show the effect insulators can have on CHO protein production, but those that do exist show little improvement. HS4 has been found to have little effect on stopping epigenetic silencing of the CMV promoter (Romanova and Noll, 2018) and insertion into a vector causes no significant increase in expression (Takagi, 2017; Saunders et al., 2015). Although HS4 has had little effect in CHO cells, the tDNA insulator described in Naderi et al. (2018) achieved up to nine fold increases in antibody titre. This exhibits the need for more diverse research into insulators for application in CHO and should be further improved through genome mining for

endogenous insulator elements (Takagi, 2017). Unlike insulators, MARS elements have had more application in CHO.

1.3.3.2 MARS

MARS elements, also known as scaffold matrix attachment regions, are regions of AT-rich sequences which are thought to bind a nuclear matrix protein (Mirkovitch et al., 1984). S/MAR regions are thought to function similarly to insulators in that they create structural loops within the genome and have been shown to function in multiple species (Harraghy et al., 2008). Congruently to insulators, one of the first MARs elements used in CHO was from chickens and led to significantly higher levels of transgene expression (Girod et al., 2005). Unlike insulators, MARs elements are thought to do more than just insulate from potentially repressive chromatin and may enhance transcription. There are several ways in which it is speculated to achieve this; i) MARs elements may act like transposons and lead to site-directed integration in transcriptionally active regions (Grandjean et al., 2011; Puttini et al., 2013; Wang et al., 2010) ii) insulation from repressive chromatin (Harraghy et al., 2011) and iii) promoter activation through binding of transcription factors to the MARs binding motif (Albrethsen et al., 2009; Harraghy et al., 2011; Girod et al., 2007). MARs not only have application in stable cultures but may be useful for transients also, with several studies reporting episomal replication of transient DNA using MARS elements (Wang et al., 2019; Wong et al., 2011a). Not only have they been used for episomal replication, but modifications have reduced the MARs element to fragments as little as 500bp in length and had higher transgene expression in CHO cells (Wang et al., 2019).

MARS elements have been tested in CHO with varying degrees of success. The MAR X_S29 described in Girod et al. (2007), achieved an average four fold increase in expression. In comparison, Zahn-Zabal et al. (2001) looked at how MARs elements could help with stable cell generation and found an increase in the proportion of high-producing cells, thus reducing the need for clonal selection. The study also found no increase in expression once selection was performed with methotrexate, which could eliminate a time-consuming process if the recombinant expression was sufficient. Kim et al. (2004) found similar results but the Beta-globulin MARS element was the best. More recently, Harraghy et al. (2011) found the S4 MAR can be used to achieve high levels of expression but found that in the presence of selection marker and backbone sequences in transient culture, the benefits were abolished. When comparing the MARs element to other stability elements, the MARs element can increase protein expression by up to two fold, which is below the UCOE but above the more unknown STAR element (Saunders et al., 2015)

1.3.3.3 STAR

STAR elements are regions of the human genome that have been found to counteract chromatin-associated repression. The mechanism by which these stability elements function is currently unknown (Kwaks et al., 2003). They range in size from 500bp to

2.1kb and a method to quickly assess their stability enhancing effects by preventing the silencing of a selection marker has been described in Hoeksema et al. (2011). Some of these elements have been tested in CHO cells, such as STAR 40, which showed little activity over the control (Saunders et al., 2015). STAR 7 has been compared to the full-length CBX3 UCOE and was found to be the most beneficial for protein expression versus all other stability elements (Otte et al., 2007). The only issue in this study is that the UCOE had no benefit on gene expression which is unusual. The usefulness of the STAR system for the generation of stable clones can be seen in the creation of the STAR-select system, which allows the rapid generation of very few, but very high-producing clones (van Blokland et al., 2007). STAR systems used in production cells have been reported to have little effect on cell survival and transgene expression (Romanova and Noll, 2018). This is in contrast to the UCOE, which has become a mainstay of cell line development and has become more effective with recent engineering developments.

1.3.3.4 UCOE

UCOEs are DNA elements which have the unique ability to “open" chromatin to ensure transcriptional activity. The first UCOE was derived from the bidirectional promoter region between the TATA binding protein and the proteasomal subunit C5-encoding house keeping genes (Neville et al., 2017). The unique characteristic of these regions is their methylation-free CpG islands which are believed to enforce a transcriptionally active state, even preventing the spread of transcriptionally repressive chromatin due to methylation and deacetylation (Antoniou et al., 2003). The most well known of the UCOE elements is from the HNRPA2b1 - CBX3 housekeeping genes and is known as the A2UCOE, which can vary in size from 1.5Kb up to 8Kb (Zhang et al., 2010). The higher mechanistic functions of UCOEs are not fully understood, though there are a few theories around this topic. One such theory is that the CpG island may confer resistance to methylation. Another is that special chromatin remodelling transcription factors such as SWI/SNF, FACT, HSF1, and histone acetyltransferases may bind and remodel the chromatin. An additional theory even states that the bidirectional or divergent transcription mechanism may confer resistance to silencing (Antoniou et al., 2003). It has also been postulated that these anti-silencing effects are due to changes in the plasmid integration profile (Betts and Dickson, 2016). These stability elements also differ from those previously mentioned in that they have shown no cell line specificity due to promoters of housekeeping genes remaining at least partially functionally active in various cell types. This anti-silencing activity has been established to work with hCMV with at least a 20 fold increase in expression for EGFP and EPO using the new promoter constructs; not only was expression increased, but stability was maintained for over 100 generations (Williams et al., 2005; Benton et al., 2002).

UCOEs have proven application in CHO cells. Boscolo et al. (2012) found that the addition of a 4Kb UCOE increases scFv-Fc production by 3-10 fold in CHO-S cells. This is in agreement with other studies which have seen a 1.5 fold to 4 fold greater increase

in expression (Betts et al., 2015; Nematpour et al., 2017) and a 1.5Kb fragment of the A2UCOE has been found to incur a 6.5 fold in expression when 3 A2UCOE elements were included in a vector (Saunders et al., 2015). Interestingly, the A2UCOE has been shown to increase expression levels even more when only used on a heavy chain plasmid, leaving the light chain without a UCOE (Nematpour et al., 2017). UCOEs have been shown to be promoter sensitive, with some promoters achieving higher expression levels using the RPS3 UCOE and others achieving maximal expression using the RNP UCOE (Rocha-Pizaña et al., 2017). The main advantage of UCOEs is that they have been proven to work in CHO and in comparison to all previously mentioned stability elements, have been shown to have increased effectiveness in relation to stability and transcriptional output (Saunders et al., 2015). The main downside to these stability elements is that their size, ranging from 1.5 Kb to 8 Kb, is a large increase in the size of the DNA load and could lead to higher DNA-based cytotoxicity when introduced during transfection. Luckily, several studies have looked at creating smaller regions of the A2UCOE that still confer resistance to repressive effects. Through analysis of the CpG islands, fragments as small as 455bp have been created, which can retain some, if not all, of the anti-silencing activity (Kunkiel et al., 2017; Zhang et al., 2017). Currently, UCOE is the most well known and established anti-silencing stability element. Still, as knowledge of the genome and stability screening strategies become more thorough, UCOEs may be replaced by an endogenous element, such as E77 from the CHO genome.

1.3.3.5 E77

Methods already exist to screen the CHO genome for promoter regions (Pontiller et al., 2008) and regulatory elements (Pontiller et al., 2010). Kang et al. (2016) investigated the possibility of applying these methods to find stability elements which could enhance CMV expression. A unique element called E77 was found to incur increased expression and stability. It was found to increase GFP fluorescence intensity by at least two fold and when expressing an antibody, raised the specific productivity five-fold while increasing the percentage of transgene positive cells nine fold. Not only that, but even in the absence of selection pressure, the expression level remained constant for 20 weeks. The E77 fragment is quite large at 3Kb and the mechanism behind how it works is entirely unknown; a hypothesis is that it has GATA binding sites in tandem, which are chromatin re-modellers and may give it some sort of stability. A reverse-orientated fragment of 1.5Kb, dubbed E77-t2, is thought to provide the stability of the main fragment. This technology and E77 itself suffer from a disadvantage in size compared to stability elements such as MARs elements, or the popularity of UCOEs, but finding endogenous stability elements has major advantages such as less variability in expression and the mechanism by which they act is guaranteed to work in the cell line in which they are found. The future of stability elements for cell and gene therapy, along with protein manufacturing, will likely rely more on endogenous screens for efficacious elements instead of tried and tested elements with large variability in effectiveness. Elements such as E77 may one day be a better option than UCOEs currently are for stables or S/MARs elements are for transient culture.

1.4 Synthetic Promoters for CHO Cell Engineering

Currently, the most utilised way to control transcription is using different characterised promoters. The most widely used promoter in CHO is cytomegalovirus immediate-early promoter (CMV-IE). This is due to it having a relatively small size of 600bp and driving a high level of transgenic expression in CHO, similar to the endogenous EF1-α promoter (Qin et al., 2010). Unfortunately, the CMV promoter does have some untoward characteristics, including cell cycle dependency and epigenetic silencing through methylation (Brightwell et al., 1997; Kim et al., 2011). Presently in industry, to change the transcriptional rate, the promoter is altered. For instance, removing the CMV promoter to replace it with EF1-α. But this only gives a swap of one transcriptional ratio to another. Instead, various methods exist to create promoters of different transcriptional strengths, but synthetic promoter engineering is the easiest and only way to fine-tune transcriptional control. Brown and James (2016) describe promoter engineering, its fundamental mechanism, and the key considerations. Essentially, to create a promoter of the desired activity, the promoter sequence must be designed in a way that can be predicted and simple to make. This is possible through engineering promoters based on endogenous transcription factors in the cell of interest. To engineer a promoter, it is first essential to understand the mechanism of action being manipulated.

1.4.1 Mechanism of Transcription

Transcription starts with the binding of RNA polymerase II (RNAPII) to the core promoter region. This allows the DNA to be converted to mRNA and then a functional protein. Transcription has four steps, initiation, promoter escape, elongation and termination. The promotor's primary function is to initiate transcription and allow RNA Pol II to bind to the gene sequence (Phillips, 2008). Unlike prokaryotes, eukaryotes require transcription factors to bind RNA Pol II (Struhl, 1999). These transcription factors bind to sites in the promotor's core, proximal and enhancer regions. Often the co-regulating activators or repressors bind the proximal and enhancer regions. The core promotor contains units such as the TATA box and Initiator (INR) element, which help to assemble the pre-initiation complex (Fuda et al., 2009). Figure 1.5 shows how these sites can interact.

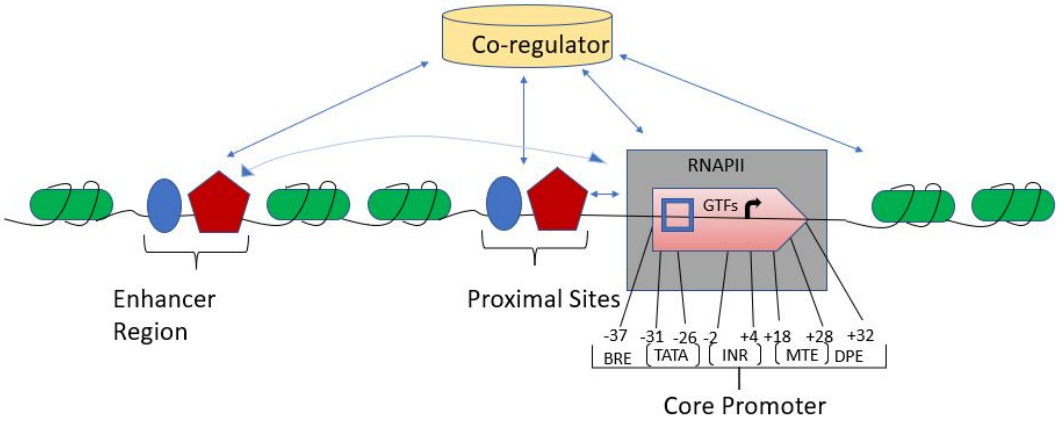


Figure 1.5: The general structure of a promoter. The core promotor element contains general transcription factor binding sites, such as the B recognition element, TATA box, Inr element, Motif ten element and downstream promotor element (Fuda et al., 2009). The transcription start site is also located in the core promotor. The red and blue shapes represent the sites where transcription factor regulatory elements (TFRE) bind to the DNA. These regulators, such as NFkB can enhance or repress transcription through interaction with the TATA binding protein or TFIID. The TFREs can also interact with the co-regulator which can then interact with the general transcription machinery or modify the chromatin structure of the DNA. Image adapted from Fuda et al. (2009).

Trans-acting elements such as the NFkB family of binding proteins have been shown to majorly affect gene transcription in the Cytomegalovirus (CMV) promotor (Brown et al., 2015). Several works have looked at altering these TFRE sites to upregulate or downregulate transcription levels within the cell. Brown et al. (2014, 2015, 2017, 2019) looked at changing the transcription levels of recombinant genes in CHO by creating the proximal and enhancer regions with randomly ligated TFREs and fusing it with the CMV IE1 core promoter. It has also been achieved through the random generation of "enhancer" regions using 10bp sequences (Schlabach et al., 2010) and similarly in Portela et al. (2017), which looked at different core promoters. However, neither of these studies was performed in CHO.

Not only have synthetic proximal and enhancer regions been generated but synthetic core promoters have also been created. Juven-Gershon et al. (2006) developed a "super-core" that showed higher transcriptional activity than the CMV core promotor (-34 to +50 relative to the TSS) that has most commonly been used in previous synthetic promotor studies (Brown et al., 2014, 2015, 2017). More recently, a "super-core promoter 3" has been developed by Even et al. (2016) and has shown increased activity over super-core promoter one and two described in Juven-Gershon et al. (2006). This study, along with the previous works on TFREs, shows that the amount of binding sites also appears to influence the transcriptional strength of the promoter Brown et al. (2014, 2015, 2017, 2019)

CpG islands represent another variable which affects the transcriptional activity and potentially how transcription works. A high CpG content has been linked with natural

bidirectional promotors found in the human genome and divergent transcription (Core et al., 2008). Divergent transcription is when a promoter produces a stable mRNA in one direction and unstable ncRNA in the opposite direction. It is essentially a bidirectional promoter that only makes one coding gene. If a promoter is found to have a high CpG content, it can also lead to increased gene silencing. Thus lower recombinant gene stability (Krinner, 2012)

1.4.1.1 Initiation

Regulation is what controls if initiation takes place and how strong the transcriptional activity is. If the TFREs create a strong attraction, then strong transcriptional activity should be seen. For this review, only RNA polymerase II (RNAP II) will be discussed as it transcribes mRNA in eukaryotes (Cooper, 2015). Figure 1.6 shows how the PIC complex is formed. Once this is achieved and in the presence of nucleoside triphosphates, strand separation occurs. This allows the C-terminal domain of RNAP II to be phosphorylated by TFIIH kinase (Hong et al., 2009). The DNA duplex then melts, forming an "open PIC". This open PIC forms the beginning of the mRNA and the RNAP II then escapes from the promoter and the TFREs (Haberle and Stark, 2018). From this point, the next step is elongation.

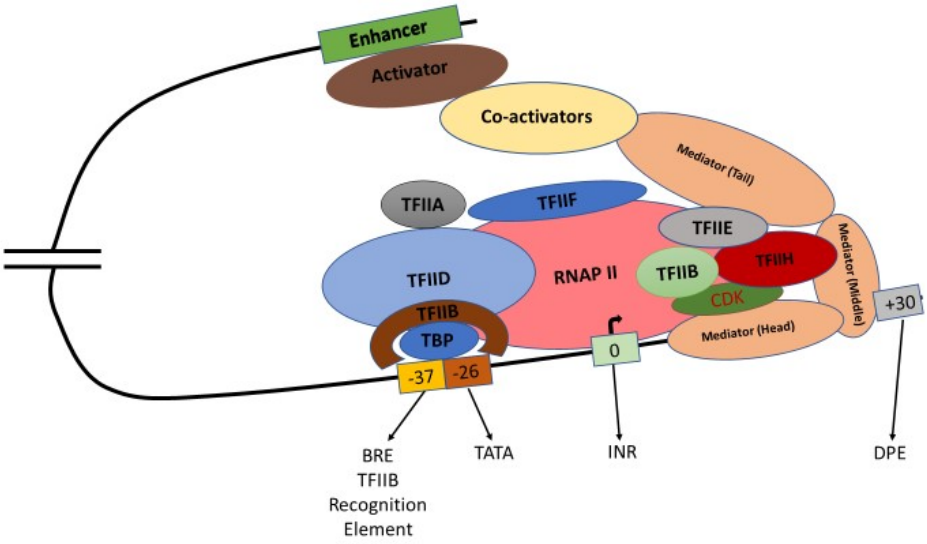


Figure 1.6: Pre-initiation complex formation. The core promotor elements shown in Figure 1.5 are bound by their respective transcription factors and the pre-initiation complex has been formed. The complex starts with TATA binding protein (TBP) binding to the TATA box, which is then followed by the binding of the general transcription factors such as TFIIB, TFIIH, TFIIF and TFIIE. The TATA box is not necessarily needed for PIC formation. The subunits can bind to other elements of the core promoter. The assembly of this complex is initiated from activators binding to the enhancers, which then recruit co-activator proteins and TFREs which can then upregulate or repress transcription. Adapted from Krishnamurthy and Hampsey (2009).

1.4.1.2 Transcription Elongation

Elongation is essentially the movement of RNAP II through the gene, creating an mRNA which can be used for translation. The PIC from the previous section has created a transcription bubble. This bubble is usually 10-12 nucleotides in size (Haberle and Stark, 2018). This transcriptional bubble runs along the length of the gene in a 3‘ to 5‘ direction. This synthesises an mRNA in the 5‘ to 3‘ directions (Griffiths et al., 2000). Throughout this stage, the mRNA begins to be processed. For instance, 5‘ capping occurs at this stage, along with 3‘ polyadenylation.

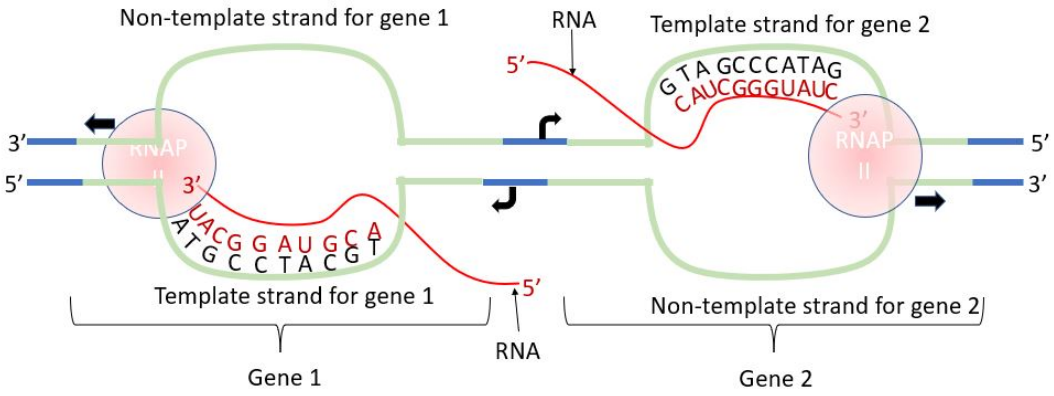


Figure 1.7: The transcription bubble. Representation of the transcription bubble moving through a gene. The RNAP II runs in the 3‘ to 5‘ direction to create an mRNA in the 5‘ to 3‘ direction. Taken from Griffiths et al. (2000).

1.4.1.3 Regulatory Factors in Elongation

Recently, it has been found that during elongation RNAP II can be paused or completely stopped while transcribing the gene. This was found by realising that Pol II levels and mRNA levels do not generally correlate in mammalian cells (Guenther et al., 2007). This breaks previous conceptions that recruiting transcription factors to create the PIC is the main rate-limiting step (Liu et al., 2015).

The pausing of RNAP II is theorised to create a structural change in the transcription complex that allows the RNAP II complex to transcribe long distances without slipping off (Core et al., 2008). The pausing of RNAP II at these sites increases the time before another RNAP II can rebind the promoter and reinitiate. It is thought that RNAP II stops at certain sequences within the gene and then requires certain transcription factors to push it off the pause site (Core et al., 2008).

One factor that has been found to be majorly involved with RNAP II pausing is positive transcription elongation factor B. This factor has been theorised to remove initiating transcription factors and substitutes them for capping enzymes, along with co-transcriptional and polyadenylation machinery (Peterlin and Price, 2006). NFkB and BRD4 have been found to be signals for Pol II release (Adelman et al., 2009; Escoubet-Lozach et al., 2011). Studies have shown that removing the pause sites can decrease the transcriptional activity of a gene as it decreases the time transcription

factors have to act on the RNAP II and can lead to improper activation (Krumm et al., 1995; Shopland et al., 1995; Fivaz et al., 2000).

1.4.1.4 Termination

Termination is currently one of the least understood stages of transcription. Due to the mechanism's unknown nature, this review won't cover it in-depth. Essentially the RNAP II will continue transcribing the genome until it reaches a terminator. There are two currently accepted models of termination. These are the allosteric and torpedo models(Rosonina et al., 2006). In the allosteric model, the RNAP II destabilisation occurs after the polyA tail is formed. The destabilization of the RNAP II is triggered by termination factors being recruited (Rosonina et al., 2006). The torpedo model differs in that when the polyA site is cleaved from RNAP II it leaves an opening for Xrn2. This enzyme then degrades the RNA from the cleavage site and snRNAs may then induce arrest of RNAP II and promote termination (Rosonina et al., 2006).

1.4.1.5 What is Transcriptional Bursting?

Transcription was once seen as a continual process, but in recent times, it has been discovered that transcription often occurs in bursts with following periods of inactivity (Chubb et al., 2006; Raj et al., 2006). Figure 1.8 illustrates how the core promoter element and enhancer affect transcriptional bursting.

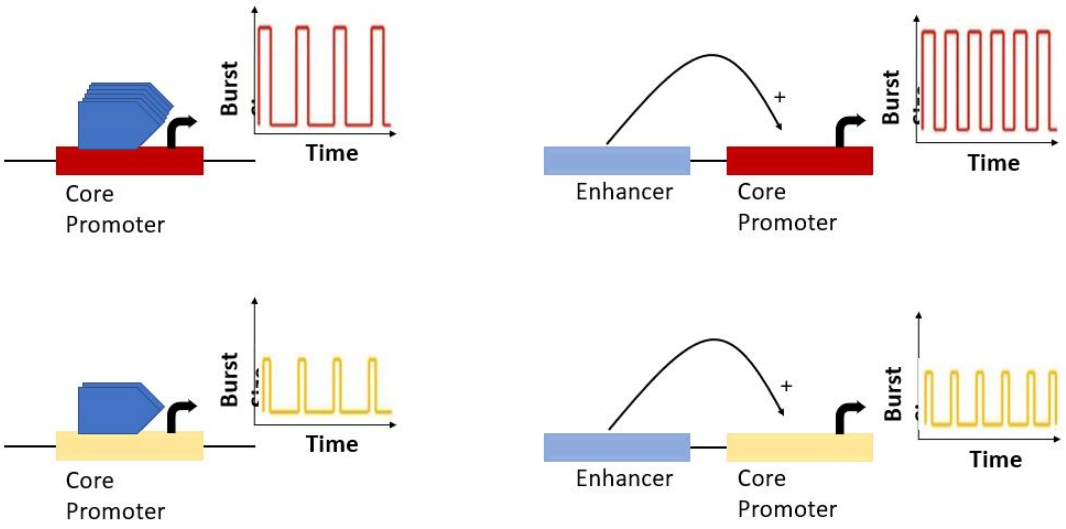


Figure 1.8: Transcriptional bursting mechanisms. The core and enhancer regions are thought to regulate transcription differently. The core promoter affects the burst size, which is the amount of RNA Pol II recruited. The enhancer region affects the frequency of RNA Pol II release.

The core-promoter has been found to be the primary regulator of transcriptional burst size (Tantale et al., 2016). A similar study using the HIV-1 gene found weakening the TATA box in the promoter caused the transcriptional activity to switch between active and inactive states faster (Miller-Jensen et al., 2013). This agreed with Tantale et al.

(2016), which found that a mutation in the TATA box could lead to long periods of inactivity.

Blake et al. (2006) found that a promoter containing the TATA box is more likely to cause transcriptional bursts than TATA-less promoters. This results from increasing transcription scaffold stability (Blake et al., 2006).

Arnold et al. (2017) investigated how different enhancers and core promoters interact. They found that enhancers have the largest effect on core promoters that contain a TATA box. This could be related to transcription factors such as NFkB promoting Pol II release (Adelman et al., 2009). Haberle and Stark (2018) theorised that this may be due to the TATA core promoter having a high burst size and the enhancer region increasing the transcriptional frequency. The effect of these two components can be additive and lead to high transcriptional output.

1.4.1.6 The Transcription Cycle

As transcriptional bursting implies, transcription is not just a one-off process. Instead, the promoter attracts more RNAP II and initiates further transcription of the recombinant gene. A possible mechanism of which is shown in Figure 1.9. The scaffold complex is formed after the RNAP II is released from the promoter (Hahn, 2004). The idea of a reinitiation complex housing TFII D, E, A, H and E was proposed in Yudkovsky et al. (2000).

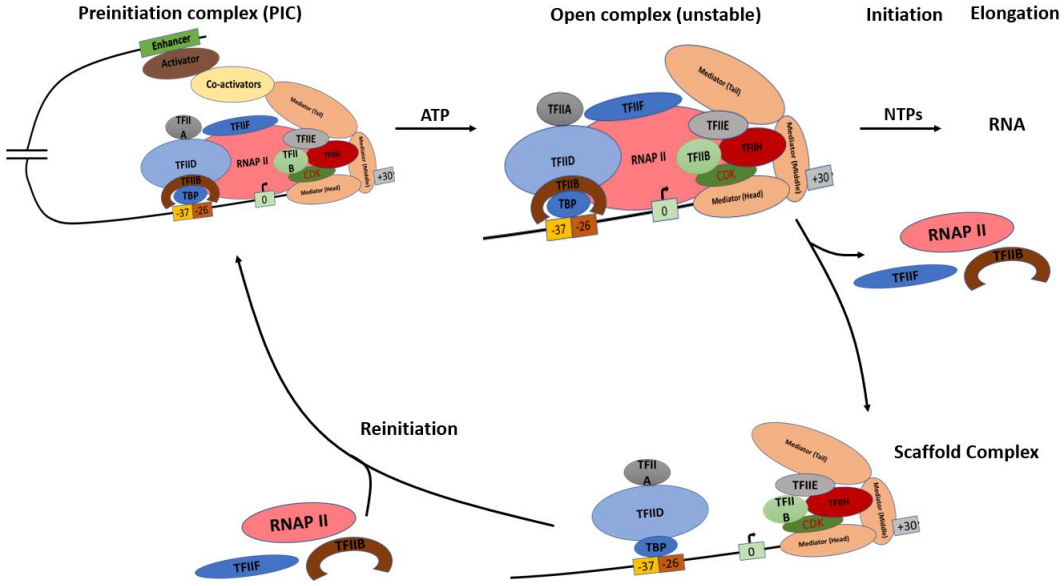


Figure 1.9: The Transcription Cycle. The idea of a scaffold complex would allow transcription reinitiation to be faster than the initial first cycle of transcription. Adapted from Hahn (2004).

The rates of reinitiation have been found to be quicker than initiation and the initial transcriptional activity may be slower than the subsequent transcriptional cycles that follow (Jiang and Gralla, 1993). The same study also theorised that reinitiation continually occurs and the efficiency of the transcription slows as they had only 4 times

as much RNA made by 30 minutes in the assay compared to the first round assay (Jiang and Gralla, 1993). This process of reinitiation outperforming initiation could be due to the bypassing of rate-limiting steps, such as when TFII D binds (Liu et al., 2014).

This mechanism of reinitiation and if it occurs is still controversial as the mechanisms of how it occurs are unclear, with many arguing if it is a mechanism that occurs at all. Other studies have shown that it is possible to increase the stability of this promoter intermediate. Yudkovsky et al. (2000) found that the activator Gal4-VP16 increased the rate of transcriptional activity by 10-fold over using no activator and 3-fold over Gal4-AH

The TATA box has also been implicated in increasing the stability of the reinitiation complex. Gralla (1997) found that the inclusion of the consensus TATA box in promoters increased the rate of re-initiation of transcriptional activity. The study found causing point mutations in the TATA sequence lowered the reinitiation rate but still increased the rate over TATA-less promoters.

Investigations into reinitiation have only recently restarted due to major advancements in experimental technologies. More recently, re-initiation has been linked with gene looping. This is essentially a process by which the promoter and terminator interact to recycle the RNAP II. This, if optimal, should increase the re-initiation rate within the cell. The mechanism by which this could happen is shown in Figure 1.10. There is some disagreement as to how this may occur. Christova and Oelgeschläger (2002) reports the presence of TFII B on the active gene promoter during mitosis. Conversely, Yudkovsky et al. (2000) reported that TFII B is not associated with the scaffold complex at the promoter. Alternatively, Singh and Hampsey (2007) suggested a new model in which TFII B disassociates from the promoter but re-associates with RNAP II at the terminator. This RNAP II-TFII B complex and TFII F then bind the scaffold allowing the function of the re-initiation complex.

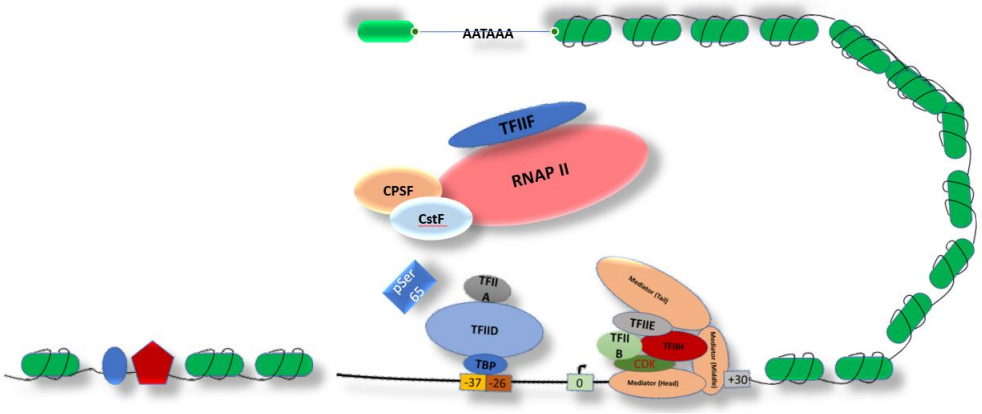


Figure 1.10: Transcription reinitiation due to gene looping. This is achieved through interaction between the promoter and terminator regions. The promoter's transcription factors, mediator and the phosphorylated TFII B interact with RNAP II, CPSF and CstF, which cause this gene looping. The terminator and promoter regions then come into contact and increase reinitiation efficiency (El Kaderi et al., 2009; Shandilya and Roberts, 2012). The figure is adapted from Shandilya and Roberts (2012).

1.4.2 Unidirectional Synthetic Promoters

The generation of unidirectional promoter libraries in CHO has been performed in several studies. Brown et al. (2014) investigated the TFREs that composed the CMV promoter and generated synthetic promoters using random ligation. NFkB and E-box were found to be major contributing factors to a promoter's overall activity. This work then investigates what regulates the transcriptional activity within the CMV promoter (Brown et al., 2015). Through scrambling the TFREs and using TFRE decoys to inhibit the binding of transcription factors, the study found that NF-kB and CRE and the largest effect on transcription activity. YY1 was found to be transcriptionally repressive as when TFRE decoys were added, the relative secreted alkaline phosphatase (SEAP) abundance increased by 150%.

Brown et al. (2017) investigated the potential to design promoters in-silico through positional insensitive, additive TFREs. This paper found the CRE TFRE to be transcriptionally repressive along with the D-box. AARE, HRE and the E-box were found to be transcriptionally inactive. Cre being found as a repressive element was interesting as in Brown et al. (2015), the element was found to positively influence SEAP production. NFkB-RE, ARE, DRE, ERSE, GC-box, X/EBP-RE and EBS1 were found to be transcriptionally active. To negate off-target effects, the study outlined the following design criteria.

- Discounted promoters containing TFREs that were inactive in heterotypic architectures (i.e. prevented unnecessary, non-functional interactions with host TFs; AARE, HRE, E-box).
- Limited the maximum copies of each TFRE per promoter to a relatively small number (5).
- Elected combinations where the copy number of the most abundant constituent-TFRE was minimized (e.g. a promoter containing one copy of four different TFREs was preferred to a construct containing two copies of two different TFREs)
- The TFREs were separated by a 6bp spacer sequence (ATTGCATCA) to limit CpG dinucleotides. (Brown et al., 2017)

The most recent study into synthetic promoters utilising CHO was conducted by Johari et al. (2019). This study generated two libraries and found the most active promoters were comprised of NFkB, GABPB and DMP1. Although it was found that these high-functioning TFRE cannot alone support high transcriptional activity. The promoter with the highest activity had 28 TFRE blocks and had a more unbiased mix compared to the promoters that had lower activity. One of the promoters with the lowest activity in library one had a heavy bias to NFkB with 6 TFRE blocks (Johari et al., 2019). Similar to Brown et al. (2017), this study also used USF1 (E-box) and Sp1 (GC-box) within the design space for the promoters. This paper also developed simple to follow models for each of the libraries. The formulas for libraries one and two are shown in Equation 1 and Equation 2, respectively.

Equation 1: Specific productivity model from library 1

$$qP = 0.93(NFkB) + 0.90(GABPB) + 0.59(DMP1)$$
$$+0.23(AhR/ARNT) + 0.18(USF1) + 0.07(STAT3) + 1.20$$

Equation 2: Model for the library 2. This library kept DMP1 and AhR/ARNT blocks at 2 and 4 blocks consistently through all promoters.

$$qP = 1.46(NFKB) + 1.21(GABPB) + 0.41(USF1) + 0.41(STAT3) + 0.57$$

Figure 1.11 shows the fold expression change of varying homotypic promoters utilising different TFRE. Both Brown et al. (2014) and Johari et al. (2019) found NFKB to be the top TFRE to affect the overall expression change. Table 1.2 shows the TFREs used throughout the literature and if they were considered high or low expressing sequences.

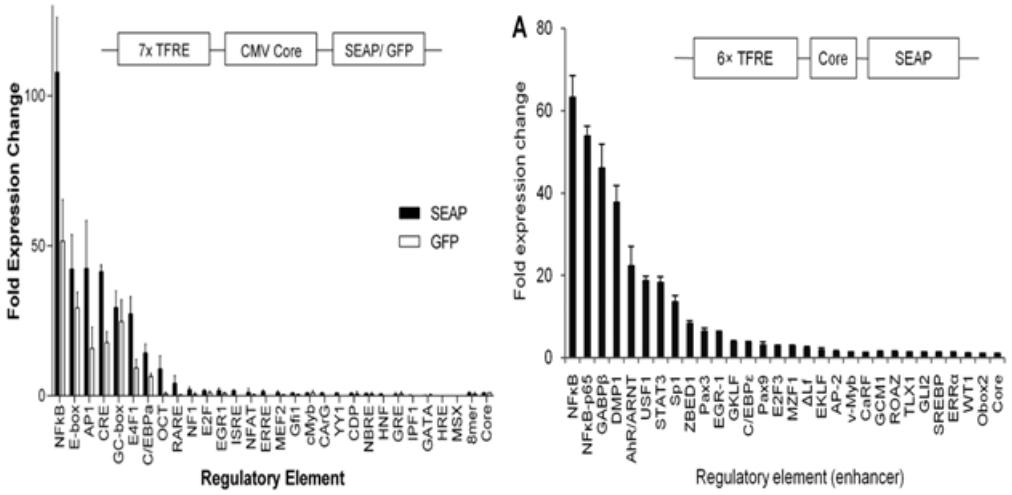


Figure 1.11: Summary of the previous literature for TFREs. Brown et al. (2014) (left) shows the expression of SEAP and GFP using a homotypic promoter. Johari et al. (2019) (right) shows the activity of other TFREs using a homotypic promoter with 6 elements. Fold change in both is relative to the transcriptional activity of minimal CMV core promoter.

Table 1.2: Table of all historically reported TFREs. The sequence and effect on transcriptional activity are also reported.

		Used in Yusuf 2019			Used in Brown et al. (2017)		Used in Brown et al. (2014)			
TFRE	**Sequence**	**Library 1**	**Library 2**	**Activity**	**Library**	**Activity**	**Library 1**	**Activity (%)**	**Library 2**	**Activity**
NFκB	TGGGACTTTCCA	Yes	Yes	High	No	-	Yes	High	Yes	High
USF1 (E-box)	CACGTG	No	No	-	Yes	Very Low	Yes	High	Yes	High
C/EBPa	CCAAT	No	No	-	No	-	Yes	low	Yes	Negative
SP1 (GC-box)	GGGGCGGGG	Yes	No	Low	Yes	Low	Yes	low	Yes	Negative
CRE	TGACGTCA	No	No	-	Yes	Negative	Yes	Negative	No	-
E4F1	GTGACGTAAC	No	No	-	No	-	Yes	Negative	No	-
ARE	ATGACACAGCAAT	No	No	-	Yes	High	No	-	No	-
NFkB-RE	GGGACTTTCC	No	No	-	Yes	High	No	-	No	-
DRE	GCTTGCGTGAGAAG	No	No	-	Yes	High	No	-	No	-
ERSE	CCAATGGCCAGCCTCCACG	No	No	-	Yes	Medium	No	-	No	-
C/EBP-RE	TTGCGCAA	No	No	-	Yes	Low	No	-	No	-
EBS1	ACCGGAAGT	No	No	-	Yes	Low	No	-	No	-
D-box	ATTATGTAAC	No	No	-	Yes	Negative	No	-	No	-
AARE	ATTGCATCA	No	No	-	Yes	Neutral	No	-	No	-
HRE	GTACGTGC	No	No	-	Yes	Neutral	No	-	No	-
GABPβ	CCCCGGAAGTGAC	Yes	Yes	High	No	-	No	-	No	-

TFRE	Sequence	Used in Yusuf 2019			Used in Brown et al. (2017)		Used in Brown et al. (2014)			
		Library 1	Library 2	Activity	Library	Activity	Library 1	Activity (%)	Library 2	Activity
DMP1	GACCCGGATGTAG	Yes	Yes	Medium	No	-	No	-	No	-
AhR/ARNT	GTTGCGTGCGAA	Yes	Yes	Medium	No	-	No	-	No	-
USF1	GGGTCACGTGG	Yes	Yes	Low	No	-	No	-	No	-
STAT3	ATTTCCCGGAAATG	Yes	Yes	Low	No	-	No	-	No	-

Overall the previous studies performed by Brown et al. (2014, 2015, 2017) and Johari et al. (2019) covered how they function, characterised them and created promoters that work better than the CMV they tested against. Johari et al. (2019) reported an increased qP under hypothermic conditions. None of the studies looked at the variation in the consensus sequences they use for TFREs, although Johari et al. (2019) did test two variants of NFkB.

1.4.3 TFRE DNA Sequence Optimization

One relatively new concept is choosing the optimum sequence for a transcription factor. Several studies have shown an increasing or decreasing affinity dependent on the sequence used for the transcription factor. Wang et al. (2018) found optimum NFkB binding sites through SELEX-seq, which had a higher affinity than the natural NFkB consensus. They then substituted these sequences into different areas of the CMV core. The relative luciferase activity (RLA) of the CMV promoter was found to change depending on the sequence used and the position of that sequence. Figure 1.12 shows the results of the study.

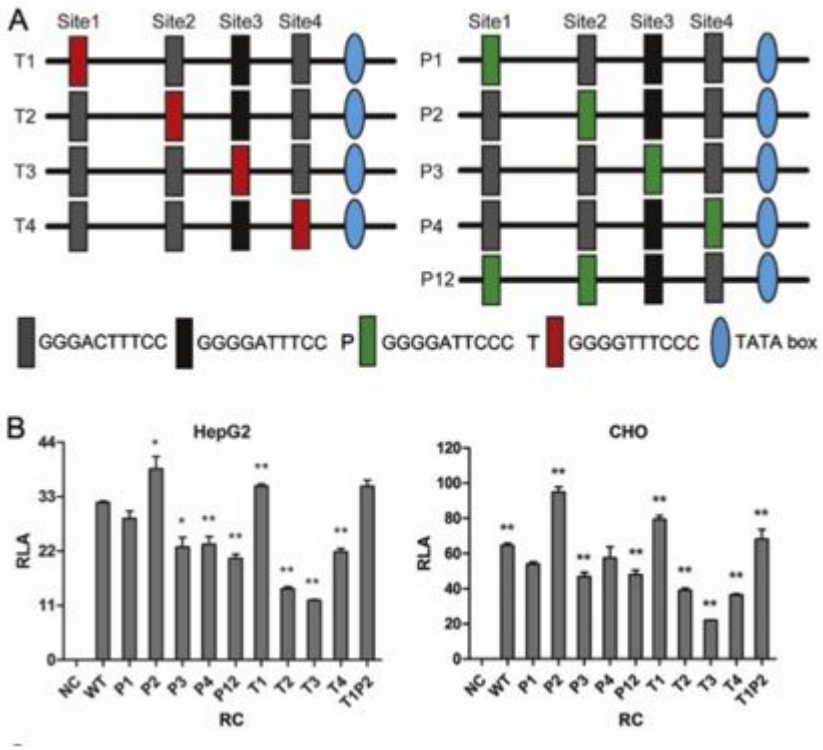


Figure 1.12: Changing the sequence of NFkB in the CMV promoter. Figure A shows the sequence changes and the position of those changes made in the CMV promoter. The sequences chosen from SELEX-seq are annotated with P(Green) and T(red). Part B shows the effect of different sequences on RLA in different positions. The graphs show sequence variation and location matter in making the largest difference in transcriptional activity. The figure is taken from Wang et al. (2018).

The NFkB DNA canonical DNA binding sequence is RGGRNNHHYYB. Wong et al. (2011a) examined how different consensus sequences may affect the binding affinity

and how NFkB may bind to non-canonical sequences. The study found non-canonical sequences to have a binding affinity that matched canonical sequences for NFkB RELA RELA (reported by z-score). The sequence of which was "AGGGGAAGTTA". This study also found that the z-score tended to increase as the sequences more closely matched the consensus, but sequences of similar differences varied in affinity. This would imply that specific changes in the sequence appear to either increase or decrease affinity.

Another study looked similarly at how transcription factor binding sites can be improved through sequence optimisation. One study looked at how Six1 binds to DNA and found specific mutations in the sequence would lead to a steep decline in affinity (Liu et al., 2012). The study showed how changing the sequence of a TFRE can increase affinity. The wild type had a Kdapp of 34.7 +/- 7.9 nM, while the new predicted binding sites had the highest affinity of 25 +/- 3.3 nM (Liu et al., 2012).

The dioxin response element (DRE) binds the Ahr/Arnt (aryl hydrocarbon receptor/Ah receptor nuclear translocator) to initiate transcription in the presence of dioxin. The CYP1A1 gene has seven DRE sequences located in the enhancer region (Liu et al., 2012). Li et al. (2014) looked at the consensus sequence of DRE, which was reported as "TNGCGTG" and "CACGCNA", where N is any nucleotide. Liu et al. (2012) found that the core sequence "TNGCGTG" initiated higher transcriptional activity than the other reported consensus sequence when induced by dioxin. The study then changed the variable nucleotide within this sequence with T, C, G and A. The study found that when Thymine or Cytosine was inserted, it improved transcriptional activity compared to when Adenine or Guanine was used.

He et al. (2015) looked at how GABPa and CREB1 can bind DNA in a cooperative manner if the ETS and CRE motifs overlap very precisely. This motif is reported as being "C/GCGGAAGTGACGTCAC". The study found that CREB1 can increase GABPa binding by approximately two-fold using the canonical sequence of "CCGGAAGT". The study then investigated different single nucleotide polymorphisms (SNPs) at the beginning and end of the ETS motif and found that binding of GABPa could be further increased by up to twenty-fold with the following mutations; "TCGGAAGT, CGGGAAGT, CCGGACGT, and CCGGAACT." Interestingly when GABPa was tested alone these SNPs lead to the lowest affinity. In this instance, none of the SNPs led to greater affinity than the canonical sequence. Still, it did show that transcription factors can work together in very precise ways to increase affinity further and decrease the effects of non-optimal DNA code.

Once a synthetic promoter library has been generated and the top TFREs selected, a further novel way to increase a promoter's transcriptional activity could be to screen the optimum DNA sequence for TF binding. This could lead to synthetic promoters being reduced in size while maintaining the cumulative affinity of the promoters with larger amounts of TFRE blocks. This could pave the way to optimised promoters with reduced TFRE blocks and hence the reduced chance of transcriptional interference between synthetic promoters.

1.5 Bidirectional Promoters

Bidirectional transcription is the same as transcription described previously but it occurs in two directions. Promoters such as these are found in nearly all eukaryotic organisms and account for 10 percent of protein coding genes (Orekhova and Rubtsov, 2013). They are thought to have a basal functional role within the cell, controlling the most basic, fundamental genes (Xu et al., 2012).

It could be beneficial for vector design due to opening a new design space that could potentially be used to avoid transcriptional interference in multigene cassettes. Transcriptional interference is competition for RNAPII between two promoters in close proximity that usually leads to one of the promoters reducing in activity (Eszterhas et al., 2002). This has been seen to occur using two of the same promoters. CMV and CMV transcriptional activity was tested together up and it was found they transcriptional interfered to reduce the activity of one promoter by up to 70%, dependent on DNA load (West, 2014).

In a bidirectional setup, the competition for the RNAPII should be less or non-existent. The shared regulatory region attracts the RNAPII and transcribes in different directions. Dependent on promoter construction different transcriptional ratios will be achieved. This could also help gene co-optimization within a vector by making the screening of the optimum ratio easier (Vogl et al., 2018) and reducing the size of the overall vector as another promoter or linker between the genes would not be required. The focus of this section will look at what makes promoters bidirectional and the factors that would be needed to create synthetic bidirectional promoters for eukaryote organisms.

1.5.1 The Structure of Bidirectional Promoters

All promoters could potentially be bidirectional (Wei et al., 2011; Andersson et al., 2015) meaning what makes a bidirectional promoter, bidirectional hasn't been clearly stated in literature yet but Xu et al. (2012) state that bidirectional promoters are genes that have their start sites less than 1kb away from each other. The distance between consecutive TSS sites is thought to impact transcription also. Literature has characterized some general features of bidirectional promoters including CpG islands (Orekhova and Rubtsov, 2013; Duttke et al., 2015), AT-rich crosslinking regions (Duttke et al., 2015) and a higher GC content of approximately 66 percent while unidirectional promotors have a lower average GC content (Orekhova and Rubtsov, 2013).

A new concept is coming forth in relation to bidirectional promoters called pervasive transcription. As mentioned previously, promoters in the human genome are intrinsically bidirectional, but this doesn't mean they code for proteins (Core et al., 2008). It is thought that when RNAPII binds to the promoter, it has an equal opportunity to initiate transcription in either direction. Still, it has been found in yeast that nascent RNA transcripts are produced far more abundantly than their divergent counterparts (Churchman and Weissman, 2011). This would indicate that it is not the binding of

RNAPII itself to the promoter which dictates its direction but regulatory sequences or signals which affect the overall direction of the promoter (Wei et al., 2011).

There are 3 ways in which a bidirectional promoter can be structured. These are overlapping, back-to-back or face-to-face setups (Johnson et al., 2018). Currently, there is no literature on the effect of these set-ups in relation to if they affect transcriptional activity. Figure 1.13 shows a potential model of bidirectional transcription. Wei et al. (2011) gives several reasons as to how the direction of the promoter could be regulated. The first potential mechanism proposed is the nucleotide composition around the promoter affecting the direction Engström et al. (2006). Another is through chromatin modification. Wei et al. (2011) describes this as previous rounds of transcription marking the orientation for future rounds. This same review proposes that if a bidirectional promoter transcribes ncRNA it could affect the directionality of the promoter through chromatin remodelling. Lastly, the 3D structure of transcription may influence the direction bidirectional promoters take. The directional memory could be maintained through mechanisms such as DNA looping if there was a favoured 3' end (Tan-Wong et al., 2009).

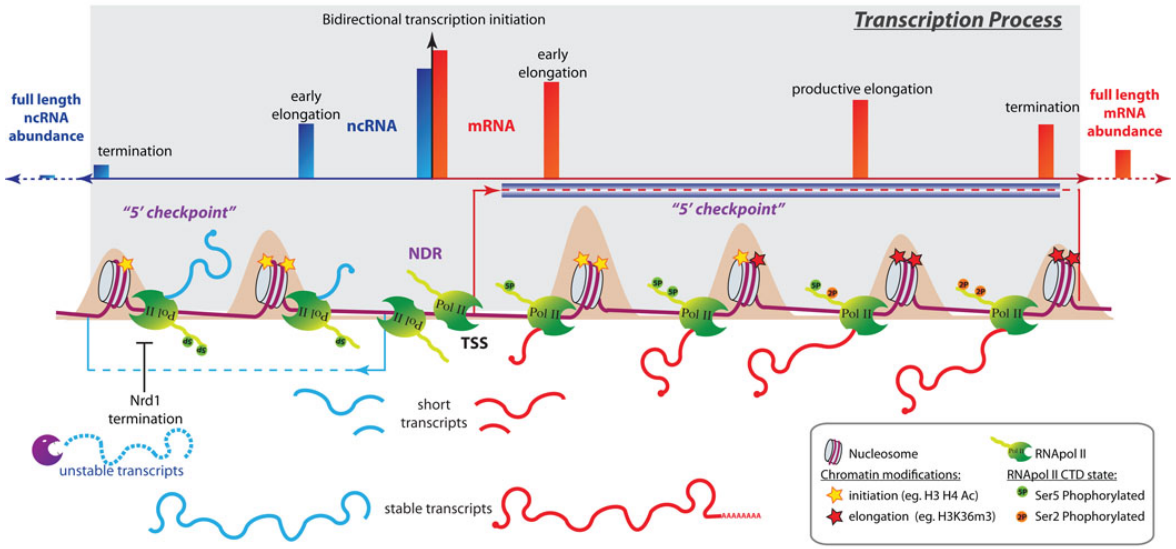


Figure 1.13: Model of pervasive transcription. The figure demonstrates a pervasive transcription model as mRNA is being produced on the right-hand side (red) and non-coding RNA is shown on the left (blue). RNAPII binds to the nucleosome depleted region (NDR) and is phosphorylated by Ser. RNAPII then reaches a checkpoint region where it pauses and can either terminate or continue to produce the functional mRNA. Termination occurs through the NRD1 pathway, leading to unstable transcripts being produced. The red and blue bars at the top of the image show the abundance of transcripts at different stages of transcription. The abundance of transcripts decreases as the RNAPII moved further along the transcript, but the abundance of mRNA in comparison to ncRNA also differs significantly. Figure is taken from Wei et al. (2011).

As previously stated, bidirectional promoters have been found to have several structural similarities but they also share many transcription factor response elements. Obayashi et al. (2012) report common transcription factors within bidirectional promoters to be "GABPA, MYC, E2F1, E2F4, NRF1, YY1, NF-Y". SP1 is also found in many

bidirectional promoters due to their high GC content (Yang and Elnitski, 2008). One major transcription factor that was found to affect bidirectionality is GABP (Collins et al., 2007). GABP has been found to be a transcription factor with high transcriptional activity in previous studies(Johari et al., 2019). The high transcriptional activity of this transcription factor could be due to it encouraging divergent transcription as well as a unidirectional promoter activity (Mikhaylichenko et al., 2018).

Collins et al. (2007) tested 121 bidirectional and 291 unidirectional promoters for GABP binding in cell lines such as Jurkat, K562, and HeLa. This was achieved through chromatin immunoprecipitation and found that GABP binds to 86.7 percent of bidirectional promoters but only 30.6 percent of unidirectional promoters. This study even introduced a single GABP site in a unidirectional library of 6 promoters and found 4 of the 6 led to significantly increased luciferase activity. The introduction of GABP inducing bidirectional transcription agrees with results obtained from Lin et al. (2007) and Patton et al. (2006), which both found GABP to be inherent to bidirectional transcription. This would make GABP one of the synthetic bidirectional promoters' most fundamental building blocks. Unfortunately, concerning the other transcription factors used in previous synthetic promoter studies (Brown et al., 2014, 2017; Johari et al., 2019) there is a distinct lack of literature on how they would affect bidirectional transcription.

Orekhova and Rubtsov (2013) mentions another transcription factor that has not been previously looked at in CHO. Anno et al. (2011) looked at 1678 genes in the human genome, of which 839 were bidirectional. What they found was that a transcription factor called hStaf/Znf143 showed a 2.4 fold increase in abundance for bidirectional versus unidirectional promoters. Anno et al. (2011) confirmed that hStaf/Znf143 was the contributing factor to promoter bidirectionality through knockdown studies and found a decrease of between 63 percent to 97 percent in mRNA produced from gene pairs when hStaf/Znf143 was knocked down.

Lastly, the transcription factor NF-Y has been linked with being involved in bidirectional transcription Zanotto et al. (2009); Bagchi and Iyer (2016). Zanotto et al. (2009) used human and mouse cells to investigate how NF-Y affects the directionality of the MRPS12/Sarsm promoter. The study deleted differing combinations of the NF-Y sites within the promoter and found deletion of the furthest 5' NF-Y led to a large increase in SARSM transcription, while deletion of the third site of four led to increased MRPS12 transcription while retaining the original SARSM transcription activity. This may agree with results obtained from Brown et al. (2015) in NF-Y was found to be a repressive transcription factor within the CMV promoter. It may downregulate the transcriptional activity in both directions. Although deletion of the furthest 5' NF-Y led to over 800 percent increase in relative expression when all NF-Y sites were deleted transcriptional activity in both directions reduced below the control Zanotto et al. (2009). This would suggest that although removal of the NF-Y sites can lead to significant increases in transcriptional activity, it may also play a small role in transcription enchantment.

Another possible mechanism from this is that deletion of the NF-Y site may push

transcription in that direction as there is less resistance for RNAPII to transcribe that gene once the site is deleted. This could have great potential for bidirectional synthetic promoters as it may allow more tunable gene expression in either direction. This may be NF-Ys functional role in nature with Häkkinen et al. (2011) reporting a conserved structure for bidirectional promoters. The study found the conservation of NF-Y binding sites in sets of four to be prevalent throughout bidirectional promoters but not unidirectional. This is the same structure shown in the MRPS12/Sarsm promoter used in the previously mentioned study (Zanotto et al., 2009).

1.5.2 Current Technologies for Synthetic Bidirectional Promoters

Bidirectional promoters have already been looked at in terms of synthetic design but not in CHO cells. Several studies have looked at synthetic bidirectional promoters in; yeast (Li et al., 2008; Partow et al., 2010; Montiel et al., 2015; Vogl et al., 2018), bacteria (Yang et al., 2013; Johnson et al., 2018) and mammalian cells such as mouse ESCs E14tg2a (Sladitschek and Neveu, 2016), neurone-derived rat pheochromocytoma PC12 cells (Liu et al., 2008), HeLa and 293T cells (Amendola et al., 2005). Currently, only one paper exists on bidirectional promoters in CHO cells Andersen et al. (2011). The study used a CHO DG44 derived cell line.

For the construction or engineering of bidirectional promoters, there are several approaches that have been carried out in literature. Vogl et al. (2018) looked at creating bidirectional promoters by screening promoters already found in Komagataella phaffii. The study used histone bidirectional promoters and mutagenesis to create a library of promoters with varying transcriptional ratios. They created inducible bidirectional promoters by incorporating parts of the MUT promoter (PDAS1-DAS2). The study then looked at the use of bidirectional promoters in optimizing co-expression of genes such as geranylgeranyl diphosphate synthase (GGPPS) and taxadiene synthase. They found the bidirectional promoters produced no taxadiene after transformation. The mechanisms behind this library could potentially be used in CHO cells as the histone bidirectional promoter architecture is conserved. Lastly, one fascinating thing from this study came in the form of bidirectional terminators. When more complex gene vectors were created, bidirectional terminators were required to prevent RNAPII collision in both directions. The inclusion of these bidirectional terminators increased fluorescence by 50-90%.

Amendola et al. (2005) were one of the few studies to look at bidirectional promoters in mammalian systems. They differed from Vogl et al. (2018) in that instead of screening the genome for natural bidirectional promoters and then engineering them, they instead created novel synthetic bidirectional promoters by fusing the human phosphoglycerate kinase promoter (PGK) and the human ubiquitin C promoter (UBIC). This was then paired with the minCMV core promoter to create a bidirectional promoter. This study showed the ability to create bidirectional promoters from two previously unidirectional promoters just with the addition of the minCMV core. This study also aimed to

create tissue-specific bidirectional promoters by replacing the PGK promoter with a tissue-specific hepatocyte specific enhancer. They found that the tissue-specific bidirectional promoter only worked in hepatocyte-derived cells and showed little to no expression in non-target cells. This, in relation to synthetic bidirectional promoters for industrial use, is not very applicable but could be extremely useful for in vivo lab studies.

Similarly to Amendola et al. (2005), Partow et al. (2010) looked at creating a synthetic promoter through the fusion of two promoters in the yeast Saccharomyces cerevisiae. This study differed in that they found different combinations of promoters were useful for different situations. For instance, the study found the tef1 promoter to be the most active during fermentation but pHXT7 to be the best in glucose limiting conditions. The whole aim of this study was to create a version of the GAL1/GAL10 promoter that did not require inducing. Unfortunately, for mammalian cells, an inducible bidirectional promoter has not been found in the genome. Instead, novel designs such as a tet-off bidirectional system have been designed that could be far more useful to an industrial setting (Unsinger et al., 2001). Still, more research is needed utilising different inducible systems to make it truly industrial relevant.

Several other studies have looked at creating synthetic bidirectional promoters by fusing two unidirectional promoters to create a new synthetic promoter (Liu et al., 2008; Sladitschek and Neveu, 2016; Johnson et al., 2018). One paper focusing on bacteria created bidirectional promoters through point mutations of a unidirectional promoter (Yang et al., 2013). The only study using CHO fused two CMV core promoters onto a single CMV enhancer region (Andersen et al., 2011). It was found that this unique promoter design created a bidirectional promoter. This was the first study to show that the CMV enhancer region could be used to create a bidirectional promoter with the addition of two CMV core regions. The study looked at 3 designs; 2xCMV TCMV-D and TCMV-C in both transient and stably transfected cells. Although the study found the CMV enhancer could be used for bidirectionality the expression was reduced compared to a CMV unidirectional promoter. The relative SEAP expression was reduced by 60 percent (Andersen et al., 2011). Another significant finding from this study was the disparity between results obtained in transient and stable expression. The study found that the best versions of the bidirectional promoter in transient and stable were completely different, with 2xCMV being one of the best promoters in transient but the worst in stable. The study hypothesised that this could be due to "orientation-dependent transcription interference of the CMV enhancer by chromatin once integrated". This agrees with findings from Huliák et al. (2012) and Seeley et al. (1997), which state, "Curiously, squelching seems to affect episomal promoters, but not those integrated into chromosomes, indicating different requirements for limiting elements of the transcriptional machinery" (Huliák et al., 2012).

The literature on bidirectional promoters is lacking in key areas of unique designs. Most studies in eukaryotes have only looked at using promoters from nature and fusing them (Liu et al., 2008; Patton et al., 2006; Sladitschek and Neveu, 2016; Johnson et al.,

2018; Amendola et al., 2005; Unsinger et al., 2001) or adding a new core promoter to the other end of the gene (Andersen et al., 2011). No entirely synthetic promoters have been created using blocks of TFREs as described for unidirectional promoters (Brown et al., 2014; Johari et al., 2019).

Other areas currently lacking investigation are the different potential designs for the bidirectional promoters. Could GABP, NF-Y and hStaf/Znf143 be used to create synthetic promoters with transcriptional activity exceeding those of a unidirectional promoter? The effect of distance, location and direction of transcription start sites is also currently lacking literature. The benefits of truly synthetic promoters could allow even greater tunable expression and higher transcriptional output than those seen in previous studies (Andersen et al., 2011). For industry, it would also have the benefit of being patentable and if required designs for inducibility could be included in the promoter to create bidirectional inducible promoters. As Vogl et al. (2018) mentioned, it could also be used to create a high throughput system to optimise co-expression rapidly. This would have great use for screening the optimum heavy and light chain ratios for antibodies within industry.

1.6 Conclusion

Literature has provided valuable insights into the current thoughts around the mechanisms of transcription. The newest theories, such as the transcription cycle and how transcriptional bursting works, give new insights into potential regulatory mechanisms that could affect the design of synthetic promoters. These insights revealed that important structures such as the TATA box and the transcription factor NFkB might play pivotal roles in increasing the transcriptional activity through promoter bursting.

The literature survey on current synthetic promoter platforms discovered essential rules for creating promoters and uncovered potential transcription factors that could be used to develop synthetic promoters that have already been tested in different CHO cell lines. These papers and the transcription factors they used led to the realisation that the same transcription factor can bind to different consensus sequences. This realisation has opened a new area for investigation as Wang et al. (2018) suggests the different sequences could be a further way to increase transcriptional output. This could be a side effect of transcription factor affinity for the other sequences (Liu et al., 2012). To investigate this, different consensus sequences for different transcription factor families such as NFkB will be investigated, and their transcriptional output will be measured through secreted alkaline phosphatase (SEAP).

One of the most exciting areas this review covers is synthetic bidirectional promoters. What bidirectional transcription is, how it works, and the conserved structure of bidirectional promoters was mentioned throughout this review. It has uncovered potential additional design rules for synthetic bidirectional promoters, such as; higher GC content, the inclusion of the transcription factors GABP, NF-Y and hStaf/Znf143, changes in the distance, location and orientation of the transcription start site and addition of core promoters on both sides of the enhancer region. The review of current synthetic bidirectional promoters revealed it is possible to create bidirectional promoters from unidirectional ones. The literature also shows it is possible to create a tuneable bidirectional promoter library and reveals the usefulness of the promoters in the co-optimization of genes and reducing transcriptional interference. The design space for synthetic bidirectional promoters is lacking compared to unidirectional design as the construction of bidirectional promoters just focuses on the fusion of two already established promoters. To investigate the structural characteristics of bidirectional promoters libraries will be created with various TFRE combinations including the transcription factors linked to bidirectionality. The library will use the top transcription factors from homotypic testing and literature informed design.

One central area of study could be applying previously proven synthetic promoter design to bidirectional promoters to increase the design space in this area further. In the future, more studies must focus on the difference in synthetic promoter design for transient versus stable production. Several studies have shown that the results can vary significantly once the synthetic promoters are introduced into a stable platform. Johari et al. (2019) showed a maximum transcriptional output of only 1.4-fold higher

than CMV in stable versus two-fold in transient.

Chapter 2

Materials and Methods

Overview

- All of the methodology used to conduct the research in subsequent chapters is contained in this chapter.
- All work was carried out in either B62, a mammalian cell culture lab, or B65, a molecular biology lab.
- Each piece of equipment will have the associated catalogue number mentioned to ensure reproducibility.

2.1 Cell Culture

2.1.1 Cell Lines

The CHO cell lines used in this study include those used at the industrial partner site (Merck Serono, 1809 Fenil-sur-Corsier, Switzerland).

- Mock-GS: A proprietary clonal cell line that has undergone MSX selection but contains no recombinant antibody.
- Clone 3: A proprietary Merck clonal cell line that has undergone MSX selection and contains a recombinant antibody but has poor stability when used in production.
- Clone 9: A proprietary Merck clonal cell line that has undergone MSX selection, contains a recombinant antibody, and has excellent production characteristics.
- CHOS: An in-house CHO cell line was used for all transient transfections. Due to IP conflicts with Merck, cell lines could not be used in this project outside of the Merck facility. All transient transfections have been performed in this CHOS cell line.

2.1.2 Cell Revival

Cells were removed from liquid nitrogen and thawed in the water bath at 37°C. These cells are then diluted into 10 ml of CD-CHO media (Gibco™ 10743029) and spun down at 200 rcf for 5 minutes. The supernatant is then decanted and the cells are resuspended in 10mL of fresh CD-CHO. Cell concentration is then measured using a Vi-Cell (Beckman Coulter) and seeded at 0.3×10^6 in CD-CHO with supplementation of 8mM of L-Glutamine (ThermoFisher 25030081).

2.1.3 Cell Passaging

Cells were measured using the Vi-Cell(Beckman Coulter) and cells were taken and placed in fresh pre-warmed CD-CHO (Gibco™ 10743029). The cells were seeded at a concentration of 0.2×10^6 if grown for 3 days and a density of 0.1×10^6 if grown for 4 days.

2.1.4 Cell Freeze down

Once cells had reached a minimum passage number of 4 the concentration was measured using the Vi-Cell. A chosen volume was decided on based on the number of cryovials needed. The calculation for how many cells are needed based on the cryovial number is shown in Equation 2.1.

Equation 2.1: Calculating the amount of cells necessary for the number of cryovials wanted

$$Volume\,of\,Cells\,required\,=\frac{number\,of\,cryovials(n)\,x\,1.5x10^{7}}{Cell\,Concentration} \tag{2.1}$$

The required volume is then transferred to a 50mL falcon tube and spun down at 200g for 5 minutes. Freezing media is then prepared using a concentration of DMSO of 7.5%. To calculate the volume of freezing media required, use Equation 2.2 below.

Equation 2.2: Calculation of the amount of freezing media required

$$Freezing\,Media\,Required\,=\frac{Volume\,of\,cells\,centrifuged\,xFlask\,Cell\,Concentration}{1x10^{7}} \tag{2.2}$$

Once calculated, pour off the supernatant from the cell pellet and re-suspend in the calculated volume from Equation 2.2. This will ensure each 1.5mL of cells contains $1x10^{7}$ cells/mL. Aliquot 1.5mL of cell culture into each cryovial and place into a Mr Frosty (ThermoFisher 5100-0050) at -80°C overnight. The next day place the cryovials into the liquid nitrogen dewars for long-term storage.

2.1.5 Cell Culture and Sampling at Merck

Cells were revived and passaged until passage four following Merck's in-house procedures. The Mock, Clone 3 and Clone 9 all went through the same process and were revived simultaneously. The cells were seeded into an Ambr15 (Satorius) and grown for 13 days in an intensified fed-batch process. Cells were grown in replicates of four in case there were any contaminations.

Cells were fed up to six times daily using the Ambr15 automated systems and samples were taken every day to measure the cell concentration using a Vi-Cell XR (Beckman Coulter), dissolved oxygen (ABL90) and lactate (ABL90). The feeding schedule and amounts were calculated using Merck's proprietary formula utilising these values. The Ambr15 also continuously monitored dissolved oxygen and temperature.

Cell pellets for RNA-sequencing were taken every day starting from day 1 at a concentration of $1x10^7$ cells. The cell pellets were spun down at 200g for 5 minutes, washed with PBS, spun down again and then stored in RNA-Protect (Qiagen 76526)

2.1.6 Transfection of Cells

High throughput nucleofection was carried out using the Amaxa SG Cell Line 96-well Nucleofector kit (Lonza V4SC-3096). The cells were passaged 3 days before transfection to ensure they were in the exponential phase of growth prior to transfection.

A DNA plate for each subsequent transfection was created with 50ul of DNA + water normalised to a value of 400ng. The plate was then sealed to ensure evaporation didn't occur and frozen at -20 °C until the transfection occurred.

Technical duplicates were performed and pooled together during the nucleofection setup steps. In each transfection 3uL of DNA was mixed with 24ul of nucleofection solution, prepared as described in the protocol provided by the electroporation kit (Lonza V4SC-3096) and 3ul of water to bring the total volume to 30ul in a 96 U bottom plate (ThermoFisher 163320).

Cells were centrifuged at 200g for 5 minutes and resuspended in non-supplemented CD-CHO to achieve a cell concentration of $2.75x10^6$ cells/well. For instance, if 90 wells are required, you would need $247.5x10^6$ cells re-suspended in 1.35 mL of CD-CHO. 30 ul of this cell solution was then added to each well of the DNA plate to result in a total volume of 60ul.

From the 96 U bottom plate, 20ul was taken and placed in each well of the 96 well nucleofector plate (Lonza V4SC-3096). Each pool in the mix plate was used for two technical replicates of each condition.

The nucleofection plate was then spun down quickly to ensure all liquid was at the bottom of the wells and electroporated using the Amaxa Nucleofector 96 shuttle (Lonza) on program FF-158. Any wells that failed were noted and excluded from future analysis. Cells were then resuspended in 80ul of CD-CHO and from this, 70ul of cells were taken and placed into 24 Shallow Well Plates (SWP) (ThermoFisher 142475)

at a seeding concentration of $1.28x10^6$ cells. These plates already contained 680ul of CD-CHO, pregassed and prewarmed. The media contained 8mM L-glut also. To ensure reproducibility, the Opentrons automated liquid handler (OT-2 Opentrons) was used to automate transfections and plate seeding.

2.1.6.1 High Throughput 24 well plate transient expression

Cells after transient transfection were cultivated in 24 Shallow Well Plates (SWP) (ThermoFisher 142475) at a temperature of 37°C, 5°(v/v) CO_2, 85% humidity and shaken at 230 rpm with a throw of 25mm. Cells were seeded at a density of $1.275x10^6$ upon transfection and were cultured for 4 days at 750ul. For instance, if cells were transfected on Monday, they were harvested on Friday. This was the same process for all transfections throughout the project.

2.1.6.2 Cleaning of Nucleofection Plates

After plates had been used, they were reused, but only in the case of biological replicates. Once used, plates were rinsed out with 100ul of isopropanol three times and then washed out with type II water once to ensure no residual was left. The plates were then left to dry in a laminar flow hood under a UV Light for an hour.

2.2 Molecular Methods

2.2.1 Gene Synthesis

Throughout various points of this project gene synthesis was used to obtain promoter constructs. Initially, Genewiz, now known as Azenta were used. Along with Geneart (Life Technologies, ThermoFisher) later in the project. Any sequences which encoded a protein, such as the heavy and light chain of a recombinant antibody, had their coding DNA sequence optimised to CHO using Genearts technologies. Every construct discussed in this thesis was synthesised during the project, except for BM1 and BM2.

2.2.2 Ligation based cloning

For all constructs, which included the analysis of SEAP, a simple restriction digest, followed by ligation, was used if the construct wasn't directly synthesised into the vector. Each construct once received, was cut using KPN1-HF (NEB) and HINDIII-HF(NEB)and ligated into the pSEAP2-CMVCore plasmid which Yusuf Johari supplied. This is shown in Appendix A.1. This is a vector which only contains the minimal CMV core promoter in front of SEAP. Following the protocols provided by NEB, 1ug of DNA was digested and run on a 1% agarose gel at 100V for one hour. The gel was stained with SybrSafe (ThermoFisher S33102) and run in Tris Acetate EDTA buffer. Usually, 12ul of SybrSafe (ThermoFisher S33102) was used per 100mL of agarose.

A metal scalpel was used to excise the required bands and washed in isopropanol between each band extraction. Once the required bands had been extracted, a gel

digest was performed using a gel extraction kit (Qiagen 28706). Ligations were then performed using quick ligase (NEB M2200L) and transformed into either sub-cloning or high-efficiency NEB cells (NEB C2988J, C2987H) following the manufacturer's protocols.

2.2.3 Golden Gate Vector

For the expression of a recombinant antibody, a golden gate system was created to allow the creation of vectors with multiple promoters. The system consisted of a heavy chain (HC) promoter vector, light chain (LC) promoter vector, LC gene, HC Gene, SV40 vector and selection marker vector. The required promoters were ligated into the promoter vectors and a golden gate reaction was carried out using the NEB BSAI kit (NEB E1601L). Each vector was added to the reaction at a concentration of 75ng and the reaction was incubated using the 11-20+ inserts protocol to ensure ligation efficiency. This consisted of being incubated in a thermocycler (Applied Biosystems, Thermo Fisher Scientific) at 37°C for 5 minutes, followed by 16°C for five minutes, repeated 30 times. This is followed by incubation at 60°C for a final five minutes as per NEBs protocol.

2.2.4 Transformation of Bacteria

All bacteria were transformed in the same method. High-efficiency NEB cells (NEB C2988J, C2987H) were taken and 1-5 ul of DNA were added to the cells. Cells were incubated on ice for 30 minutes and then heat shocked at 42°C. The cells were then placed back on ice for a further five minutes. 950 ul of room temperature SOC media was added to the cells (NEB B9020) and incubated at 37°C for 1 h in a thermomixer (Eppendorf, Stevenage, UK), shaking at 900 rpm. Cells were then plated on an agar plate which contained the relevant selection marker based on the plasmid.

2.2.5 Plasmid DNA Amplification

For transfections, all DNA preps were done at midi prep scale to ensure transfection grade DNA was achieved. Midi preps were performed by taking a single colony and inoculating it into 50mL of LB broth supplemented with the relevant selection marker. Cells were incubated at 37°C at 200 rpm for between 16-18 hours. Once grown, 5 mL of culture was taken and purified using Qiagens miniprep kit (Qiagen 27106X4). Once purified, this DNA was quality checked using a nanodrop spectrophotometer 2000 (ThermoFisher). The DNA was then sequenced using appropriate primers and if correct, a midiprep was performed using the Qiagen midiprep kit (Qiagen 12945). This was done to save money and avoid the re-transformation of DNA.

2.3 Assays Used for Quantification of Results

2.3.1 Secreted Alkaline Phosphatase

SEAP assays were carried out from harvest material on day 4 of batch culture. Cells were spun down at 200g for 5 minutes and the supernatant was extracted and placed in a 96 well plate. The assay used for measuring SEAP was the Sensolyte pNNP SEAP reporter kit (Anaspec AS-71103), but instead of using the kit substrate, a cheaper alternative was used. The substrate used was pNPP (Immunochemistry Technologies 6279). Samples generally had to be diluted using the 1x assay buffer up to a 400 fold dilution to ensure the samples were within the linear range. Apart from using a different substrate, the protocol was followed per the manufacturer's instructions. The SEAP samples were placed in a 96 flat bottom plate and the assay was performed in the Opentrons liquid handler in the dark to ensure light did not interfere with the results. Absorbance was then measured at 405nm using a SpectraMax ID5 (Molecular Devices). Two technical replicates of each standard were repeated on every plate to ensure the assay worked correctly.

2.3.2 Measuring IgG Concentration

Supernatant was collected using the same method as the SEAP experiment. Cells were spun down at 200g for 5 minutes and the supernatant was taken and put into a 96 flat well plate. When IgG concentration was measured, a fast ELISA was used(RD-Biotech expRDB 3257-5:).The collected supernatant was already within the linear range without dilution.

20ul of samples was placed into the ELISA plate and 20ul of standard in technical duplicate was also added. Once the samples are added, 100ul of peroxidase conjugated anti-human IgG was added and incubated for 15 minutes at room temperature. The peroxidase was then removed from the wells and the plates were then washed with 300ul of wash solution. The plates were hit off tissue to remove the liquid and repeated another two times. Once washed, 100ul of TMB substrate is added to each well and incubated for 10 minutes at room temperature. The reaction can then be stopped using 100ul of stop solution. The absorbance was then measured at 450 nm and 620nm on the SpectraMax ID5 (Molecular Devices).

The data was then put into the analysis software Prism(Graphpad) to fit a cubic spline curve to interpolate the values.

2.3.3 DDPCR

2.3.3.1 RNA Extraction and Cell Pellet Storage

Cells were transferred to an autoclaved 1.5mL tube (Eppendorf) spun down at 200g for five minutes, media was decanted off and the cells were placed in the -80°C until RNA extraction was performed.

RNA extraction was performed using an RNaeasy mini kit (Qiagen 74106). Cells were defrosted on ice and the RNA was extracted according to the manufacturer's protocol. RNA quality was checked using a Nanodrop spectrophotometer 2000 (ThermoFisher). An acceptable ratio was having 260:230nm between 2.0 - 2.2 and a 260:280nm ratio of 2.0.

2.3.3.2 Reverse Transcription

Samples were all normalised to a value of 600ng/ul before undergoing cDNA synthesis. The QuantiTect Reverse Transcription Kit (Qiagen 205311) was used, which has a first step to eliminate gDNA and a secondary reverse transcription step. All steps followed manufacturer's recommendations except for the reverse transcription step. Samples were left to incubate for 30 minutes instead of 15 to try and increase cDNA yield.

2.3.3.3 ddPCR

Digital droplet PCR was carried out to quantify the ratio of heavy and light chain produced during transient transfection. Primers for ddPCR were designed using Integrated DNA technologies PrimerQuest tool. The heavy chain was designed with a Fam probe and the light chain was designed with HEK for multiplex ddPCR. cDNA was taken and 2ul was placed into a reaction which included 4ul of the heavy chain probe set and 4ul of the light chain probe set to ensure a concentration of 900nM of primer and 250nM of the probe. This was along with 10ul of ddPCR supermix for probes (Biorad 1863026).

Once samples had been prepared, 20ul of the ddPCR mix was transferred to DG8 cartridges (Biorad 1864008) and 70ul of droplet generation oil (Biorad 1863005) was added to each sample and a gasket (Biorad 1863009) was placed on top and put into the QX200 droplet reader (Biorad 1864003) to generate oil droplets. From there, 40ul of oil + sample were taken and placed into a 96 well PCR plate (Biorad HSL9601) and sealed using the PX1 PCR Plate Sealer (Biorad 1814000) with PCR heat seal foil (Biorad 1814040).

The ddPCR plate was then placed into a C1000 Touch™ Thermal Cycler with 96–Deep Well Reaction Module (Biorad 1851197). The PCR cycle was carried out as per the manufacturer's instructions. Through optimisation, an annealing temperature of 55 °C was chosen based on previous optimisation experiments. Samples were then placed into the QX200 droplet generator (Biorad 1864002) to be analysed. A minimum droplet count of 10,000 droplets was required for analysis.

2.4 RNA-sequencing

Cell pellets stored in RNA protect were sent to Genewiz(Known as Azenta now) for RNA extraction and sequencing. The sequencing was carried out on an Illumina Hi Seq (Illumina) with a configuration of 2 x 150bp to achieve a minimum mapping depth

of 20-30 million reads per sample. The heavy and light chain sequences were added to the CHO PICR genome obtained from Ensemble to measure the gene expression for the components of the antibody. Reads were provided in an unstranded configuration once received and went through the following pipeline on the shARC HPC cluster at the University of Sheffield:

- FastQC: Quality Control to ensure samples were of correct quality to continue analysis.
- Trimmomatic: To remove adapter contamination from the sample.
- STAR: Fast genome aligner to align the RNA-seq reads to the CHO-PICR genome.
- FeatureCounts: To annotate the aligned reads and convert them to counts.

From here, counts were produced that could be used for differential expression analysis using DESeq2 or gene expression comparison using a metric known as Transcripts per Million (TPM). This will be discussed in further detail in Chapter 3, along with some informatics analyses of the differences between cell types. The code for all analyses performed is in Appendix B.

Chapter 3

Bioinformatic Analysis of Chinese Hamster Ovary Cells

Overview

- The chapter discusses the bioinformatic survey of the differences between a producer and non-producer cell line by looking at RNA-seq data.
- The objective was to discover any conserved differences between the producer versus non-producer cell lines and discover unique characteristics the producers have.
- The collection of the RNA-Seq samples is described in Section 3.2, followed by some general informatic observations presented in Section 3.3.
- A hypothesis-driven approach was taken and KEGG pathways were used to analyse the differences between producer and non-producer on days 2, 5 and 10. These are presented in Sections 3.5, 3.6, 3.7 and 3.8.
- Lastly, discussed in Section 3.9, a novel approach was taken to study the pathways relating to the production of a recombinant antibody, looking at overall fold changes in the pathways.

3.1 Introduction

This thesis aimed to expand synthetic promoters' design space. The caveat is that even though knowledge of how to create synthetic promoters is widely available, the optimal implementation, design and stoichiometry are still unknown. The use of RNA sequencing was hypothesised as a potential avenue to avoid or at least minimise the large amount of empirical testing that currently exists.

Although the primary objective was to create synthetic promoters, the retrieval of the RNA-seq data presented a unique opportunity to decipher the genetic differences between a producing and non-producing CHO cell. The chapter aimed to look at key areas, such as

- A general bioinformatic survey of the producer and non-producer cell line.
- A hypothesis driven bioinformatic approach was taken to find the differences between a producer and non producer on day 2, 5 and 10.
- An investigation into what pathways related to protein production are up or downregulated.

3.2 Process information and Sample Collection

Sample collection was performed at Merck. The process is called an intensified fed-batch process and was seeded at 7 million cells/mL, as described in Section 2.1.5. Figure 3.1 shows the growth curve for all 3 cell lines. The Mock cell line is the only cell line which is significantly different in terms of the general trend of the growth curve.

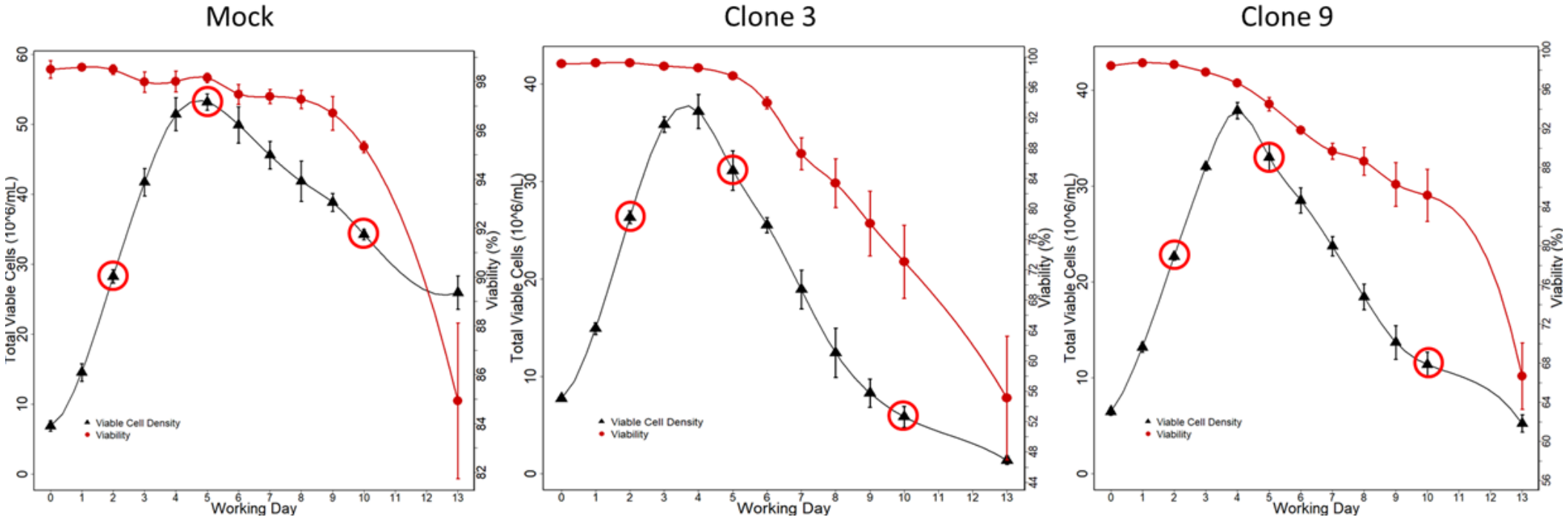


Figure 3.1: Growth curve from the Ambr15 showing the Mock growing much better than Clones 3 and 9. The growth and viability profile of the 3 cell lines tested in Merck is shown. The total viable cell density is shown on the right y-axis and the viability (%) on the left y-axis. The working day of culture is shown on the x-axis. The Mock generally grew better than Clone 3 and Clone 9. The viability of Clone 9 and Clone 3 dropped off sooner than the Mock. The red circles indicate day 2, day 5 and day 10 at which RNA-sequencing was performed. Error bars are standard deviation (sd) with 4 biological replicates.

Figure 3.1 indicates the Mock cell line producing no antibody appears to have a much higher capacity for growth, reaching a maximum cell density of approximately 51 million cells/mL, while both Clone 3 and Clone 9 failed to reach a maximum concentration of 40 million cells/mL.

Figure 3.2 shows the protein titres obtained using an IQUE (Sartorius). The Mock shows the experimental noise presented by the IQUE. Samples were taken on day 2 and then day 5 to day 13 due to the minimum volume requirements of the Ambr15. Although variable, the data fulfilled the requirement of giving a rough estimation of how much protein the cells produce on each day of culture. Unfortunately, as seen in the HPLC results discussed later, they do not correlate very well.

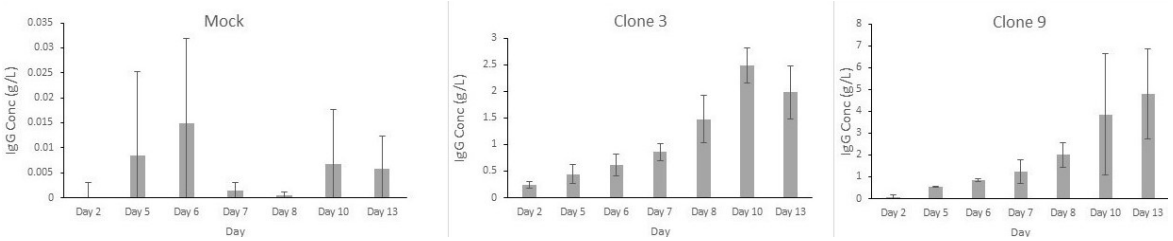


Figure 3.2: Protein titres acquired from the IQUE show considerable variability. The Mock, Clone 3 and Clone 9 are shown. The y-axis is IgG concentration measure in g/L and the x-axis is the day of culture. Error bars are standard deviation with a replicate count of n = 4.

Taking Figure 3.2 and Figure 3.1 together. It looks as though once the cells start dying on working day 5, as they start producing more protein. This could be due to the cells being more stressed. The CMV promoter drives expression and in stress, it can increase transcriptional activity (Bruening et al., 1998). It could also be due to recombinant protein not being secreted from the cell in earlier days and when they lyse due to cell death, they release the protein previously trapped within the cell.

Due to the variable nature of the IQUE measurements, final titres were measured in-house by the HPLC department in Merck. These are reported in Table 3.1. The HPLC data correlates poorly with what was observed from the IQUE assay. The Mock showed no measurable amount of IgG, while Clone 3 showed approximately 4.8 g/L and Clone 9 showed approximately 6.5 g/L. All biological replicates showed little variation between them.

Table 3.1: The final protein titres measured by HPLC on day 13 for each Clone.

Sample Name	Concentration (mg/L)
D13_Mock_R-1	<0.100
D13_Mock_R-2	<0.100
D13_Mock_R-3	<0.100
D13_Mock_R-4	<0.100
D13_Clone_3_R-1	4930
D13_Clone_3_R-2	4940
D13_Clone_3_R-3	4800
D13_Clone_3_R-4	4670
D13_Clone_9_R-1	6820
D13_Clone_9_R-2	6880
D13_Clone_9_R-3	6370
D13_Clone_9_R-4	6270

Obtaining the titres gave increased context to the RNA-seq data. It reassured that the Mock was producing no IgG and that Clones 3 and 9 produced the recombinant antibodies, albeit at different amounts. The inclusion of Clones 3 and 9 was to compare a "good" and "bad" producer.

3.3 Quality Control of Count Data

Before differential expression analysis, it was important to ensure the quality of the data. This was done through raw read analysis in FastQC. Still, more importantly, once counts had been generated from FeatureCounts, it was important to check the number of counts, along with PCA and to check if the normalisation was working. The minimum amount of reads required for differential express can be as low as 5M, but with lower reads, there is less of a chance of finding smaller changes in the clones and the statistical power of the experiment is reduced. All code for subsequent analysis is shown in Appendix B.

Figure 3.3 shows the counts for all 27 samples in the RNA-seq. The sample number correlates to different clones and different days. For instance, samples 1-9 are the Mock, samples 10-18 are Clone 3 and samples 19-27 are Clone 9. As can be seen, all samples are over 15 million reads, with the minimum sample 19, having a read count of 17,692,378. These are the annotated read counts. This means that they are the reads which aligned successfully with features from the CHO-PICR genome obtained from ENSEMBL. The graph also shows why normalisation is important. If you were to compare raw counts right now, the depth of the sequencing would have a significant impact on the results.

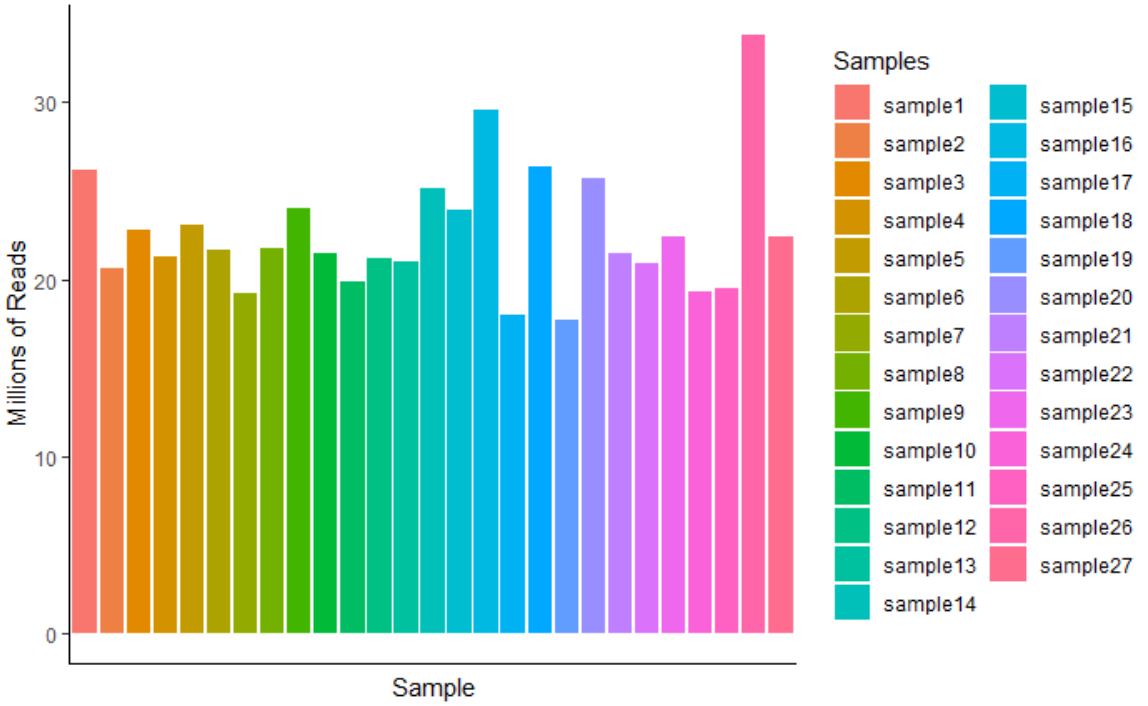


Figure 3.3: Non-normalised annotated counts from the RNA-seq. The number of reads obtained once the samples have been aligned to features in FeatureCounts is shown on the y-axis. Samples 1-9 are the Mock, samples 10-18 are Clone 3 and samples 19-27 are Clone 9. Sample names were changed to numerical values to make the analysis easier to align using loops in the bash scripting process in Linux.

Visualising this data on a boxplot with $\log_2$ counts per million, it can be seen why normalisation is required. Figure 3.4 shows the counts before they undergo normalisation to account for library size dependencies. As can be seen, the median is in different places in each sample, meaning if compared, their base level is different. DeSeq_2 always normalises counts before carrying out differential expression analysis.

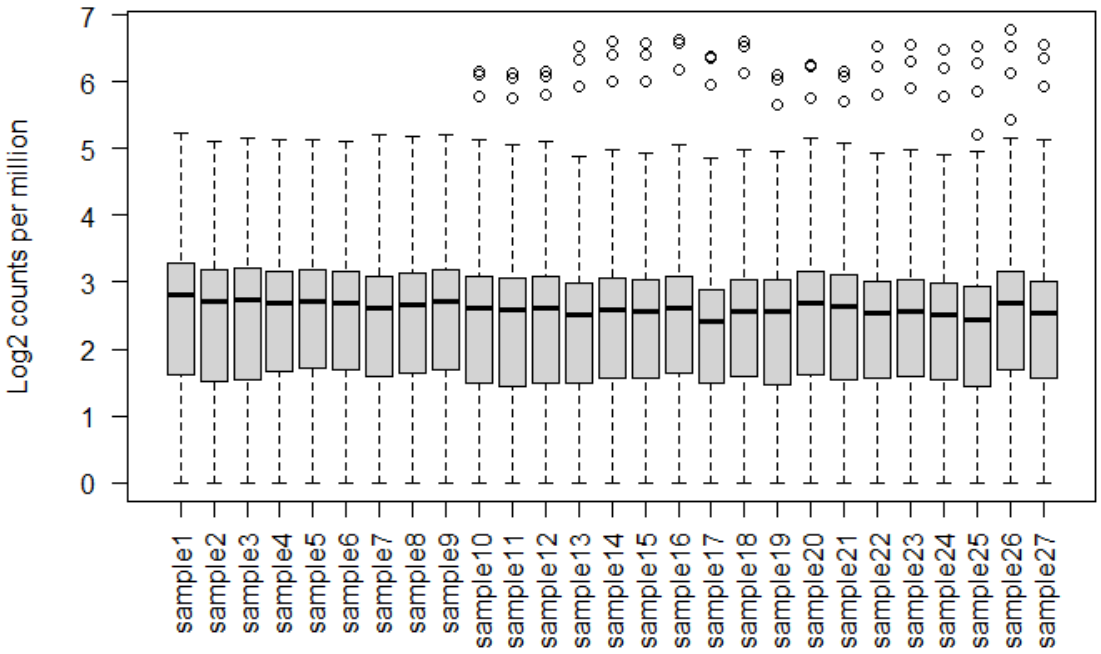


Figure 3.4: An example of the data distribution of non-normalised counts and why normalisation is needed. The y-axis shows the Log_2 counts per million and the x-axis shows the generalised sample names. The boxplot shows the median, upper quartile and lower quartile. The graph shows the median for each sample is in a different place; thus, the comparison of counts is different in each sample.

The vst function from DeSeq2 stands for variance stabilising transformation and results in the count matrix data being homoskedastic. This is just used for visualisation but ensures that when DESeq_2 normalisation is performed, it can be assumed that it has also led to each sample having comparable variance. As shown in Figure 3.5 once variance stabilising transformation has been applied, the median values are equal throughout the samples. Once the normalisation is carried out, the variance in each sample should be similar.

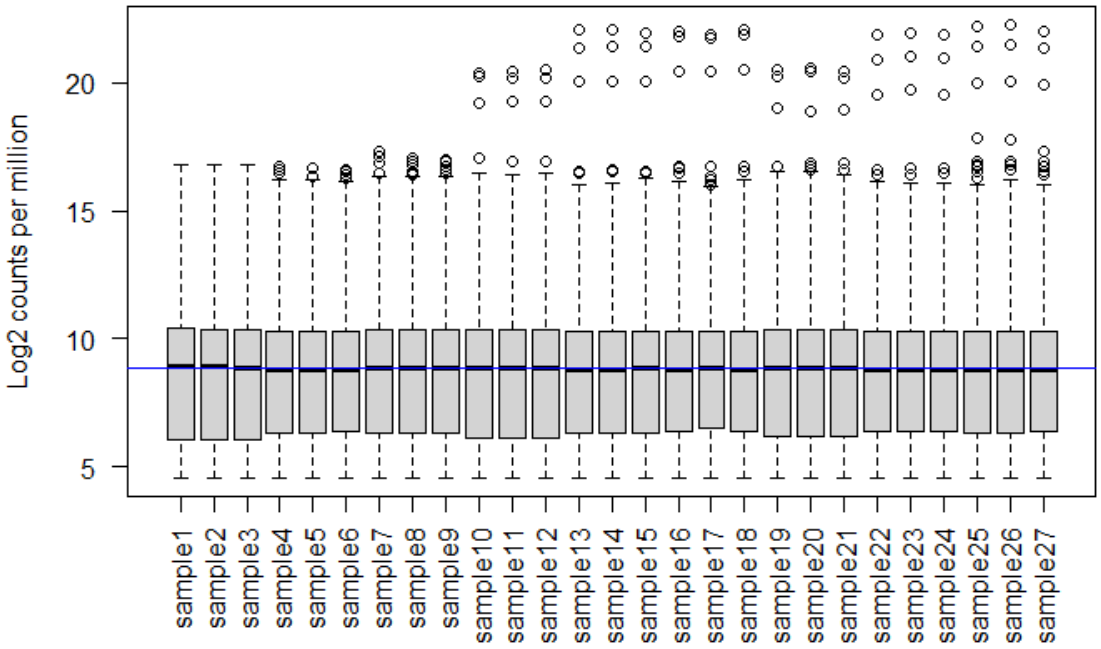


Figure 3.5: Normalised counts after undergoing vst normalisation. The boxplot shows the median, upper, and lower quartile. Compared to Figure 3.4, the median values are now much more closely aligned, as indicated by the abline.

The following important check is to ensure there was no sample mix-up. This is usually done through principal component analysis and a heat map of sample distances. Figure 3.6 shows the principal component analysis (PCA) that was performed. The grouping of biological replicates in this graph is important to visualise the sample clustering and variance between samples.

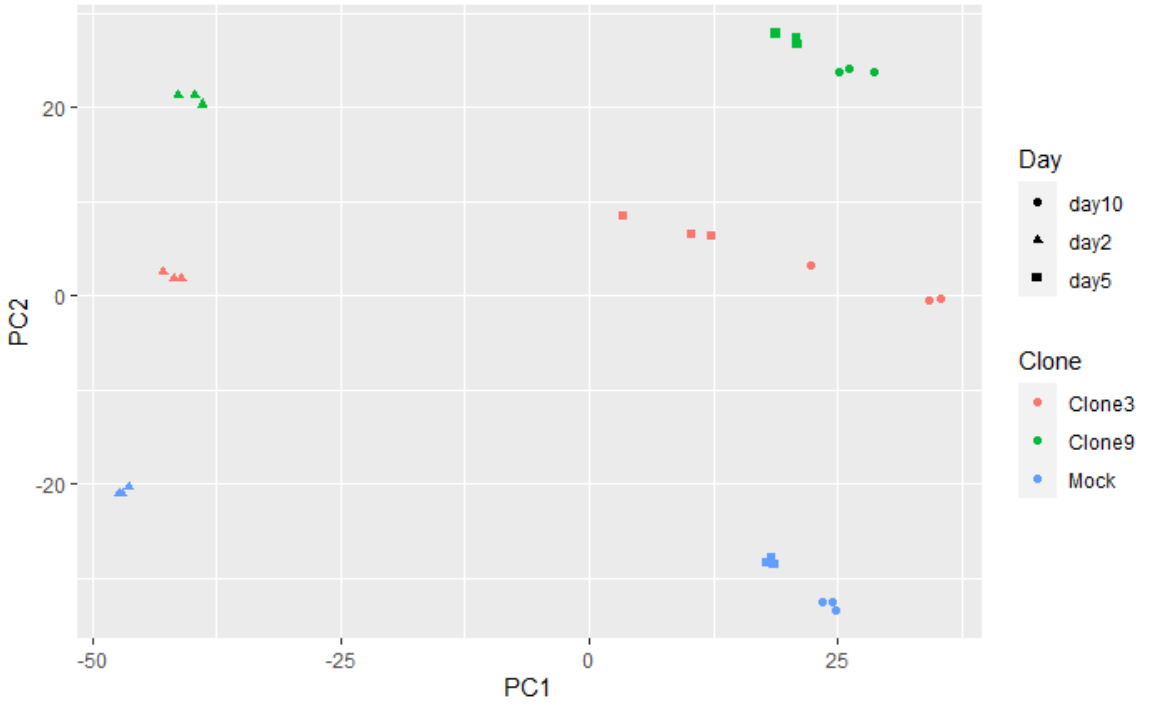


Figure 3.6: PCA plot of samples showing replicate clustering. The individual colours identify each clone and the shapes indicate the day. For instance, the Mock is blue triangles on day 2, blue squares on day 5 and blue circles on day 10. As seen from the PCA plot, all samples are clustering as expected.

From Figure 3.6 it can be seen that all samples are clustering. Clone 3, which was chosen because it is unstable when used for protein production with highly variable titres depending on the batch, appears to have the most variability on day 5 and day 10, with the data points showing more spread. This could indicate that this clonal cell line has more genetic heterogeneity or even has a higher mutation rate than the Mock and Clone 9, which could potentially lead to genetic drift.

Lastly, for quality control, the sample distances can be examined using a heatmap. Figure 3.7 shows the sample distance heatmap with a stronger blue colour indicating less variation and the lighter the blue, the more variation there is.

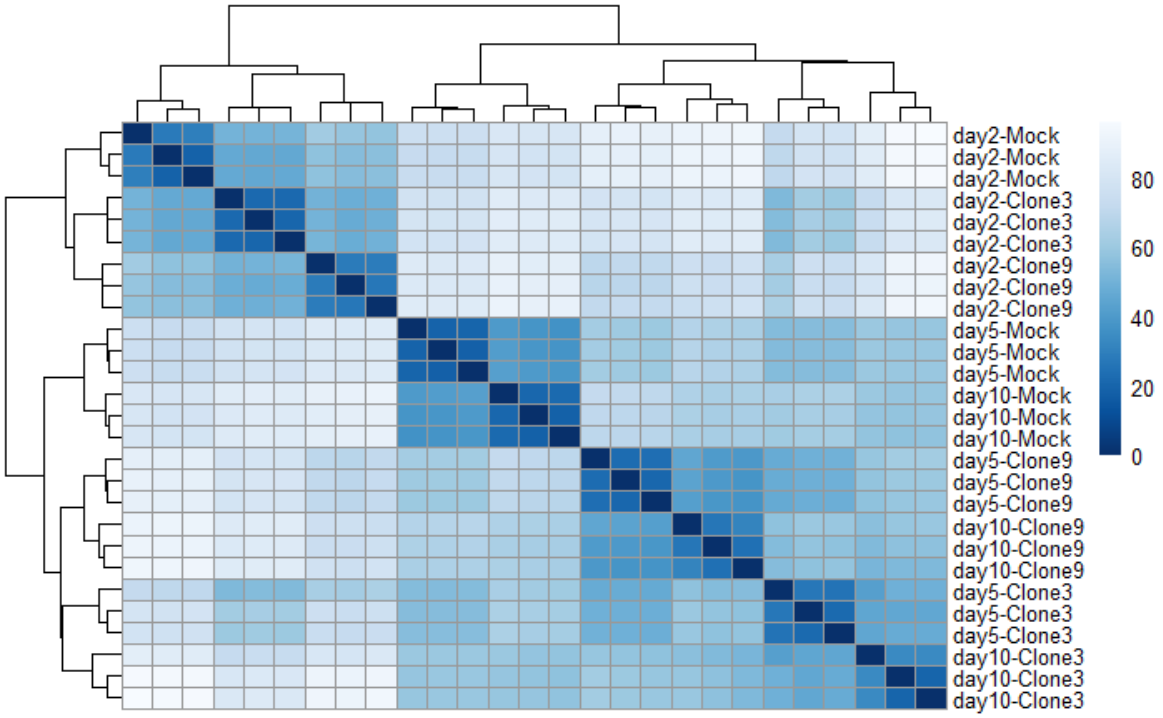


Figure 3.7: Heatmap showing the sample distances. The heatmap shows sample to sample distance (euclidean) by colour. To interpret a heatmap imagine there is also every sample name on the x-axis so the correlation of each sample on itself is zero and the box next to it is its correlation to the next sample on the list. The map was generated in R using the pheatmap function

The heatmap matrix showed all biological replicates are grouped. On day 2, it appears that the genetic variability among the samples is not very high. This is interesting as it shows that after day 2 the samples from their base genetic level change drastically. The most important observation from this data is that it correlates with the PCA plot, showing all biological replicates are grouped together as expected and none need to be removed before proceeding onto $DeSeq_2$ normalisation and differential comparisons.

3.4 General Observations

One of the most important things to do when looking at differential expression data is to get an idea of the overall differences between the conditions to be looked at. This section will overview some general data on the comparisons made between a producer and non-producer and Clone 3 versus Clone 9.

It's easy to observe the landscape of differential expression between the samples using histograms. It gives an idea of how broad the change is between the samples. A broad histogram is produced if there is a large difference between samples. If the samples are very similar, the peak will be narrow. Looking at the producers versus non-producing comparison on day 2, there are 2482 genes upregulated and 2442 genes down-regulated. While Clone 3 versus Clone 9 has 1664 upregulated and 1802 downregulated of 13636 genes. The genetic differences between the two producing clones compared to the mock was surprising as it was thought producing clones may share genetic similarities. The histograms can be seen in Figure 3.8 below. The main focus of the analysis was on a producer versus a non-producer because the variation between the two clones was so significant.

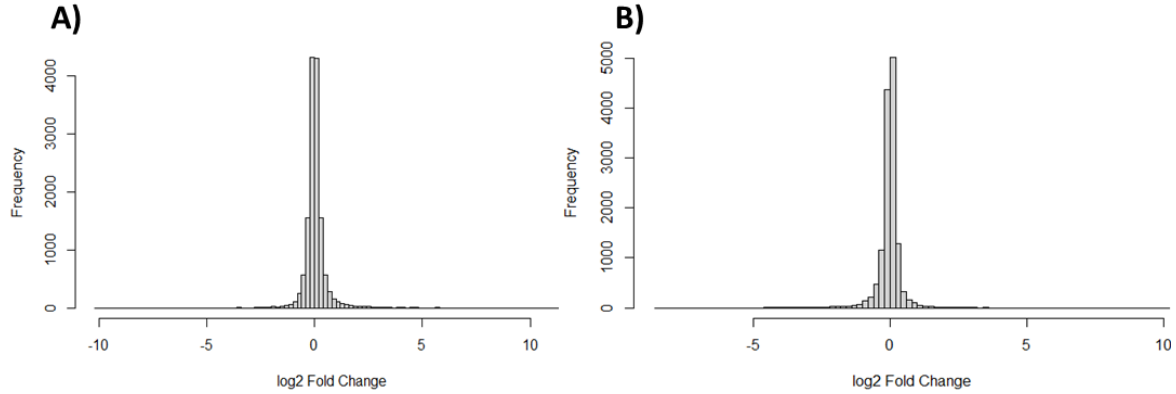


Figure 3.8: Histogram of differential expression for the producer versus non-producer and Clone 3 versus Clone 9. Part A) shows the distribution of differentially expressed genes in the comparison of producer versus non-producer. B) shows the distribution of differentially expressed genes in the comparison of Clone 3 and Clone 9. The Log_2 Fold change is on the x-axis and the frequency at which they occur is on the y-axis. Taking both figures together, it shows very little difference between comparisons A) and B). The genetic variation between the clones is high. For this reason, it was decided to take the clones into account as a group to try and decipher the differences between the producing cells and the non-producing Mock cell.

The conserved changes between Clone 3 and Clone 9 on day 2, when compared to the Mock further indicate many variations between the clones. Figure 3.9 shows the amount of conserved upregulated and downregulated genes compared to the Mock. For instance, only 33% of upregulated changes are conserved and only 36.37% of total changes are conserved when downregulated. This shows a large variation between the Clone 3 and Clone 9 samples, even on day 2 when the samples are at their closest relational distance in terms of similarity.

For this reason, looking at Clone 3 versus the Mock or Clone 9 versus the Mock was not considered; instead, only the producer versus non-producer comparison will be discussed for the pathway analysis with emphasis on looking at what is conserved between the Clone 3 line and Clone 9 line. Pooling these samples together as producers is done automatically by DeSeq2 analysis. If more cell lines were tested, it would be possible to break down the differences in clones based on productivity if similar trends were seen in other high and low producers.

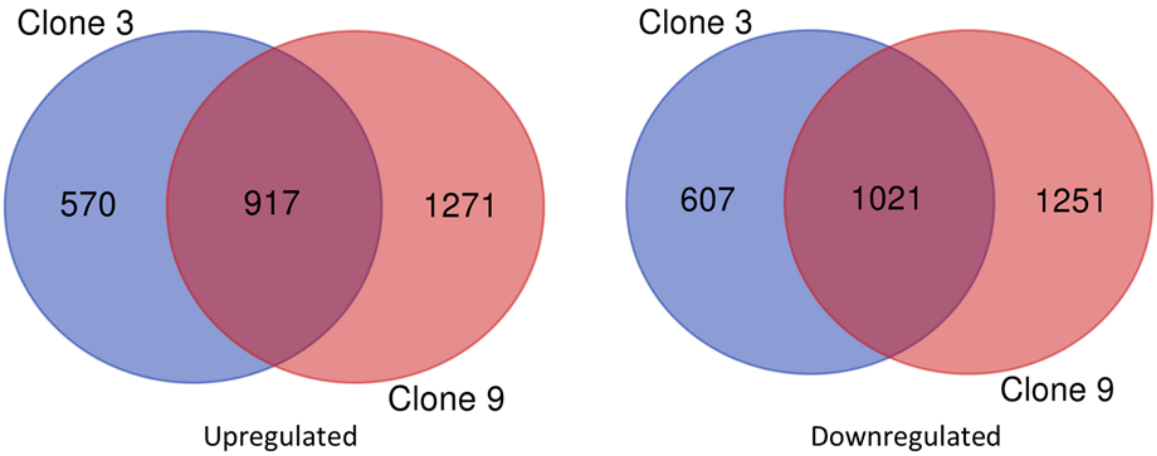


Figure 3.9: Conserved changes between Clones 3 and 9. This pie chart depicts the statistically significantly up and downregulated genes compared to the Mock on day 2 of culture. Less than 50 percent of genes are similar between the samples.

Taking the quality control PCA and the general observations of PCA into account, its no surprise that there is a lot of variation between Clone 3 and Clone 9. Figure 3.6 even shows the nearly equal variance between the Mock, Clone 3 and Clone 9 on day 2, with more significant variance occurring on day 5 and day 10 but the variance between Clone 3 and Clone 9 remains approximately constant.

3.5 Day 2 Producer versus Non-Producer

This section focuses on the differences that occur throughout culture between the producing cell lines and the non-producing cell line. By grouping the Clone 3 and Clone 9 samples, it should allow the conserved changes to be observed with respect to what is needed to produce a recombinant antibody and how producing such a metabolically demanding molecule affects the cell transcriptomically. For all KEGG pathway analyses and all analyses discussed in this section, only genes with a p_{adj} value of less than 0.05 were mapped to the pathways to ensure high stringency interpretation of results.

One of the main issues with gene ontology and RNA-seq, in general, is that with so much data, it is hard to know where and what to look at. Gene ontology can look at pathways that aren't necessarily of interest. To account for this and what is currently known about producing recombinant protein in CHO there was certain hypothesis that could be proposed, such as

- RNA encoding transcription factors between the Mock and Producer will not be significantly different as the bottlenecks of protein production occur after transcription.
- For translation: Genes linked to ER Stress, folding proteins, ribosome biogenesis, ribosome function and secretion should be increased in the producing cell lines.
- Translocation: Protein secretion should be upregulated and vesicle transport also.

These hypotheses aim to track every step of creating a recombinant protein, from transcribing the gene to protein folding and secretion. By following the pathway of a recombinant protein, one would expect changes in these pathways for a cell producing a highly complex protein versus one that is just producing a glutamine synthetase molecule for survival.

3.5.1 Day 2 Transcription

RNA encoding transcription factors showed no significant differences between a producer and a non-producer, as shown in Figure 3.10 below. The basal transcription factor gene TTDA was upregulated, but this is associated with DNA repair. RNA Pol II showed no significant difference between a producer and GS null. Spliceosome showed no significant difference in complex formation and upregulation of only CTNNBL1, which is involved with nuclear localisation and activates the deaminase enzyme.

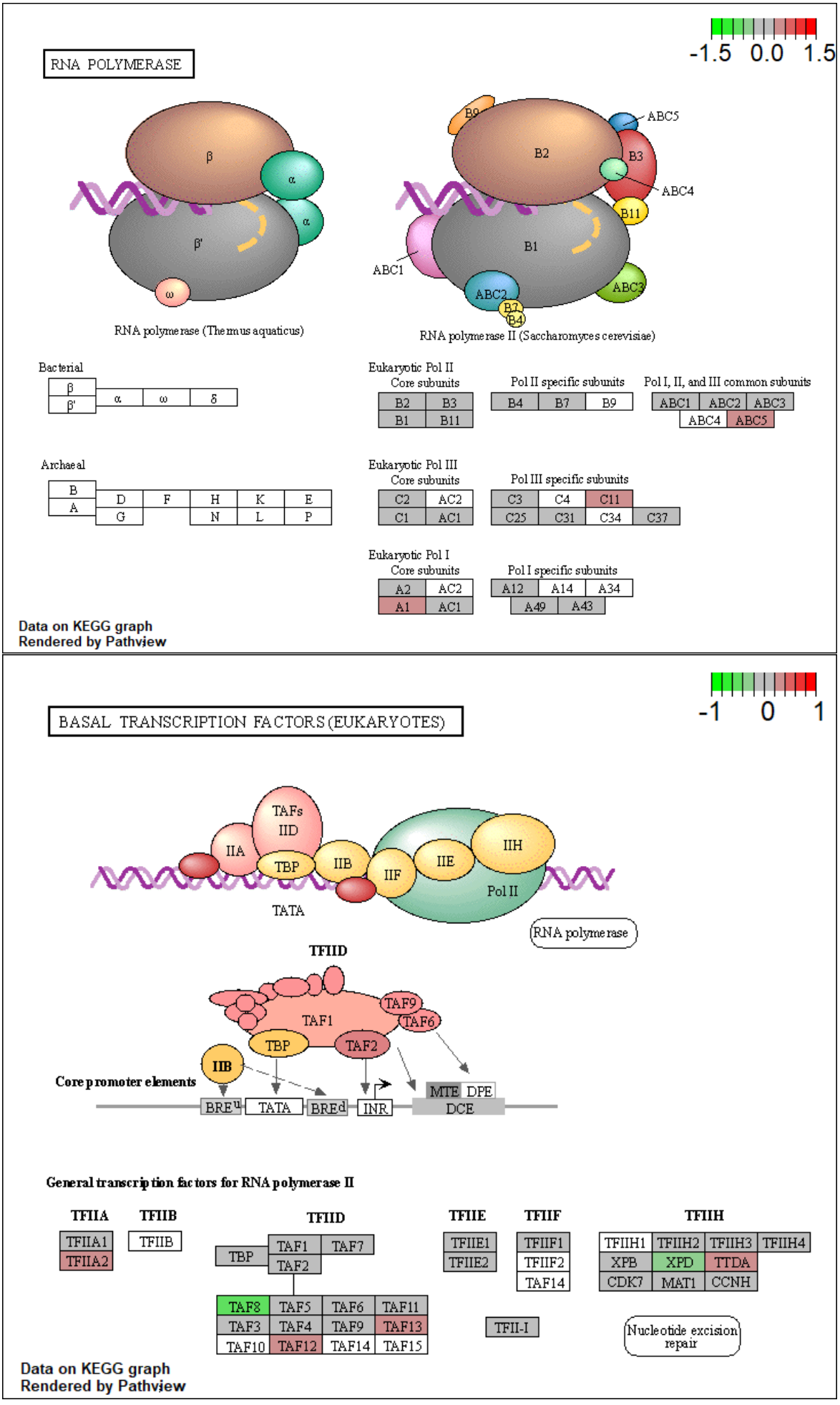


Figure 3.10: Continued on next page

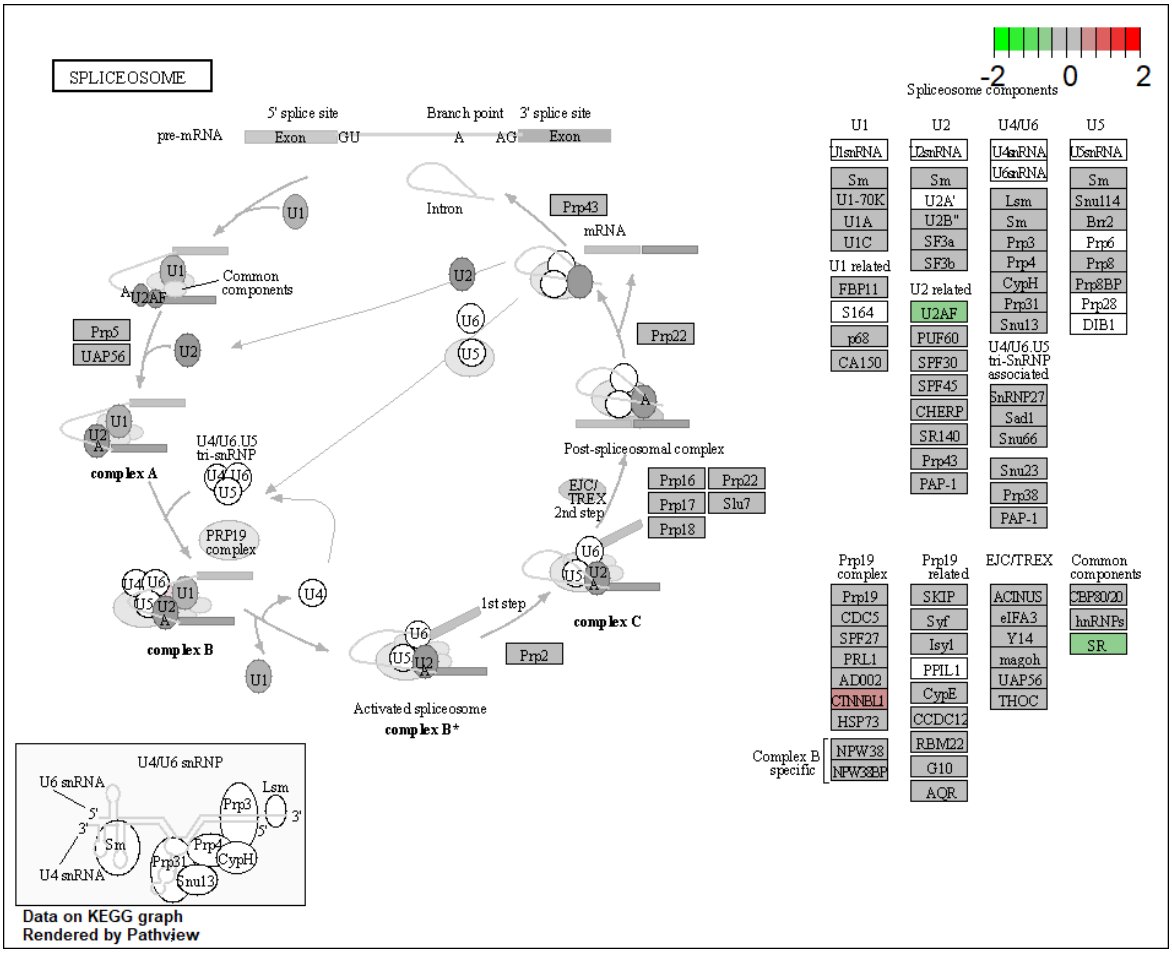


Figure 3.10: The comparisons show little change in transcriptional landscape on day 2 for the producer versus non-producer. KEGG pathway summary for transcription. As can be seen from the heat mapping, which goes from green being downregulated to red being upregulated.Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. This was also shown by none of these KEGG pathways having a p-adjusted value of less than 0.05 when pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function.

3.5.2 Day 2 Translation

In the mRNA Surveillance Pathway, *CSTF1* and *PABP* are upregulated and statistically significant with a $p_{adj} < 0.05$. *CSTF1*, 2 and 3 combine to form cleavage stimulation factor (*CSTF*). 2 and 3 are upregulated but not statistically significant. *PABP* is a polyadenylation binding protein and escorts mRNA through the nucleus; it is vital for translation initiation. The combination of *CSTF* and *PABP* indicates that in the producing protein cell lines, there might be more production of mature mRNA transcripts. Figure 3.11 shows these pathways. The only notably downregulated gene is *PP2A* and this could likely be due to the Mock cell line growing better than the producing clones. On day 2, Clones 3 and 9 have slower growth rates than the Mock.

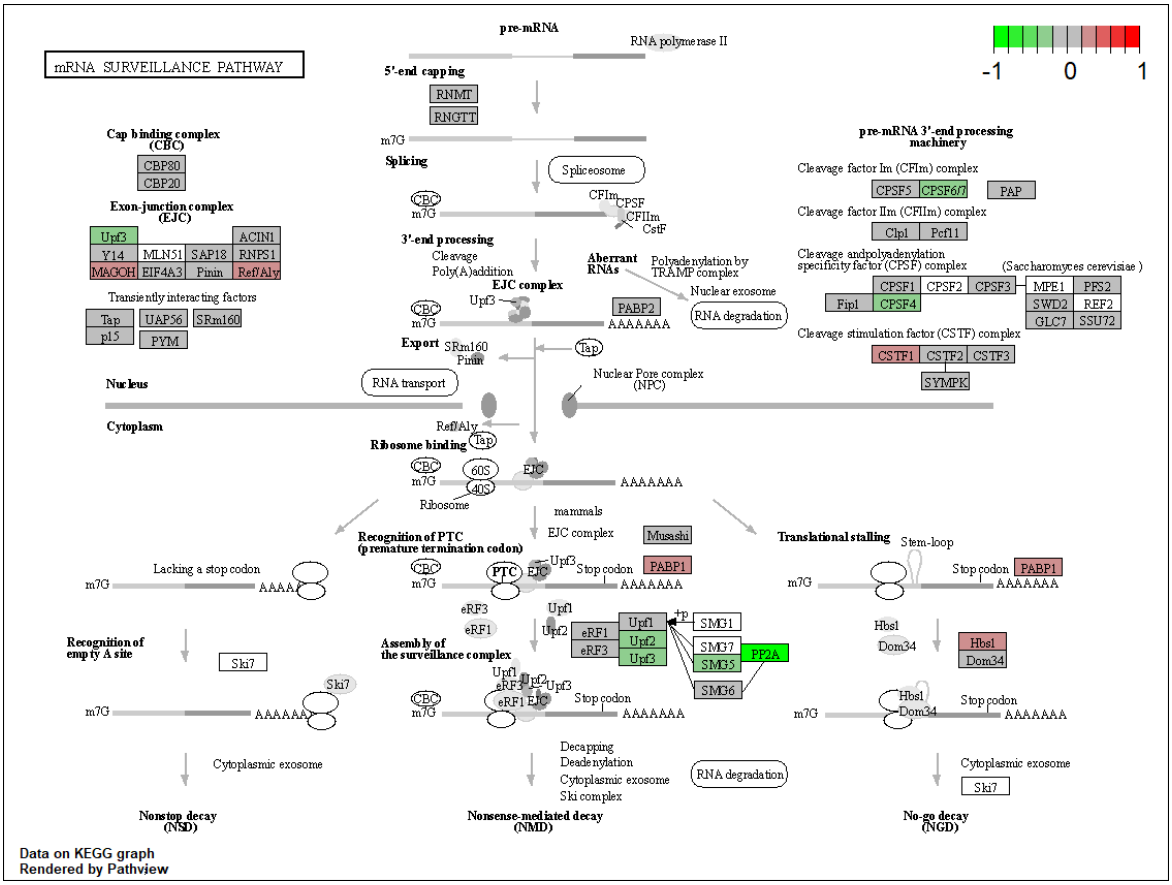


Figure 3.11: The mRNA surveillance pathway shows very little difference for the producer versus non-producer on day 2. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function.

Given that there are no major changes to the mRNA surveillance pathways, which functions to look at the quality of mRNA and the production of mRNA, it would be expected that mRNA transport is not greatly affected. Figure 3.12 shows little difference between a producer and a non-producer on day 2. The most notable gene upregulated in the RNA transport pathway is *eEF1A* which is responsible for the enzymatic delivery of aminoacyl tRNAs to the ribosome. It is statistically significant ($P_{adj} < 0.05$) and may indicate an increased need for amino acids for translation in the producing cells. This would make logical sense as the cells producing the recombinant proteins would have a greater amino acid requirement due to their production burden. Interestingly, *RAE1* is an mRNA shuttling protein and is upregulated, supporting the idea that more mRNA shuttling from the nuclease might be occurring and this may indicate that at a base level, the producing clones have more mRNA shuttling and translation occurring, but transcription is largely not what characteristics are important for high antibody-producing clones.

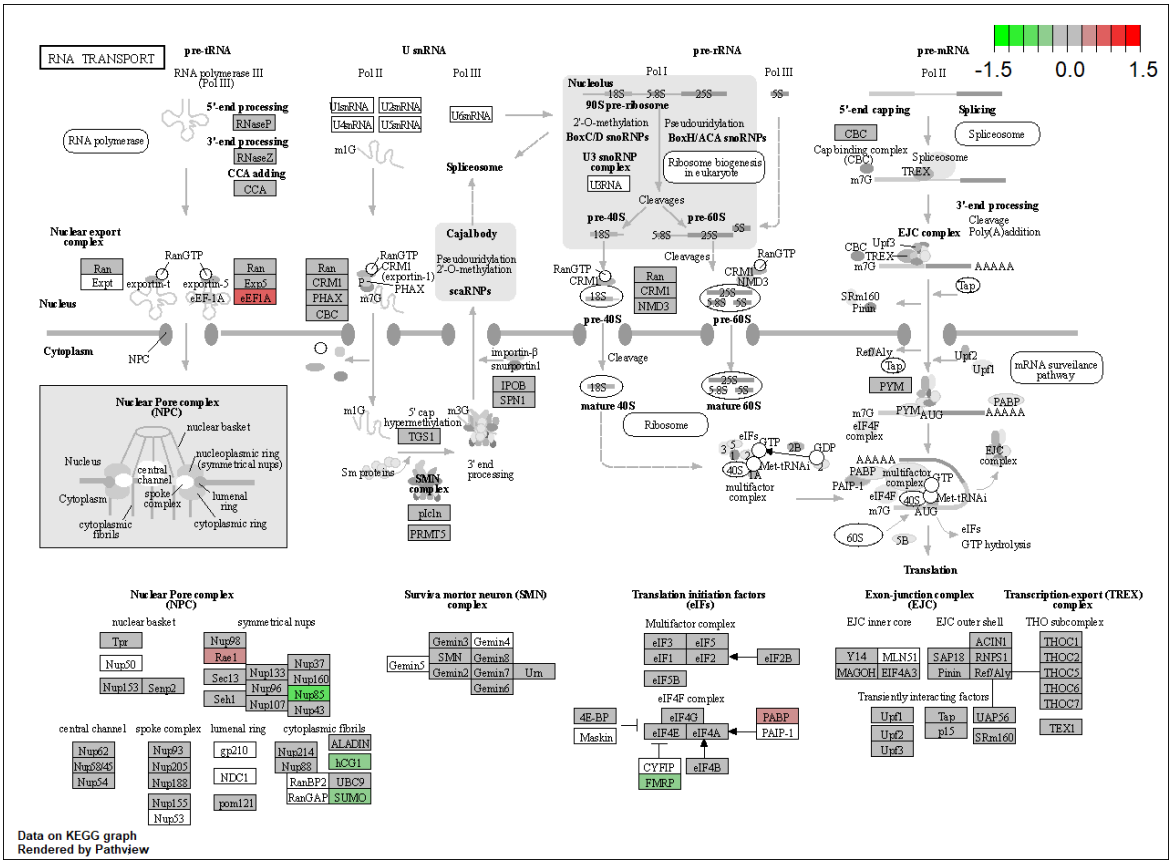


Figure 3.12: There is very little change in many pathways involved with RNA transport for the producer versus non-producer on day 2. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function.

As shown in Figure 3.13, ribosome biogenesis shows little upregulation with only 2 genes upregulated and one downregulated to any degree. Individual ribosomal proteins show some upregulation but not to the expected extent. Due to the increased metabolic burden of the recombinant protein, it was expected that the producing cells would have increased ribosome biogenesis due to more ribosomes being required for increased protein production demands. Since *eEF1A* was upregulated it was also worth looking into Aminoacyl tRNA biosynthesis which showed upregulation of L-Tryptophanyl and L-Threonyl tRNA, which could indicate increased uptake of these amino acids.

The lack of increased ribosome biogenesis, along with the upregulation of genes associated with tRNA synthesis and tRNA utilisation could indicate that the producer cells have increased shuttling capacity instead of more ribosome capacity. Glutamine selection with the antibody present could preferentially select cells with more efficient use of the resources they already have instead of selecting cells with larger capacities for translation.

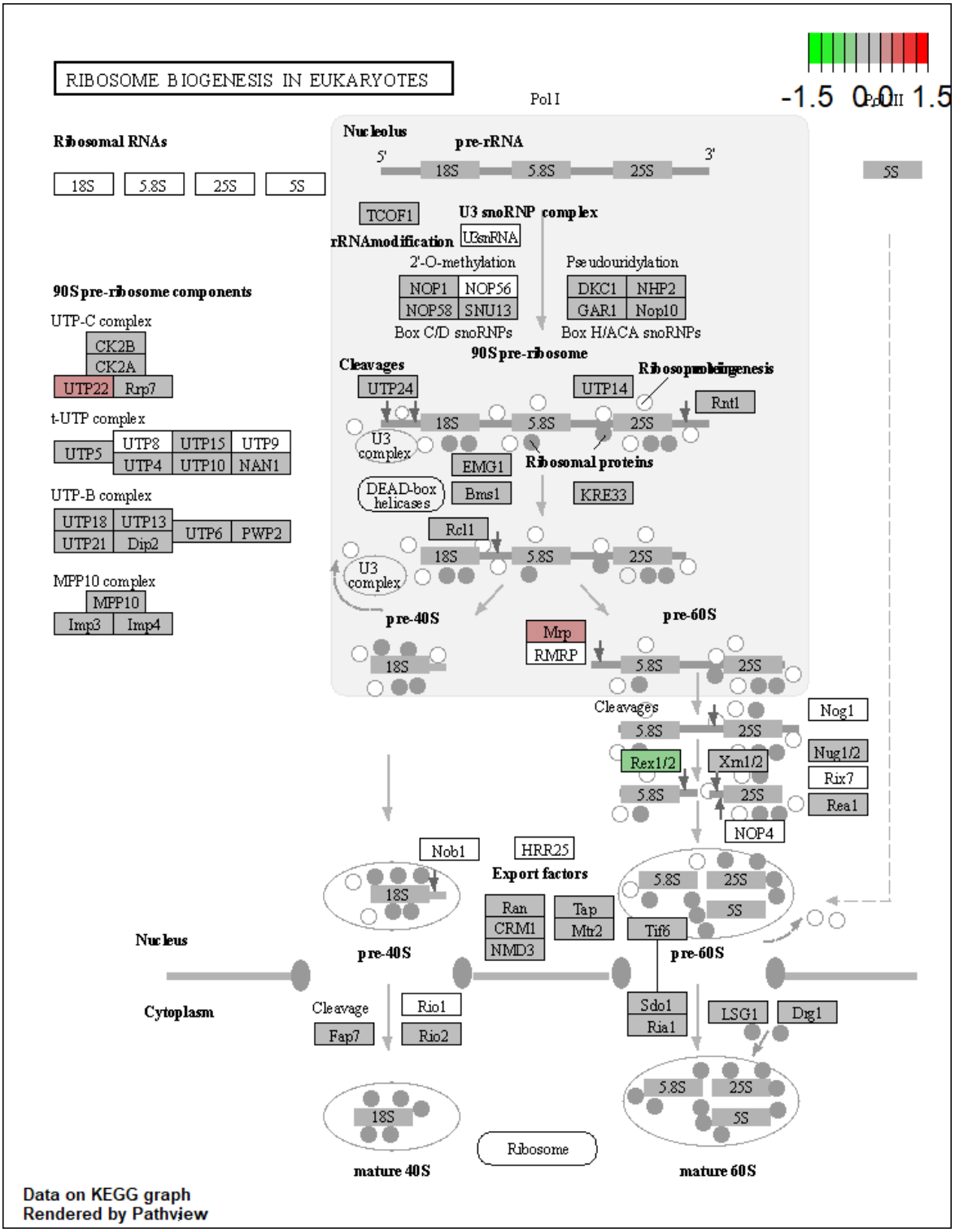


Figure 3.13: Continued on next page

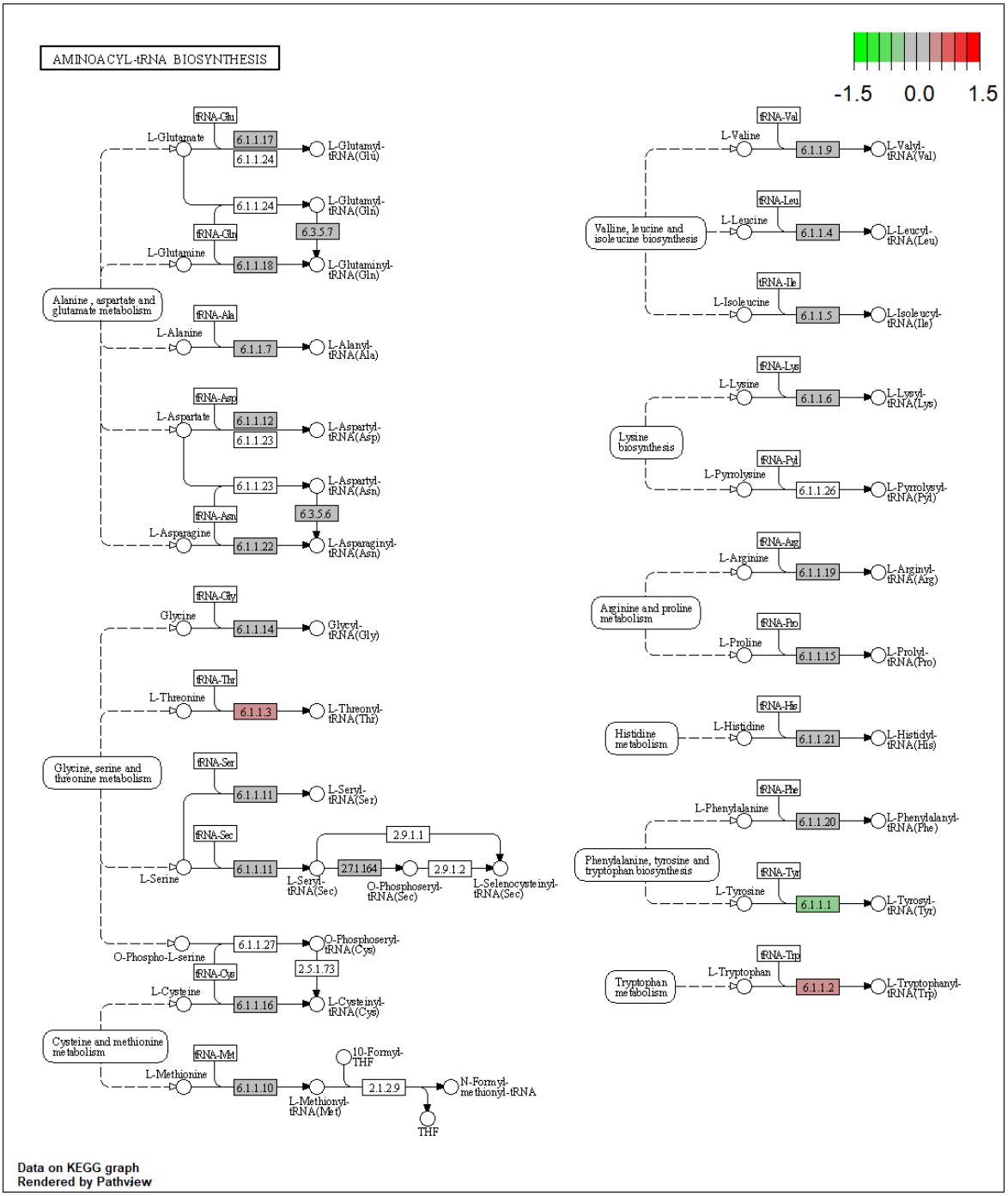


Figure 3.13: Continued on next page

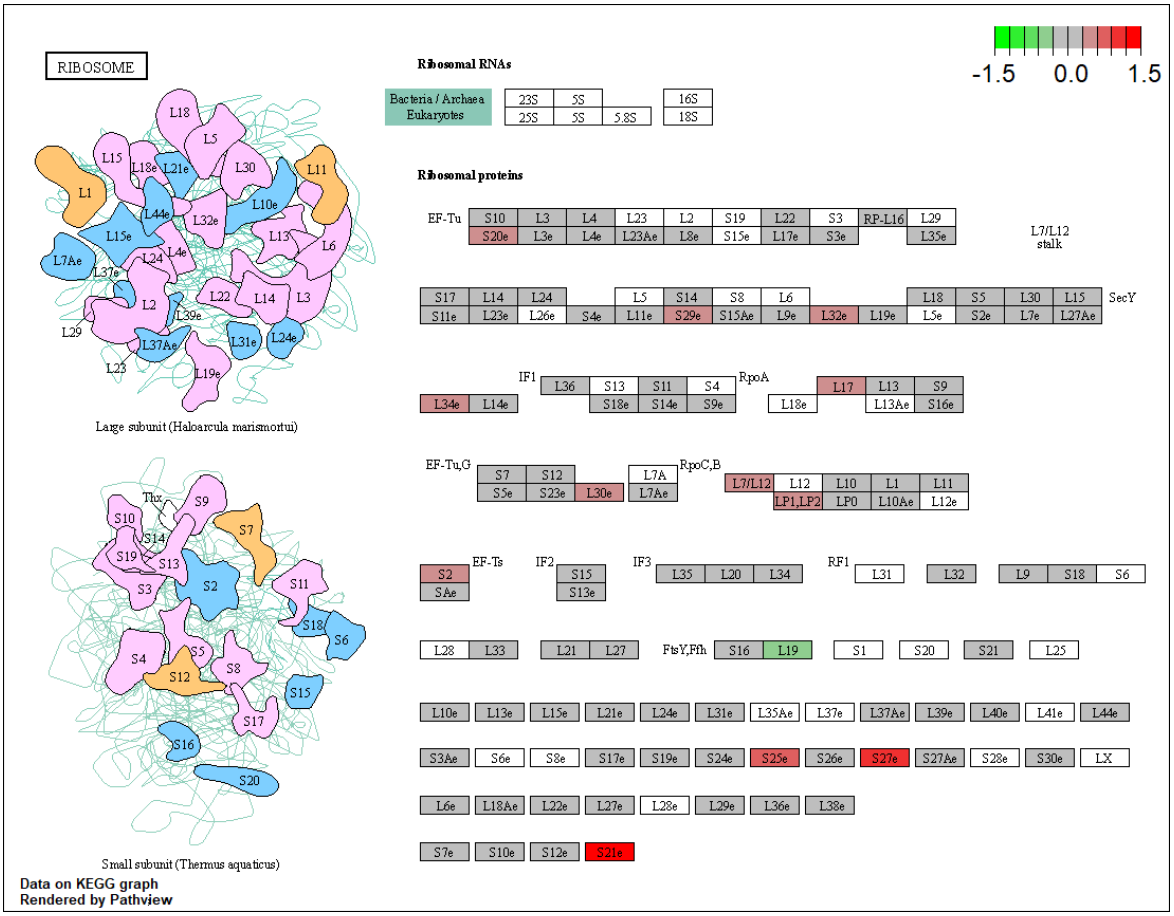


Figure 3.13: Pathways relating to Translation show little difference for the producer versus non-producer on day 2. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function.

3.5.3 Day 2 Protein Folding, Sorting and Degradation

Figure 3.14 shows the protein export, protein processing in the endoplasmic reticulum and SNARE interactions in vesicular transport. Protein Export shows several key genes upregulated. Genes upregulated in this pathway are *SRPR* which is a subunit of the endoplasmic reticulum signal recognition particle receptor that, in conjunction with the signal recognition particle, is involved in the targeting and translocation of signal sequence tagged secretory and membrane proteins across the endoplasmic reticulum. It appears to play a crucial role in the insertion of secretory and membrane polypeptides into the endoplasmic reticulum. This protein was found to be tightly associated with membrane-bound ribosomes, either directly or through adaptor proteins.

Overall, proteins involved in the ERAD pathway are downregulated, along with quality control steps like the reglucoslates (*UGGT1*) gene. Overall, genes involved in translocation, protein folding, disulphide interchange reactions and ER associated ribosome proteins are upregulated, while genes associated with the ERAD are downregulated. Genes looked at in this pathway where $p_{adj} < 0.05$.

SNARE interactions in vesicular transport show statistically significant upregulation of several genes. The STX1-4 genes appear to be downregulated, which may not agree with literature, which found when looking at some of the genes involved in this pathway, including VAMP, that over expression increases protein titre (Mozzanino, 2018).

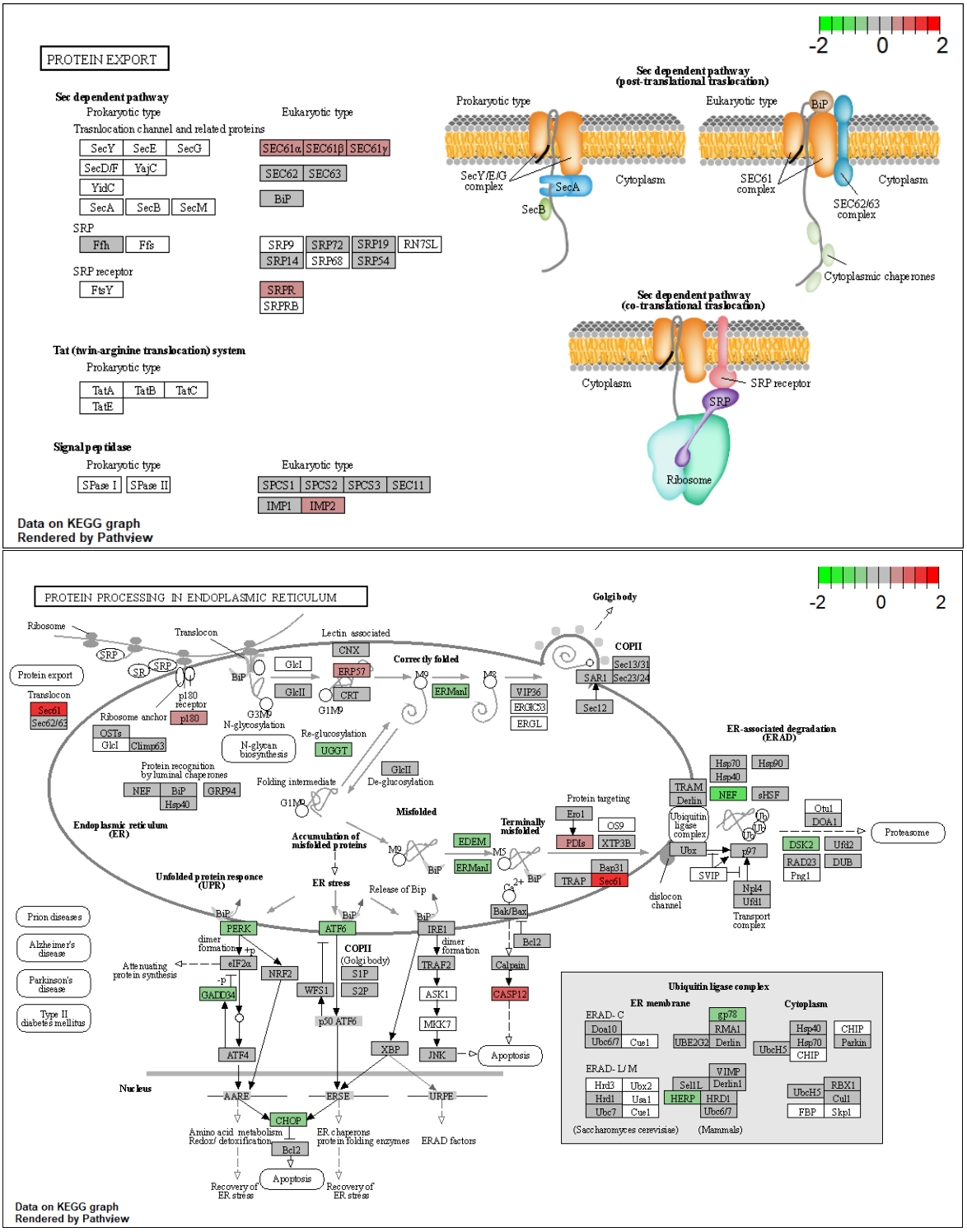


Figure 3.14: Continued on next page.

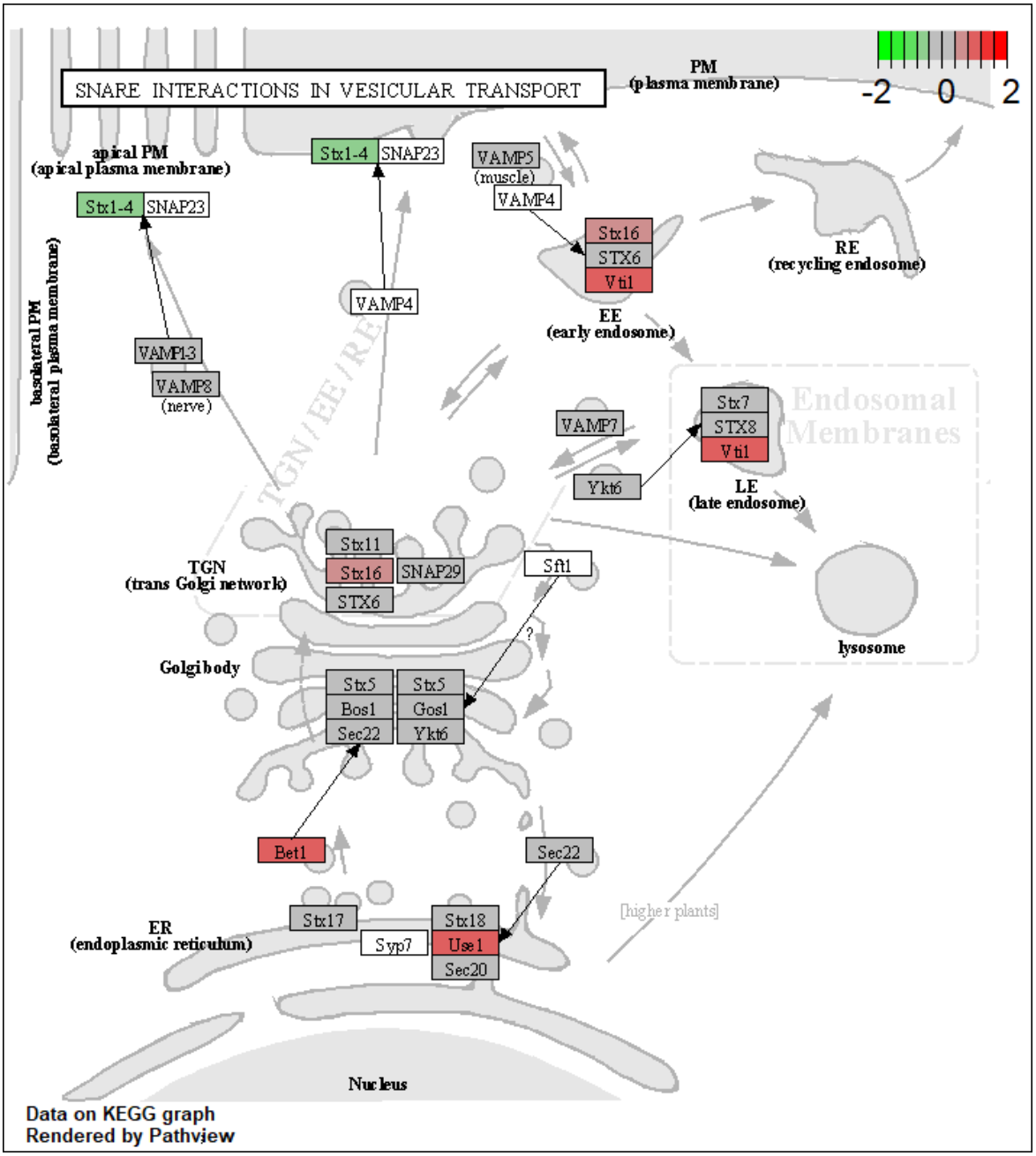
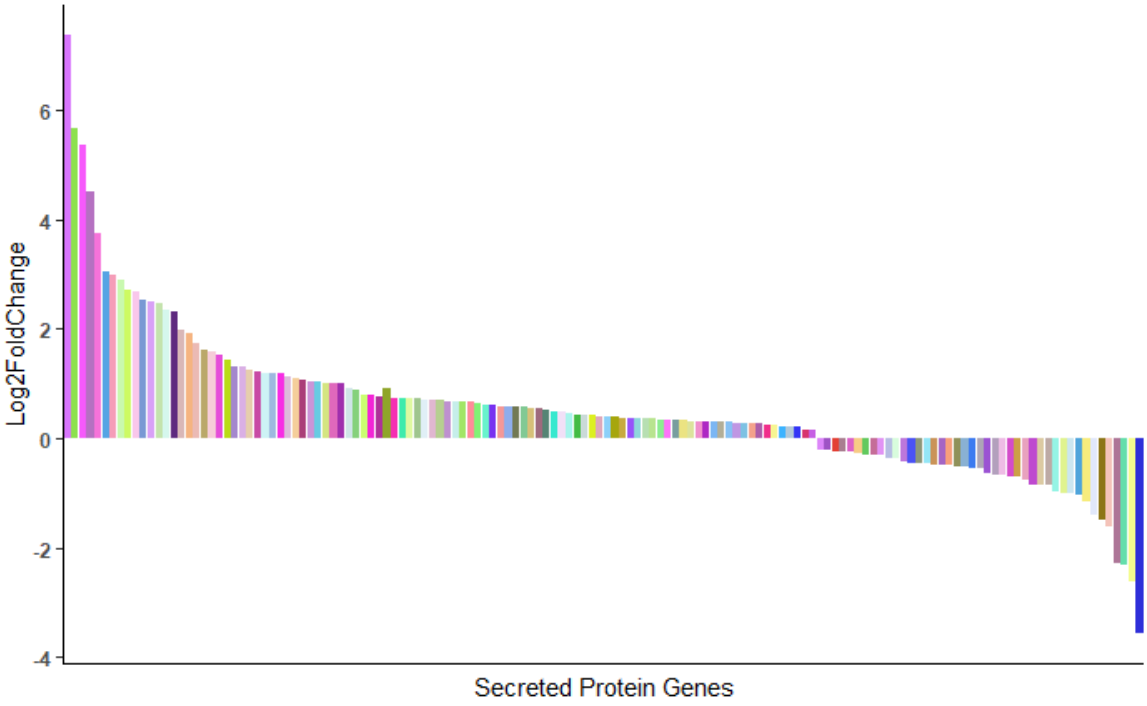


Figure 3.14: The non-producer is less efficient at moving the protein into the endoplasmic reticulum and out of the endoplasmic reticulum on day 2. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function.

3.5.4 Day 2 Secreted Proteins

To get an idea of how many secreted proteins the cell was producing, the mouse secretome was taken from an online tool called MetazSecKB and converted to CHO orthologues. Only the highly likely secreted genes were taken to reduce false positives. This was to check how the expression levels of these proteins were changing throughout culture when looking at the producer versus non-producer comparison, giving insight into how much more protein the cells were producing that was secreted compared to the non-producing cell line.

Figure 3.15 shows that most of the genes that are likely to be secreted in mouse cells are expressed higher in the producing cells versus the non-producing cells. This is interesting as it gives an idea of how much capacity of the secretory pathway is already taken up at a basal level in the producer versus non-producer before the recombinant protein is taken into account. Unexpectedly, the producer had higher expression of proteins that are secreted and the genes with the largest expression change were *TNC* and *DPT*, which are both associated with TGF-beta signalling, which is often upregulated in cancer cells and may indicate increased cellular stress.



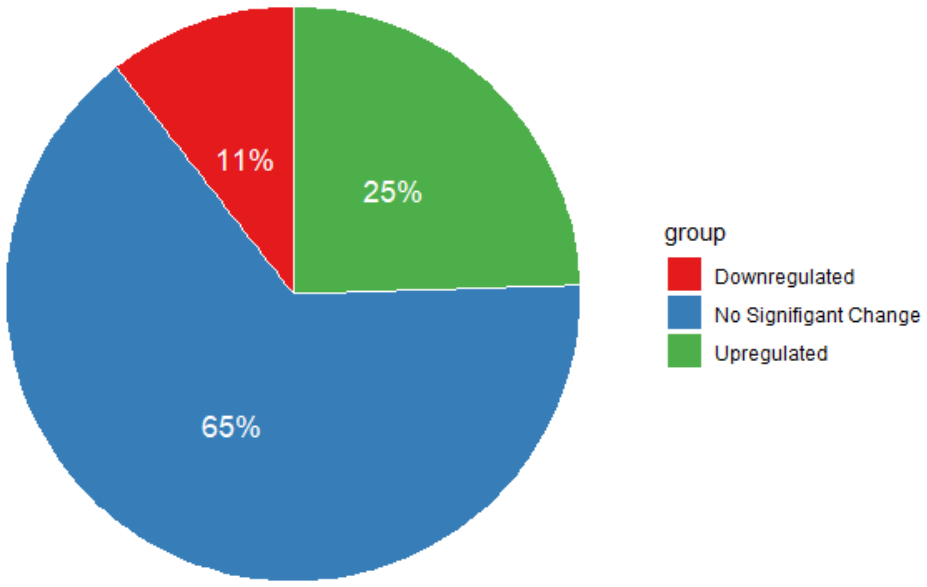


Figure 3.15: The producing cell lines transcribe more secreted protein genes on day 2. The majority of secreted proteins between the producing cell lines and non-producing cell lines show no difference in expression. There are more upregulated secreted proteins than downregulated. The difference in up and downregulated genes on day 2 indicates the producing cell lines are secreting more protein. All genes stated as up or downregulated have a $p_{adj} < 0.05$.

3.6 Day 5 Producer versus Non-producer

This section contains the analysis of the differences between the producers and non-producers on day 5 and attention will be drawn to how the differences have changed from day 2. By investigating this, it will show what has changed due to the producers making more protein and how the cells changing to stationary phase changes the transcriptomic landscape.

3.6.1 Day 5 Transcription

Figure 3.16 indicates little has changed in the area of transcription compared to other pathways. Overall, for the RNA polymerase subunits, there is more downregulation. In contrast, the basal transcription factors show upregulation of *TBP* and *TAF4* which may indicate upregulation of TATA box-based transcription in cells producing a recombinant antibody.

The splicesome showed very little difference in gene expression on day 2. On day 5 this pathway has shown several genes to be upregulated. Genes associated with the *PRP19* complex and *TREX* complex are upregulated. Comparing this to day 2, genes related to the *PRP19* complex are upregulated in the producer on day 5. The *PRP19* complex is related to many biological processes in the cell including splicing, transcription elongation, genome maintenance, lipid biogenesis and recruitment of ubiquitylated proteins to the proteasome. It is also suggested that the *PRP19* complex is indirectly associated with transcription initiation and mRNA export (Chanarat and Sträßer, 2013). This may coincidently relate to the upregulation of the *TREX* complex, which functions to connect transcription elongation to correct 3' end formation for nuclear transport.

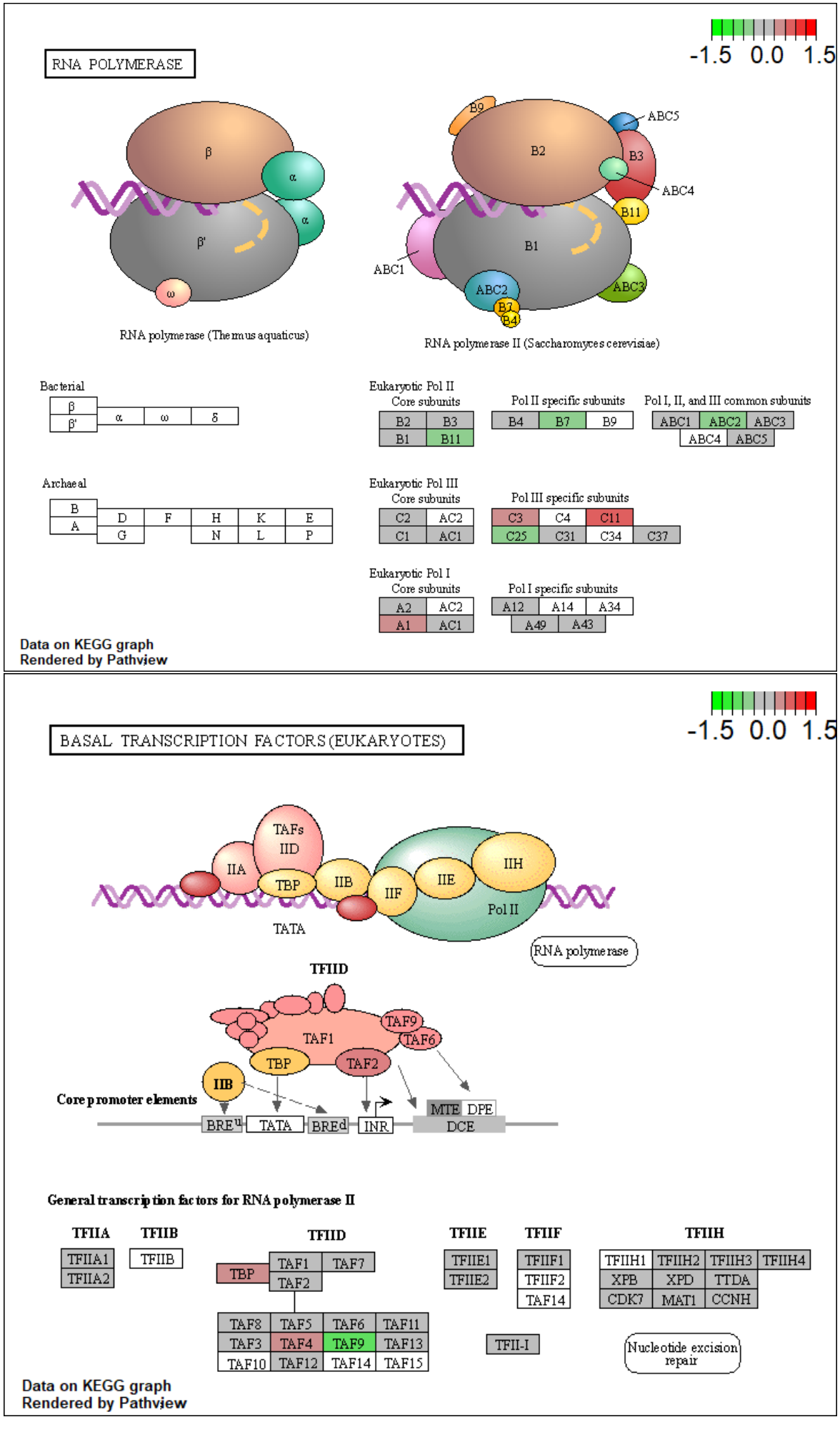


Figure 3.16: Continued on next page

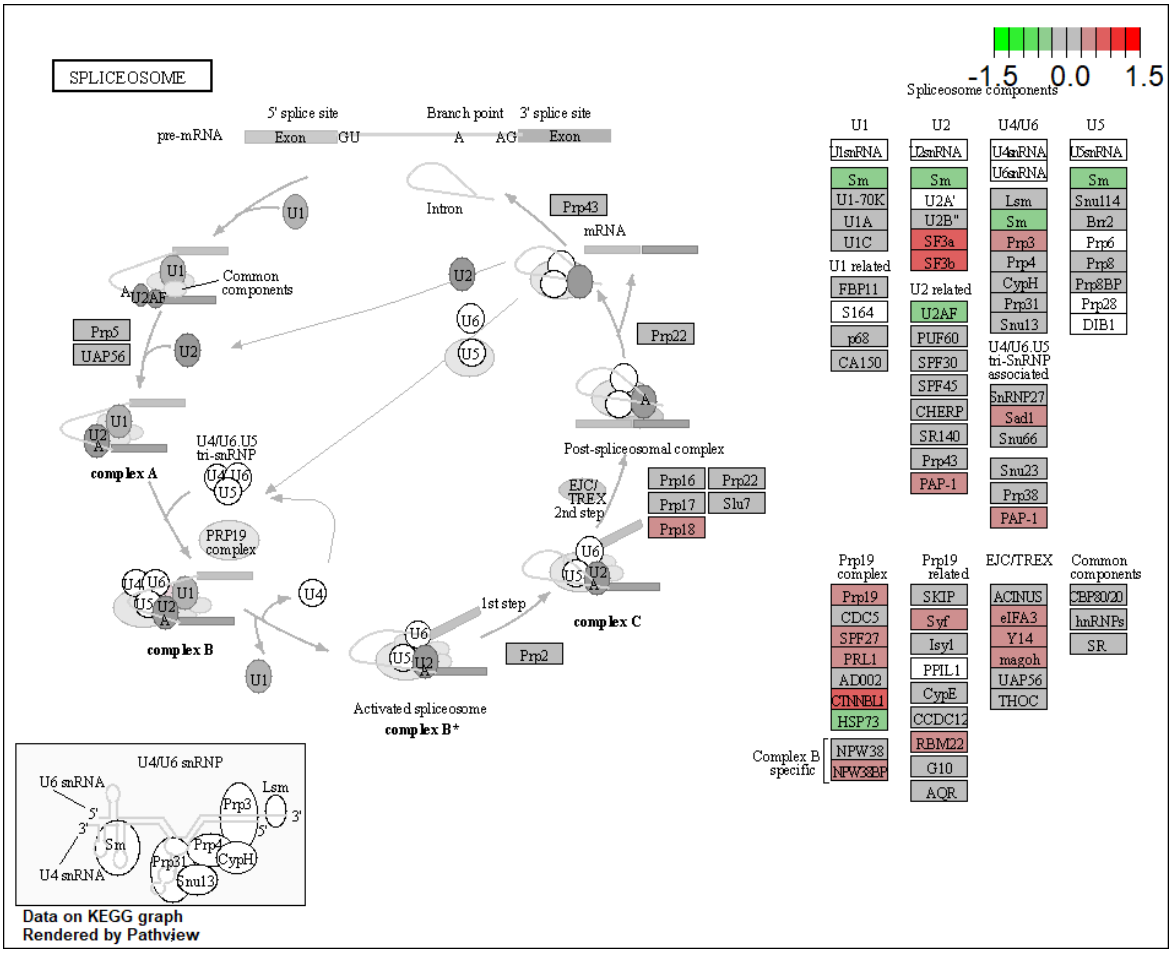


Figure 3.16: Pathways relating to transcription show increased variation from day 2 to day 5. All genes that have been mapped with a fold change have a $p_{adj} < 0.05$. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function.

3.6.2 Day 5 Translation

The mRNA surveillance pathway, as shown in Figure 3.18 is similar to the trend seen in transcription. There appears to be far greater differential expression between the mock and producing clones on day 5. Upregulation is seen in the exon junction complex, which was also seen in the spliceosome. Most significantly, *PP2A* has gone from being downregulated on day 2 to being overexpressed in recombinant cells on day 5. Genes such as these would be expected to be downregulated as the recombinant cells have a reduced growth rate at this stage of culture and *PP2A* is linked with being a growth suppressor.

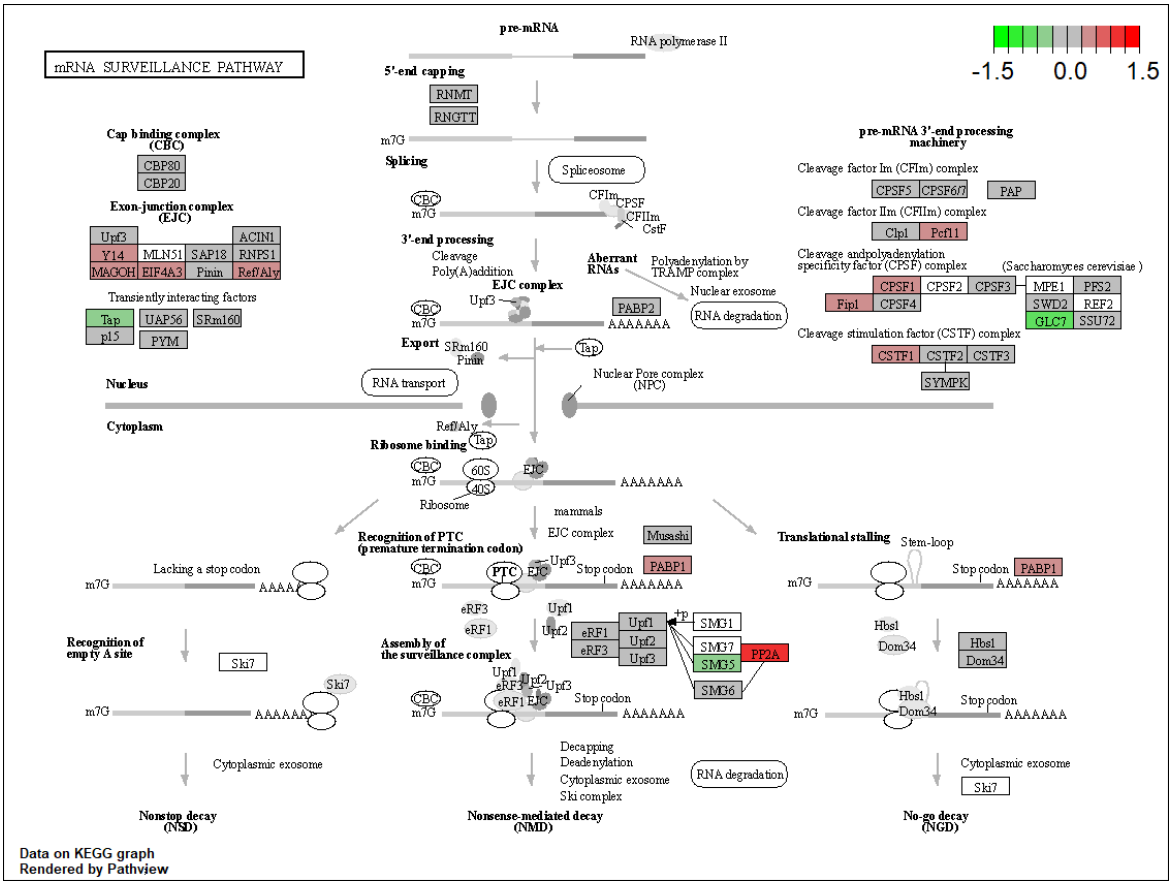


Figure 3.17: The mRNA surveillance pathway showed increased differential expression compared to the non-producing cell line on day 5. All genes that have been mapped with a fold change have a $p_{adj} < 0.05$. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function.

Figure 3.18 shows the RNA transport pathway. Similar to the previous pathway, day 5 shows much more upregulation than seen on day 2. Looking at the KEGG map, complexes such as translation initiation factors and the exon junction complex seem to be upregulated along with the upregulation of certain specific genes. For instance, *IPOB*, is a member of the iron/manganese superoxide dismutase family.

It is involved with shuttling snRNA back into the nucleus after processing and heavily indicates that the producer cells have increased splicing compared to the non-producing cell. This could be due to the selection for high-producing clones, as it has been shown that increased splicing led to enhanced translation (Nott et al., 2004) and at this point, the producer cells are starting to produce larger amounts of protein. It also appears that along with increasing splicing, the genes associated with the translation initiation complex are upregulated.

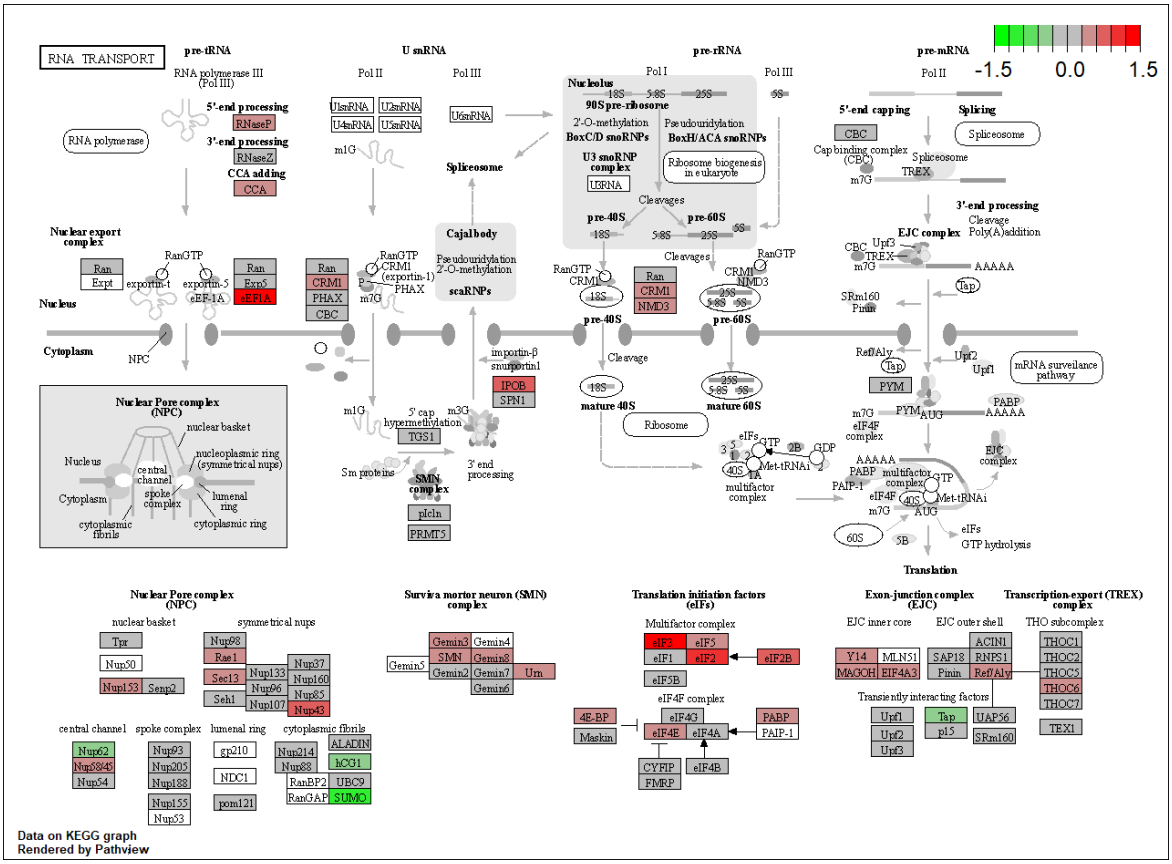


Figure 3.18: Genes involved in RNA transport appear to show increased upregulation in genes related to tRNA,snRNA and rRNA. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function. All genes that have been mapped with a fold change have a $p_{adj} < 0.05$ on day 5.

As previously noted, day 2 ribosome biogenesis appeared to have little upregulation. This was against the hypothesis that more ribosomes would be required in a protein-producing cell as more protein synthesis would be occurring. On day 5 broader changes are seen, with 19 genes upregulated versus 2 on day 2. The most upregulated of these are *TCOF1*, *RCL1*, *NUG1/2* and *DRG1*. *TCOF1* is linked with ribosomal DNA transcription, while RCL1 is linked with rRNA cleavage. This can be seen in Figure 3.19.

The genes associated with the ribosome are also differentially regulated between day 2 and day 5. More ribosomal proteins are upregulated on day 5. Most noteworthy is *L44E*, which showed no large fold change for day 2 but is heavily upregulated on day 5. This gene is associated with helping in protein synthesis within the mitochondria.

Consistently with the previous two pathways, tRNA biosynthesis also shows major upregulation compared to day 2. All of the tRNA translation between the producer and non-producer clones on day 5 is very different. This could indicate an increased demand for tRNA due to increased protein production within the cell and increased translation occurring.

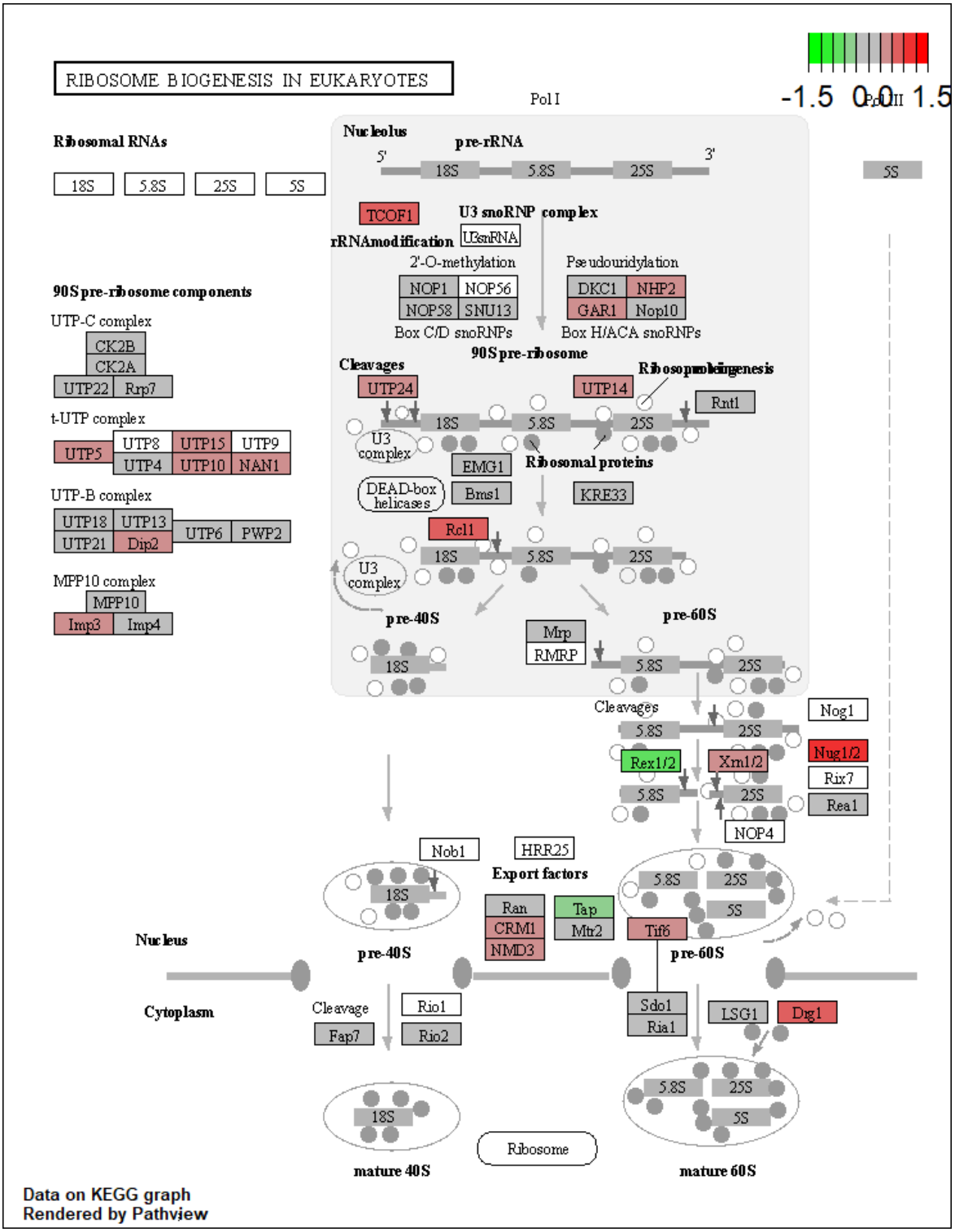


Figure 3.19: Continued on next page

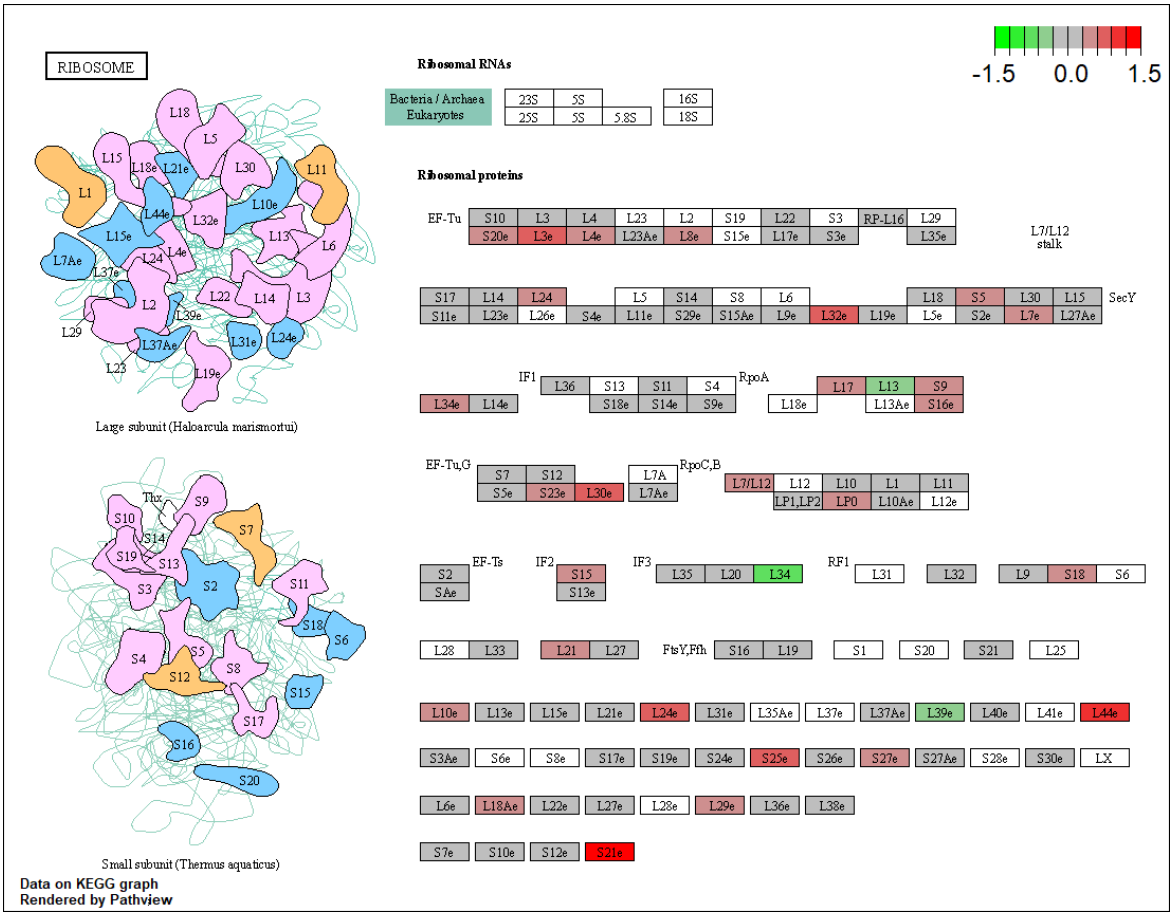


Figure 3.19: Continued on next page

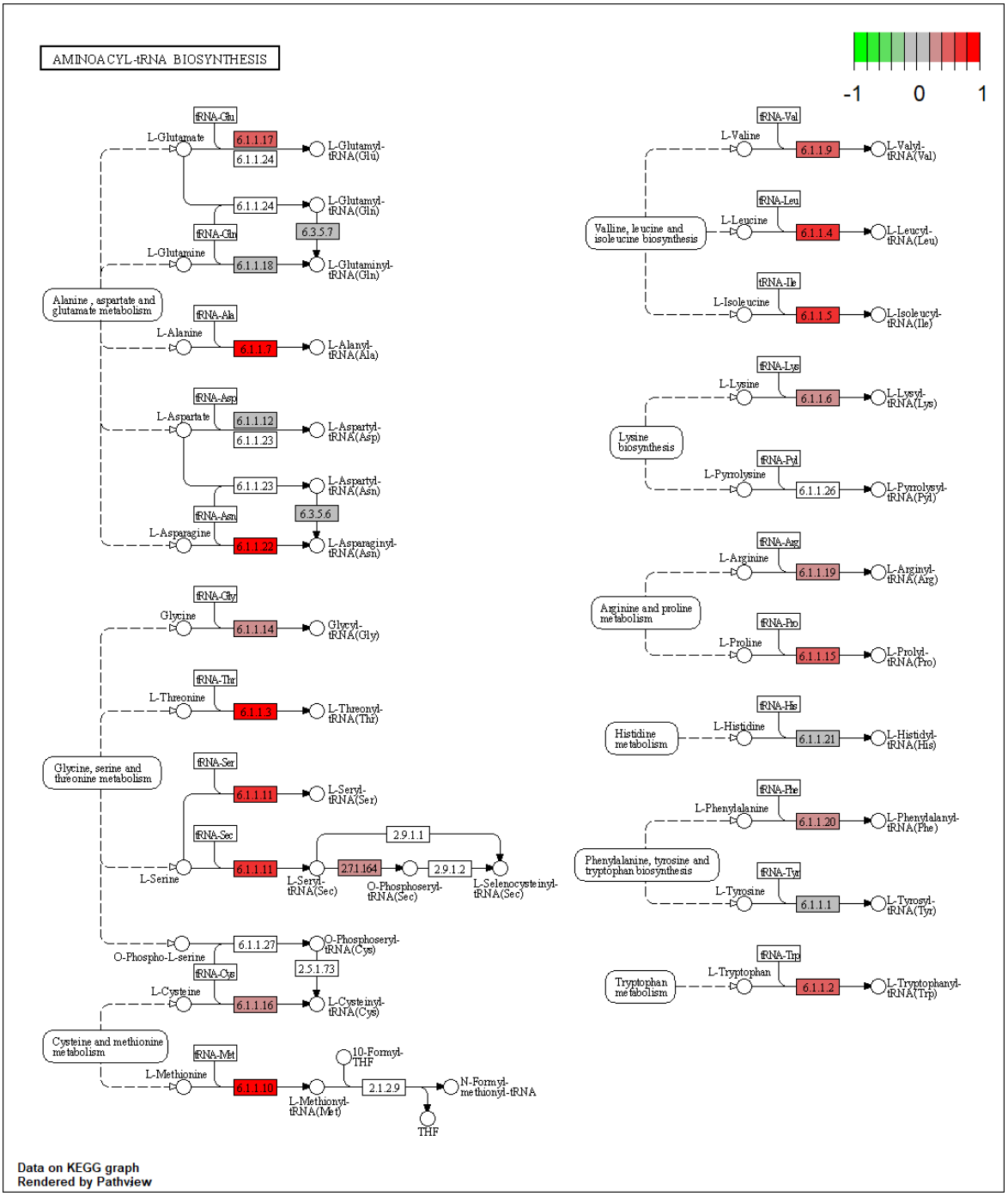


Figure 3.19: Pathways related to translation show increased upregulation in ribosome biogenesis and tRNA synthesis. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function. All genes that have been mapped with a fold change have a $p_{adj} < 0.05$ on day 5.

3.6.3 Day 5 Protein Folding, Sorting and Degradation

Figure 3.20 shows the differential expression of genes in the protein folding, sorting and degradation pathways. Protein export is very similar to day 2. The only differences are increased upregulation of *IMP2* and *SRP54*. *SEC61B and y* have lower fold changes compared to day 2 also. *CHOP* is heavily upregulated indicating, that CHOP induced apoptosis is likely to occur soon.

Protein processing in the endoplasmic reticulum has large changes compared to day 2, possibly due to increased protein production on day 5. Interestingly, *SEC61* has no significant change. *SEC61* was upregulated on day 2 and the protein targeting genes such as *PDIs* and *TRAP* show reduced differential expression between day 2 and day 5. On day 2 these were heavily upregulated. *CASP12* also switches from being heavily upregulated to no change; instead, *CALPAIN* is downregulated. This is interesting as *CALPAIN* is linked to calcium dependency and may indicate a depletion of calcium inside of the ER or could even indicate non-functioning of the ER (Mekahli et al., 2011). A further interesting gene being upregulated is *CHOP* which was downregulated in protein-producing cells on day 2. This gene has been shown to promote apoptosis, which may be the reason for the decreasing cell growth. A paper has shown deleting the *CHOP* gene may increase the ability of the ER to produce a protein (Marciniak et al., 2004).

Overall, it appears that many of the genes related to protein processing are downregulated. On day 5 many of the functional components of the ER are being expressed less. This was not expected as with the increased burden of producing the recombinant antibody, one would expect increased expression of these genes to deal with the production of protein. Still, perhaps the cells have reached their limit and as such *CHOP* induced apoptosis is occurring and the endoplasmic reticulum is trying to self regulate itself by slowing down.

The only differences within the SNARE interactions in vesicular transport are the reduced expression of *VAMP8*, *STX 11* and *USE1*. Instead, *STX5* is upregulated, which again coincides with increased autophagy/apoptosis that may be occurring as it is a key regulator in this process.

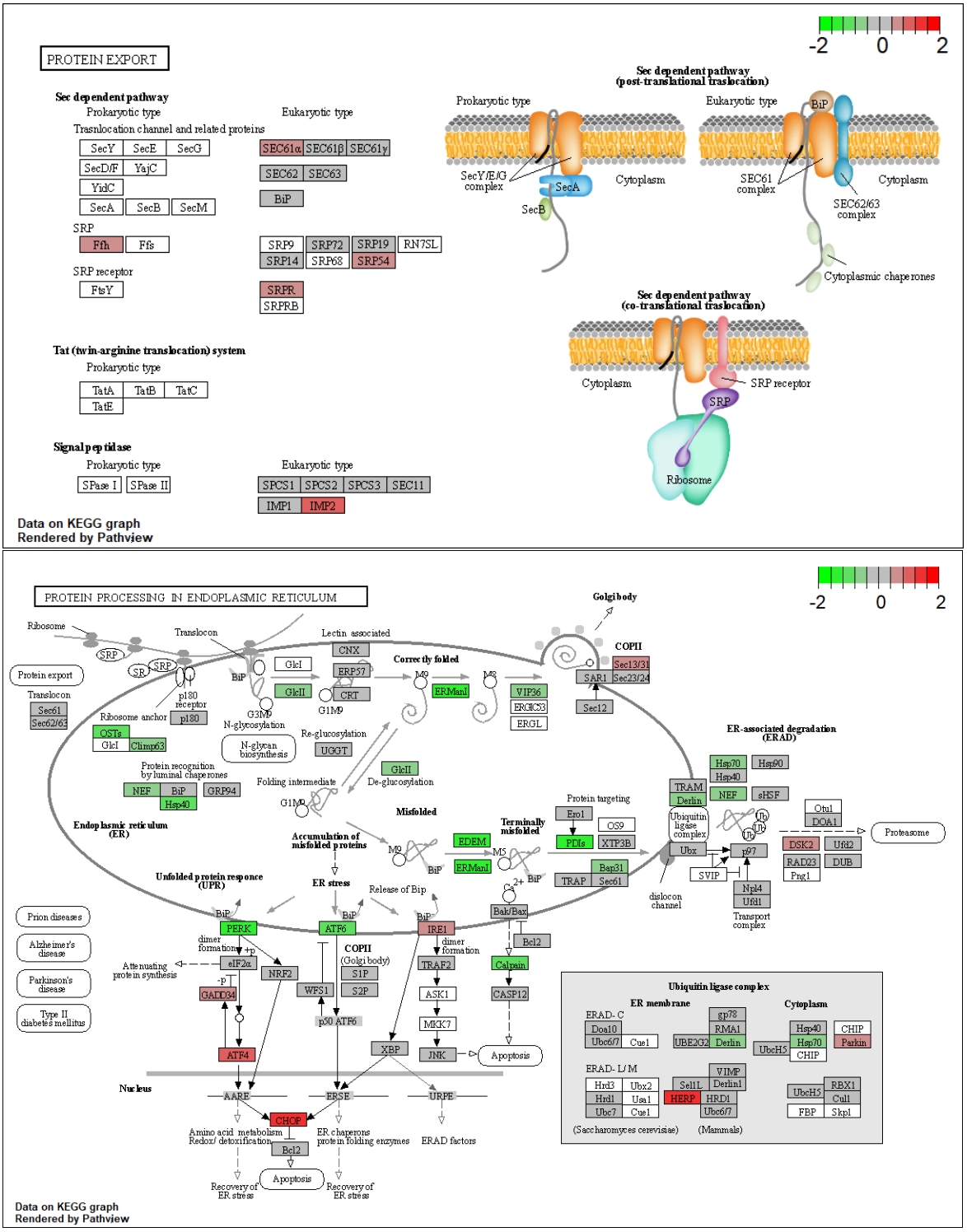


Figure 3.20: Continued on next page

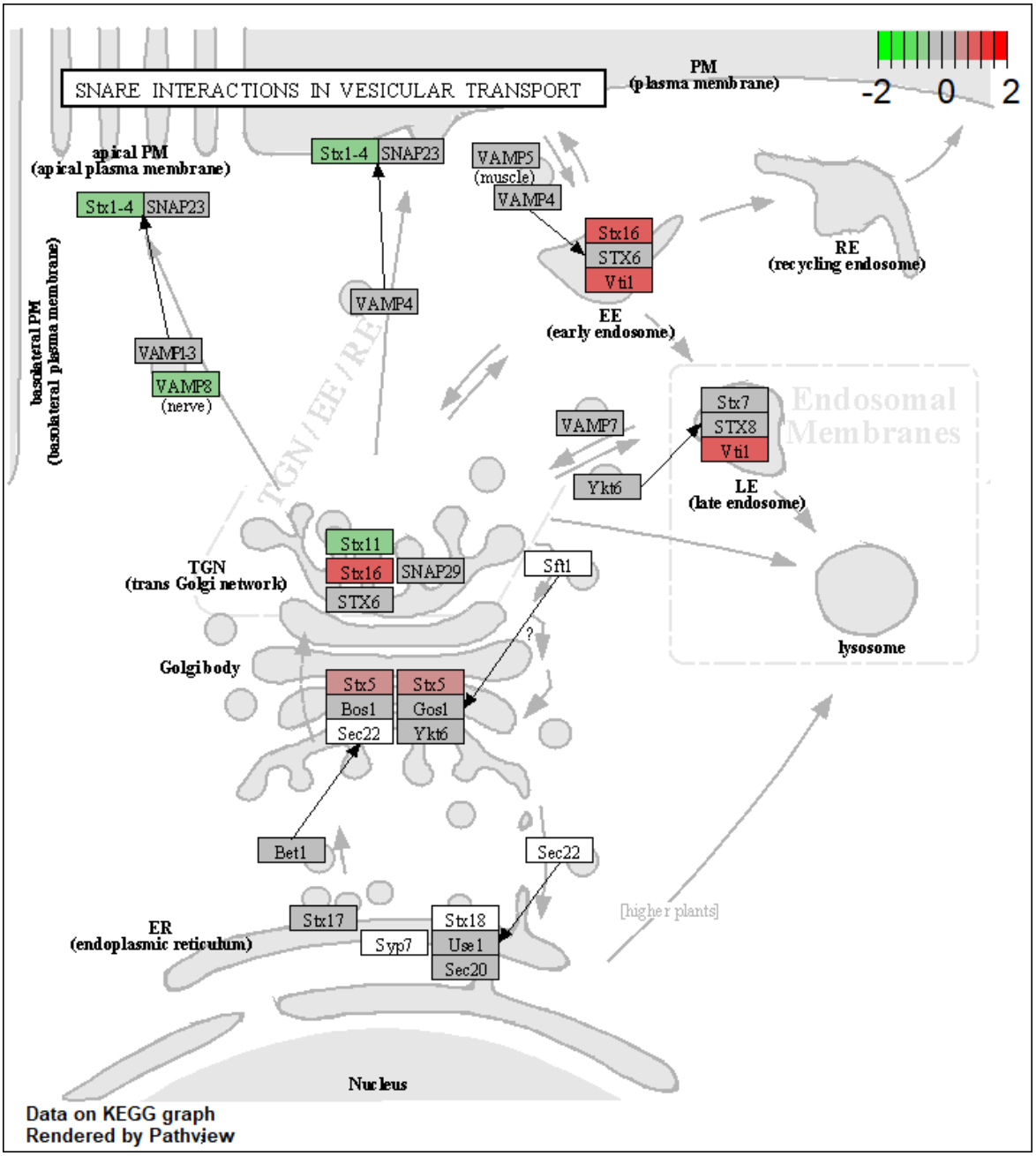
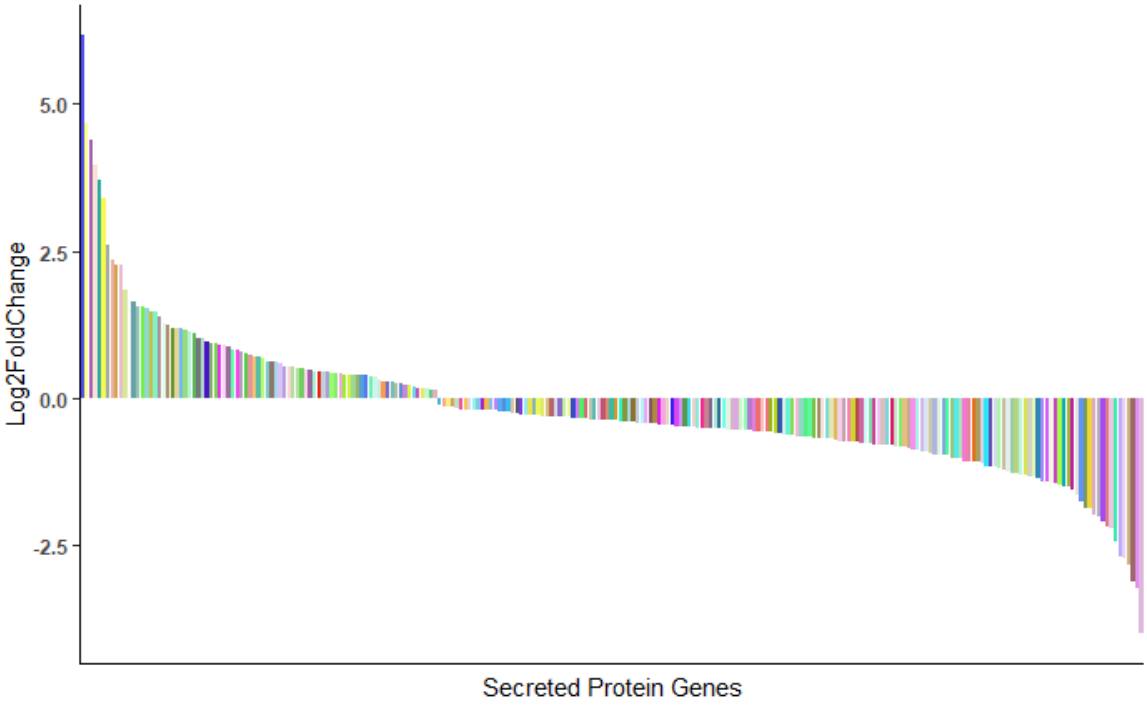


Figure 3.20: Genes in this pathway appear to show increased protein export over the non-producing cell line and increased vesicular transport via the SNARE pathway toward the Golgi and lysosome on day 5. All genes that have been mapped with a fold change have a $p_{adj} < 0.05$ on day 5. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function.

3.6.4 Day 5 Secreted Proteins

The secreted proteins within the CHO cell are shown below in Figure 3.21. A large shift can be seen from the day 2 results in which secreted proteins appeared to be upregulated 25% compared to an upregulation of 18% and a downregulation of 35%, compared to 11%. This could be due to endoplasmic reticulum failure meaning previously secreted proteins can no longer travel through the ER. Thus, there is an overall downregulation to attempt to stop ER dysregulation and restore homeostasis. Although, it would be expected that transport genes and vesicular transport would also be downregulated if the amount of protein passing through the ER is reduced. More likely, this is caused by the cells dying and no longer communicating with one another due to the harsh conditions they are enduring.



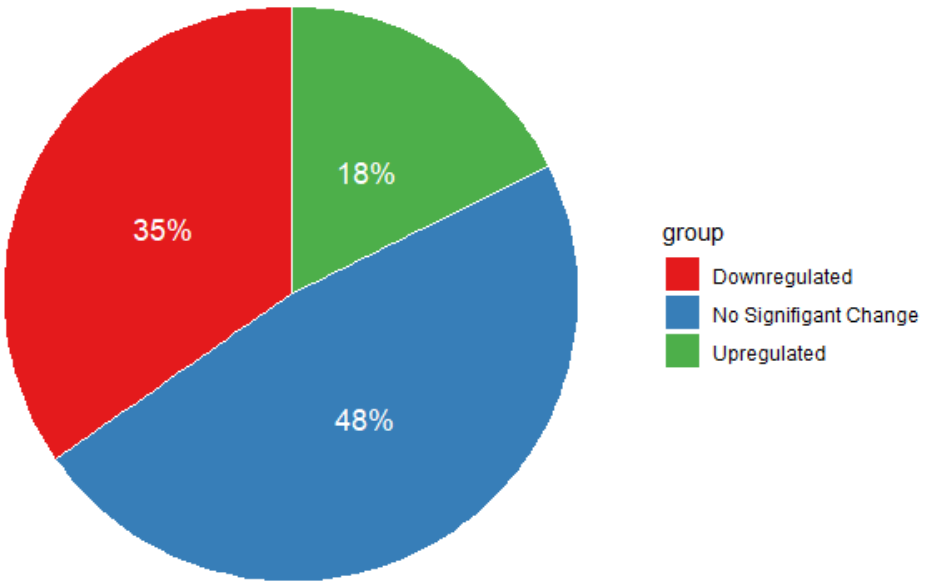


Figure 3.21: Genes likely to be secreted show a trend of downregulation in the producers on day 5. This could be due to many mechanisms, but the most likely is that because the cells are no longer growing, they are secreting fewer signals to one another. All genes that have been measured have a $p_{adj} < 0.05$.

3.7 Day 10 Producer versus Non-producer

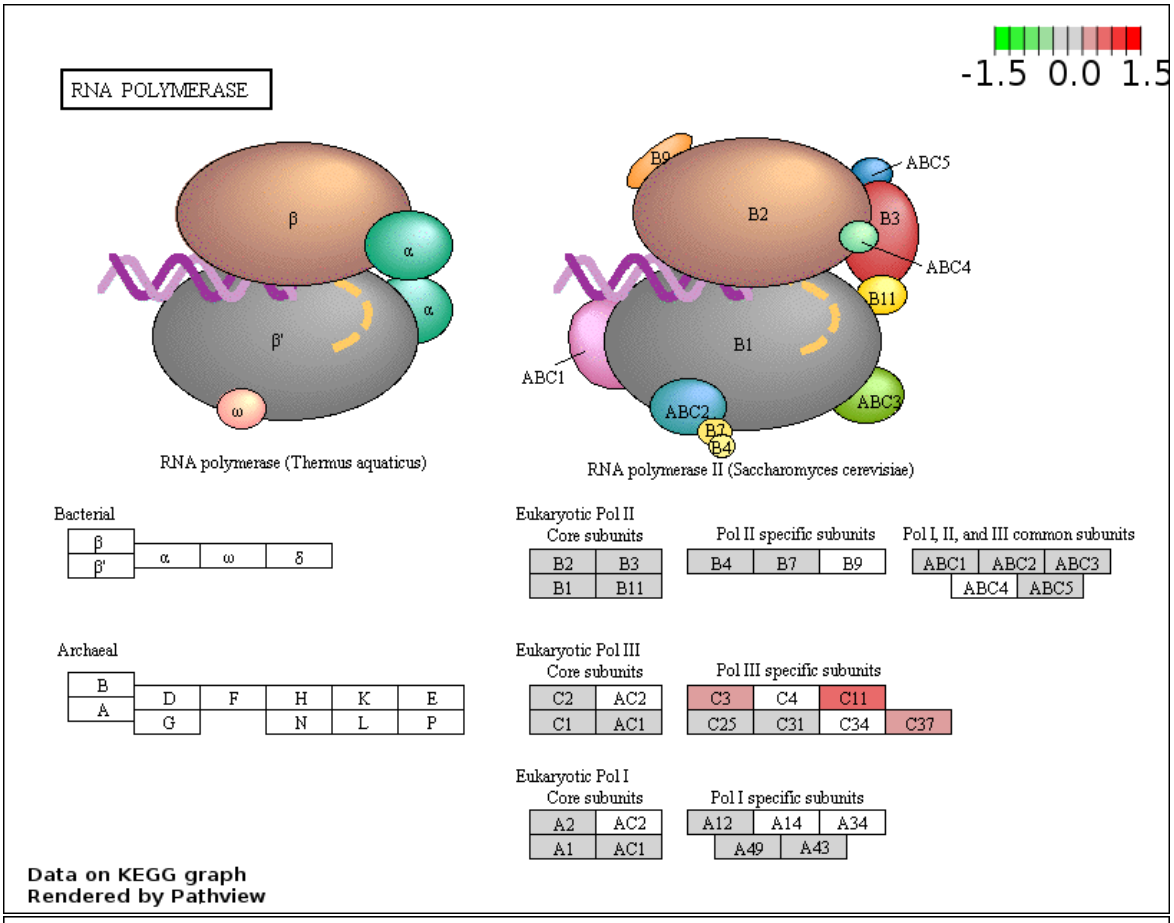
The majority of the recombinant protein appears to be created in the latter part of the process, which can be seen in Figure 3.2. This is also occurring at the same time as the cell density and viability have been dropping for five days straight. This is unexpected as it appears the majority of the protein is produced when the cells are most stressed. This may indicate that the cells at this point have shifted fundamentally compared to day 5 and are more primed to produce protein. Although, it could also be the case that the cells are releasing more protein due to lysis and what is being measured is intracellular IgG that was once trapped in the cell.

3.7.1 Day 10 Transcription

Transcription does not show much difference between the producer and non-producer cells on day 10. Some downregulation in genes such as *B11, B7* and *ABC2* has vanished, while *C3* and *C11* are still upregulated. The shift appears to suggest that genes related to RNA Pol II activity are no longer downregulated. However, transcriptional units are associated with RNA Pol III and thus may indicate the translational pathway is still upregulated at this point in culture compared to the non-producing cells.

The basal transcription factors shown in Figure 3.22 show little difference from what was seen on day 5. The only relevant shifts in gene expression that can be seen is the loss of overexpression of the *TBP* gene which may indicate a reduction in TATA box transcription and the increase in expression of *CDK7* which could potentially indicate that there is DNA damage at this stage of culture or at least another apoptosis signal occurring.

The splicesome also shows a shift compared to day 2. Components of the *EJC* and *PRP19* complex are still upregulated, but there is less overall upregulation in the recombinant cells versus on day 5.



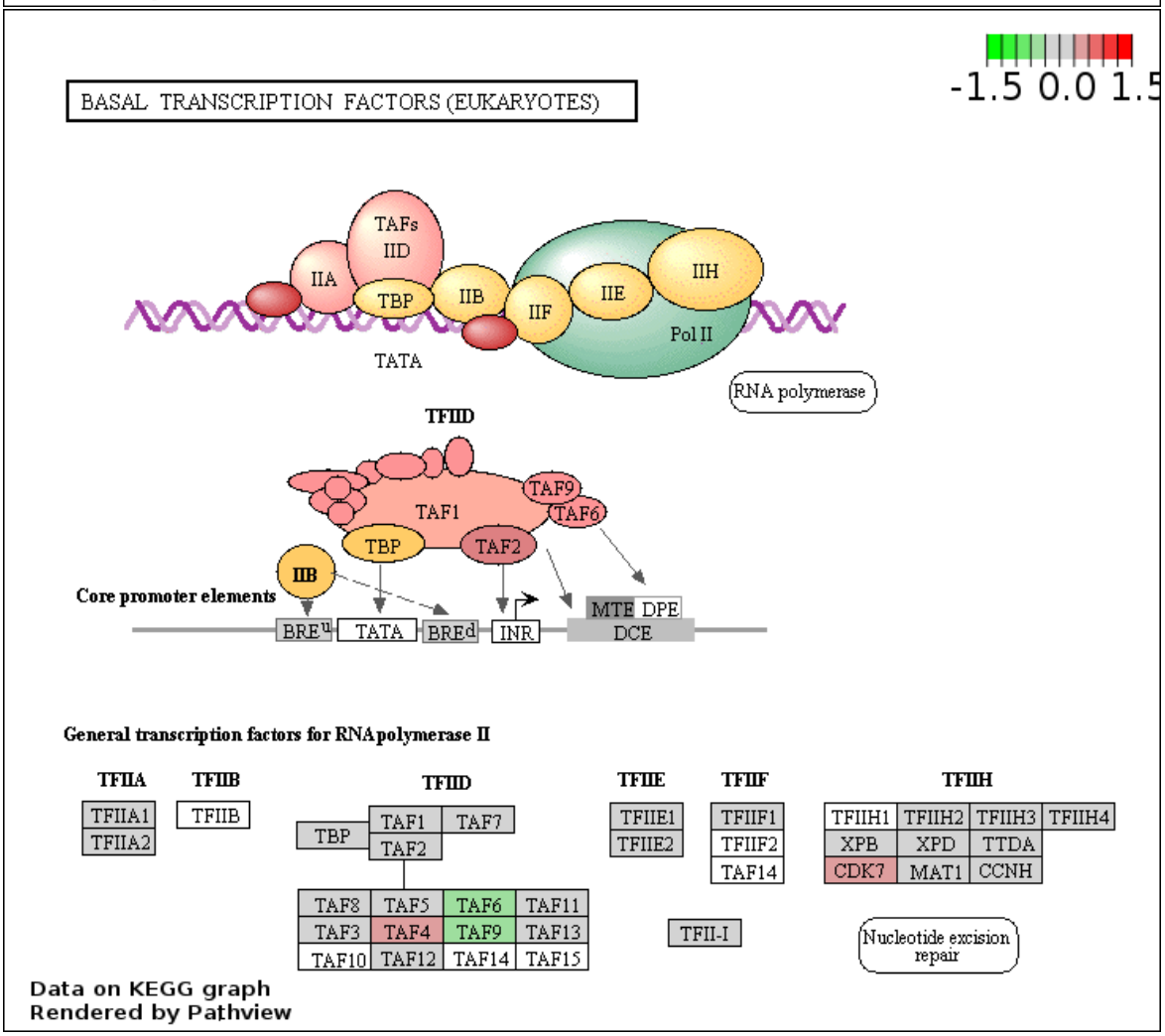


Figure 3.22: Continued on next page

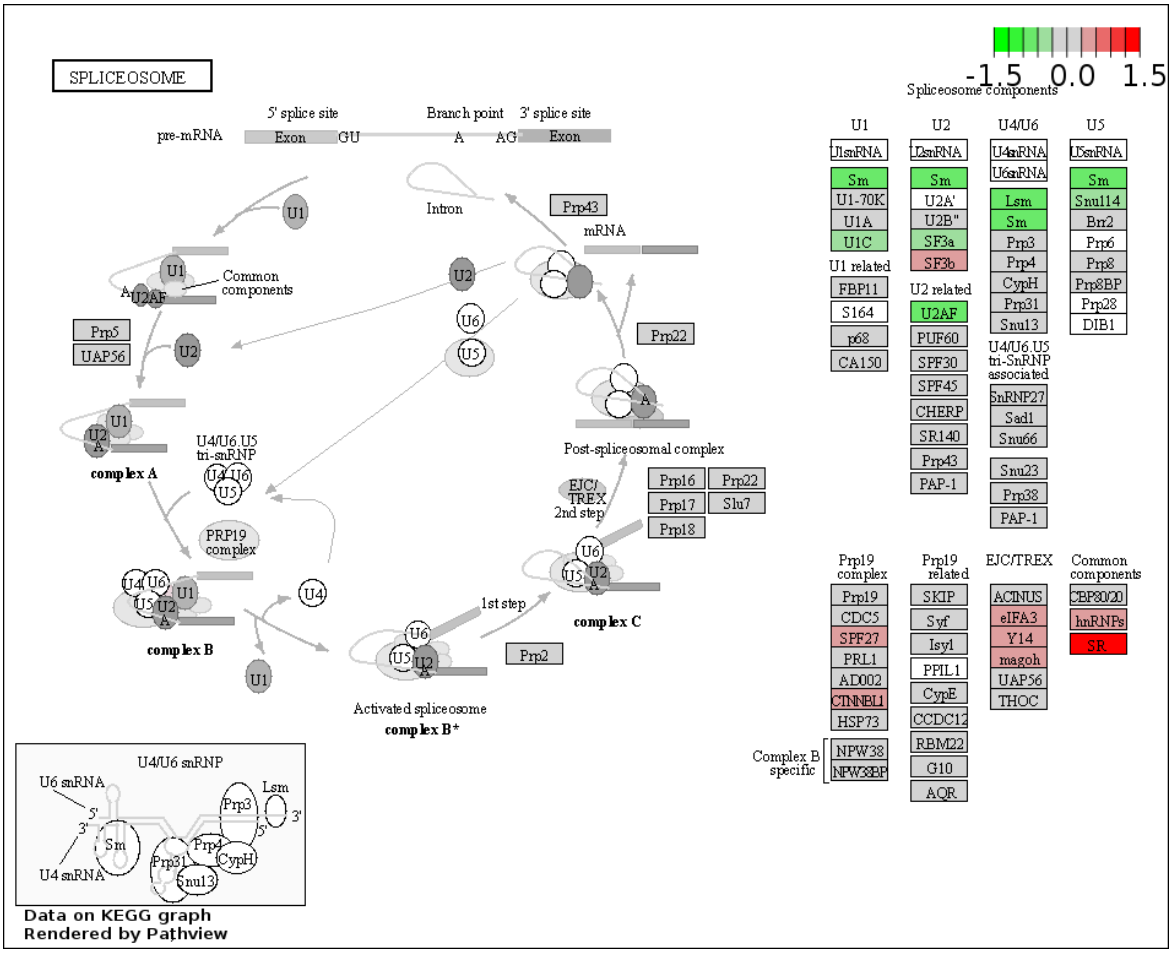


Figure 3.22: Pathways relating to transcription show little variation between day 5 and day 10. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function. All genes that have been mapped with a fold change have a $p_{adj} < 0.05$ on day 10.

3.7.2 Day 10 Translation

The differences between day 5 and day 10 for the mRNA surveillance pathway are not drastic. The most notable changes are an increase in expression of *PERF1*, which is associated with negative regulation of endoplasmic reticulum to Golgi transport and an increase in expression can also be seen from increasing calcium concentrations. The upregulation of *HBS1* may indicate that translational stalling occurs due to many ribosomes stalling on the mRNA. This could be due to the mRNA being produced at this time having an unfavourable secondary structure or adverse conditions for the ribosomes to function. Figure 3.23 shows these changes mapped to the associated KEGG pathway.

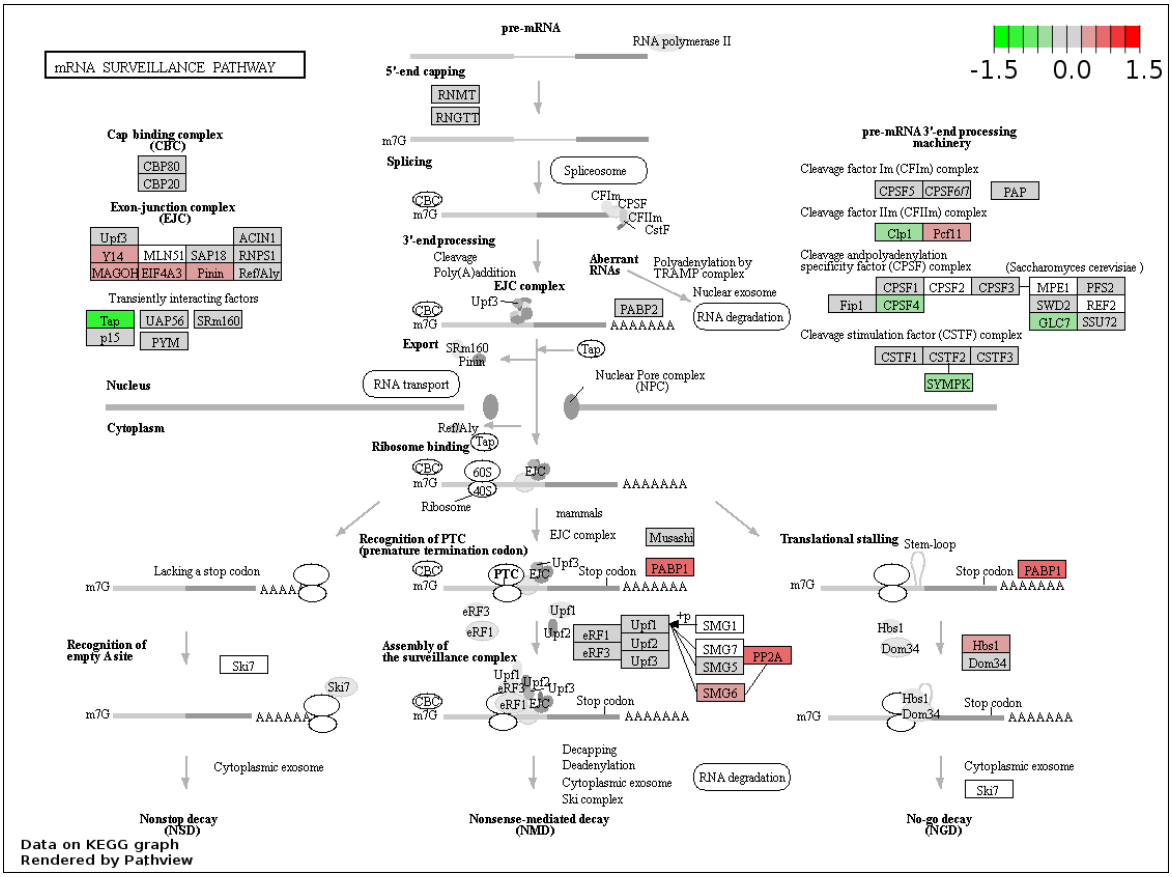


Figure 3.23: Pathways relating to mRNA Surveillance for the producer versus non-producer on day 10. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function. All genes that have been mapped with a fold change have a $p_{adj} < 0.05$ on day 10.

Gene expression in the RNA transport pathway between day 5 and day 10 of this comparison has changed very little, as shown by Figure 3.24. The only notable change is the decrease in the expression of genes related to the survival motor neuron complex. This indicates that there is less snRNA being produced and perhaps, less splicing is occurring. Genes associated with translation initiation are upregulated similar to day 5. The critical difference is the downregulation of a gene called Maskin which downregulated *4E-BP* by binding to it. Once *MASKIN* is no longer binding to *4E-BP*, it activates cytoplasmic polyadenylation element mRNAs which are associated with polyA tail lengthening and thus further attraction of more ribosomes to the mRNA. Hypothetically this could indicate that this is an adaptation to try and produce protein in a cell where the ribosomes are starting to fail and No-go decay is occurring.

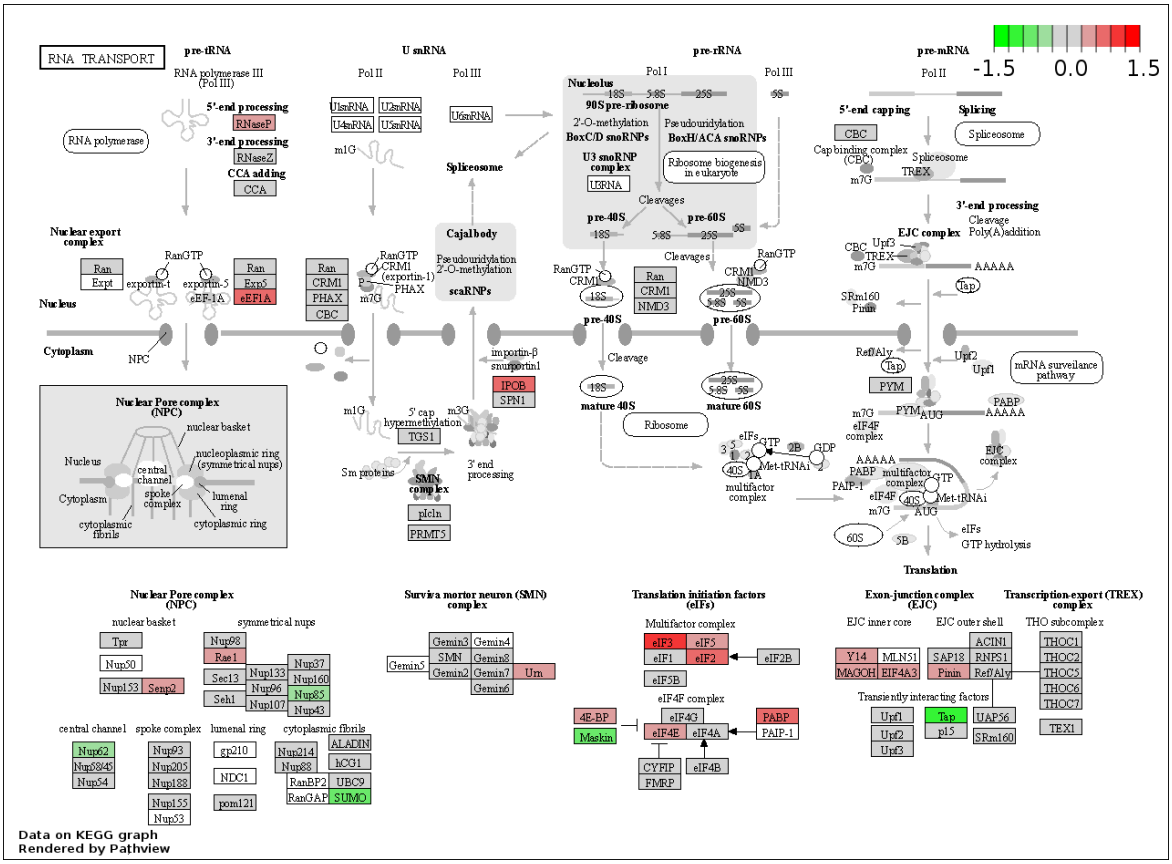


Figure 3.24: Pathways relating to mRNA Transport for the producer versus non-producer on day 10. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function. All genes that have been mapped with a fold change have a $p_{adj} < 0.05$ on day 10.

Regarding ribosome biogenesis, there is little difference between the day 5 comparison and day 10. Notably, it appears that the cells are still trying to make ribosomes at this stage of culture, which is unexpected as the cells are dying. At such low levels of viability, it was expected that the cells would try to slow down protein production.

The proteins of the ribosome show a shift and many of the genes which were upregulated on day 5 are now insignificant versus the non-producing cell line on the same day. *S21E* is the only gene that is still heavily upregulated.

The tRNA pathway shown in Figure 3.22 shows some downregulation compared to what was seen on day 5. The Histidine and L-phenylalamine tRNAs have been downregulated compared to day 5, although a lot of the tRNA synthesis is still high compared to the non-producing cells.

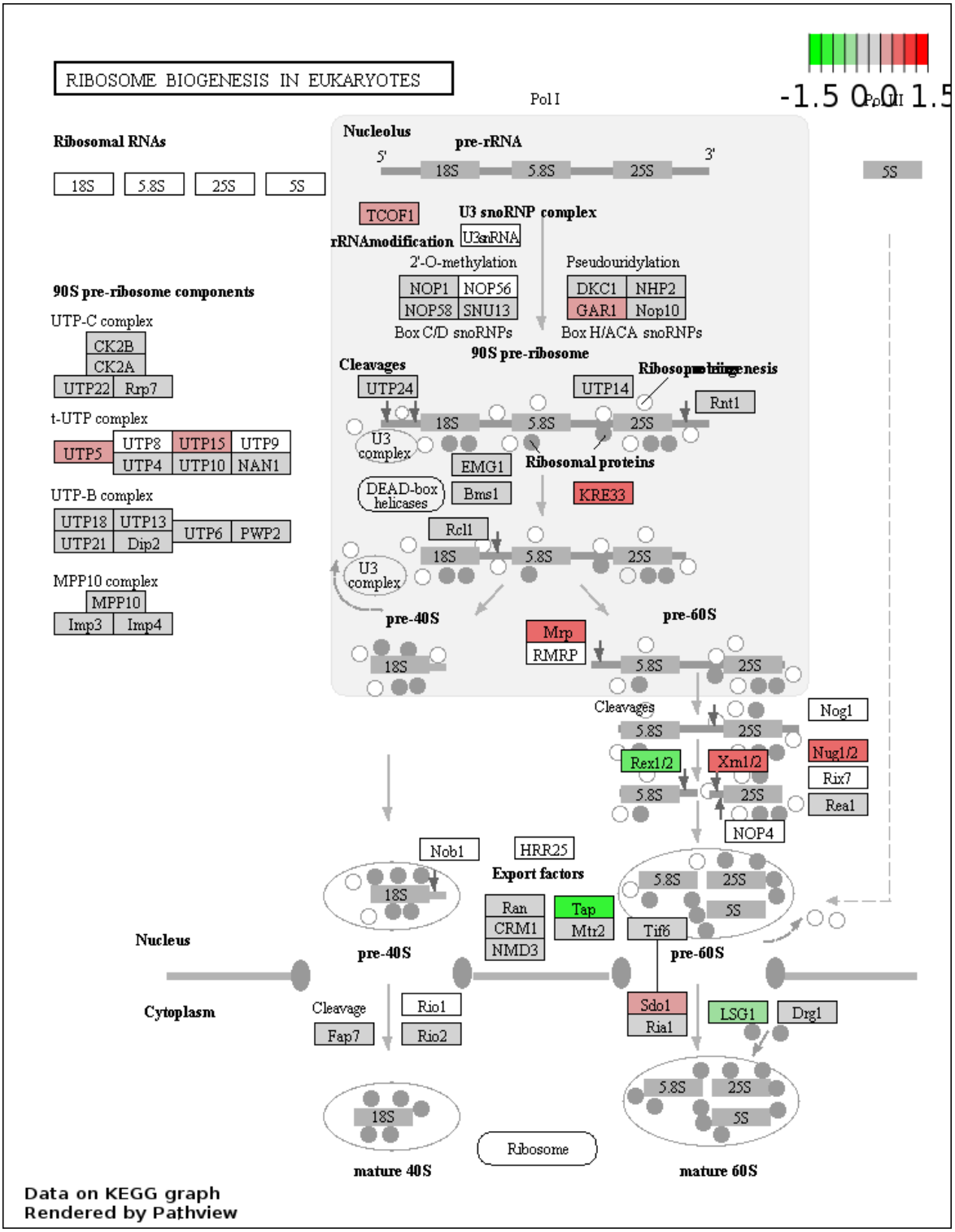


Figure 3.25: Continued on next page

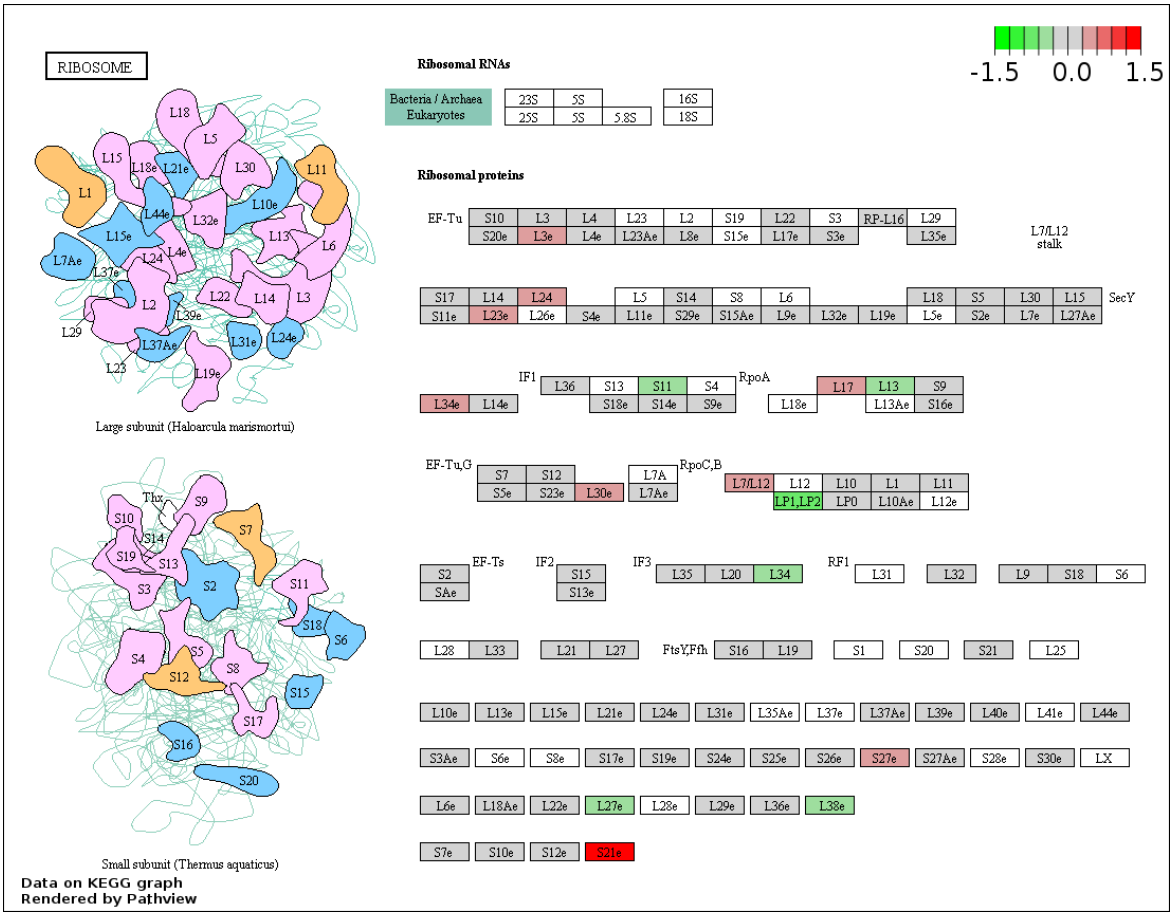


Figure 3.25: Continued on next page

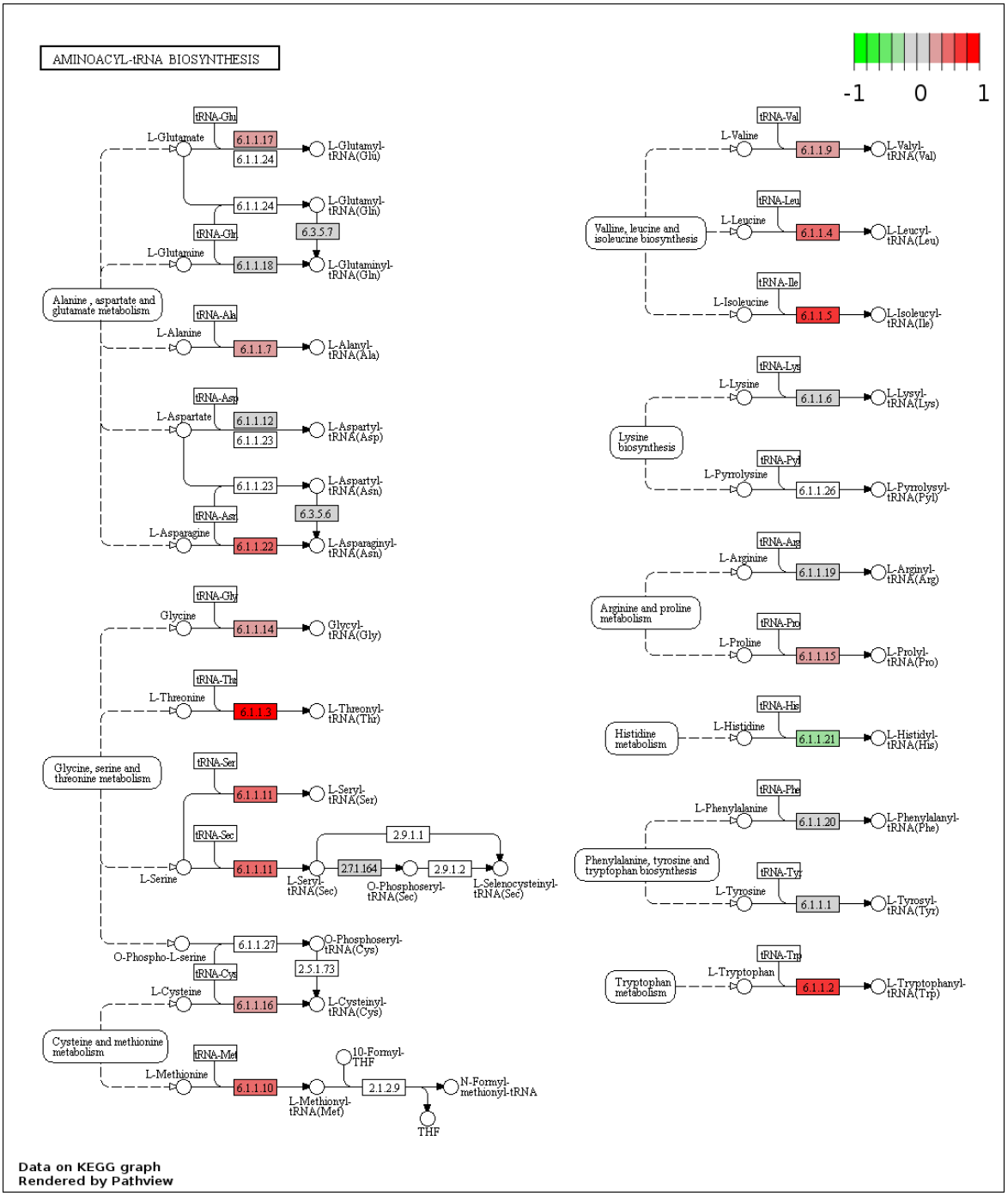


Figure 3.25: Pathways relating to Translation show little difference for the producer versus non-producer on day 10. Genes shown in green are upregulated and red downregulated. Grey indicates no change and white indicates a gene is not found. All comparisons are against the Mock, so any genes in red are upregulated in the cell lines producing protein. Pathway enrichment was performed using a package called clusterprofiler and using the enrichKEGG function. All genes that have been mapped with a fold change have a $p_{adj} < 0.05$ on day 10.

3.7.3 Day 10 Protein Folding, Sorting and Degradation

Figure 3.26 shows pathways relating to protein processing on day 10 for the producer clones against the non-producer. Protein export shows a minor shift with less upregulation of genes such as *SEC61 alpha, PFK, SRP54* and *IMP2. SEC61 beta* is the only gene that has become upregulated.

Protein processing in the endoplasmic reticulum shows less upregulation of proteins

such as *SEC13* and *SEC23*, which may indicate less transport out of the ER. Other notable changes are the increase in *HSP40*, which may mean there are more unfolded proteins in the ER or the cell is undergoing stress. Upregulation of *BCL2* is also occurring. This is interesting because the CHOP protein downregulates *BCL2* but when *BCL2* is expressed, it functions to try and prevent apoptosis in many scenarios. This is interesting as *PARKIN* can have the same function as *BCL2* in trying to inhibit apoptosis.

Vesicular transport shows a very similar trend to day 5 except there is now upregulation of *USE1* and *BET1* which may indicate increased interchange in the ER. The *STX1-4* and *VAMP8* proteins are no longer downregulated, but genes related to the early endosome and late endosome are still upregulated.

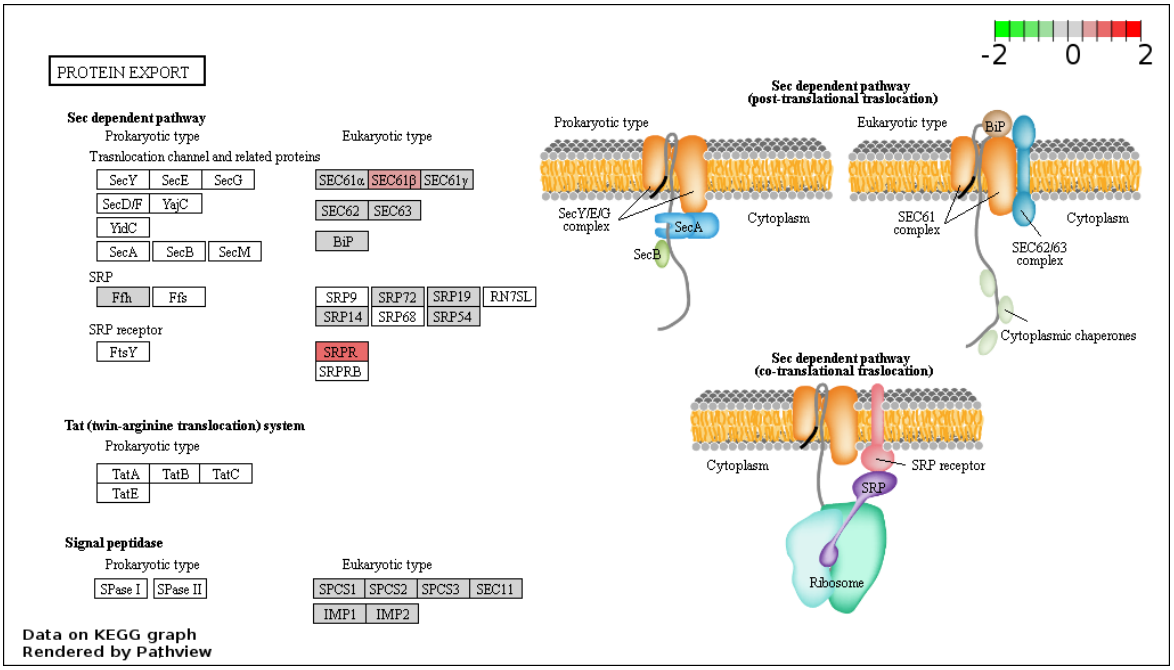


Figure 3.26: Continued on next page

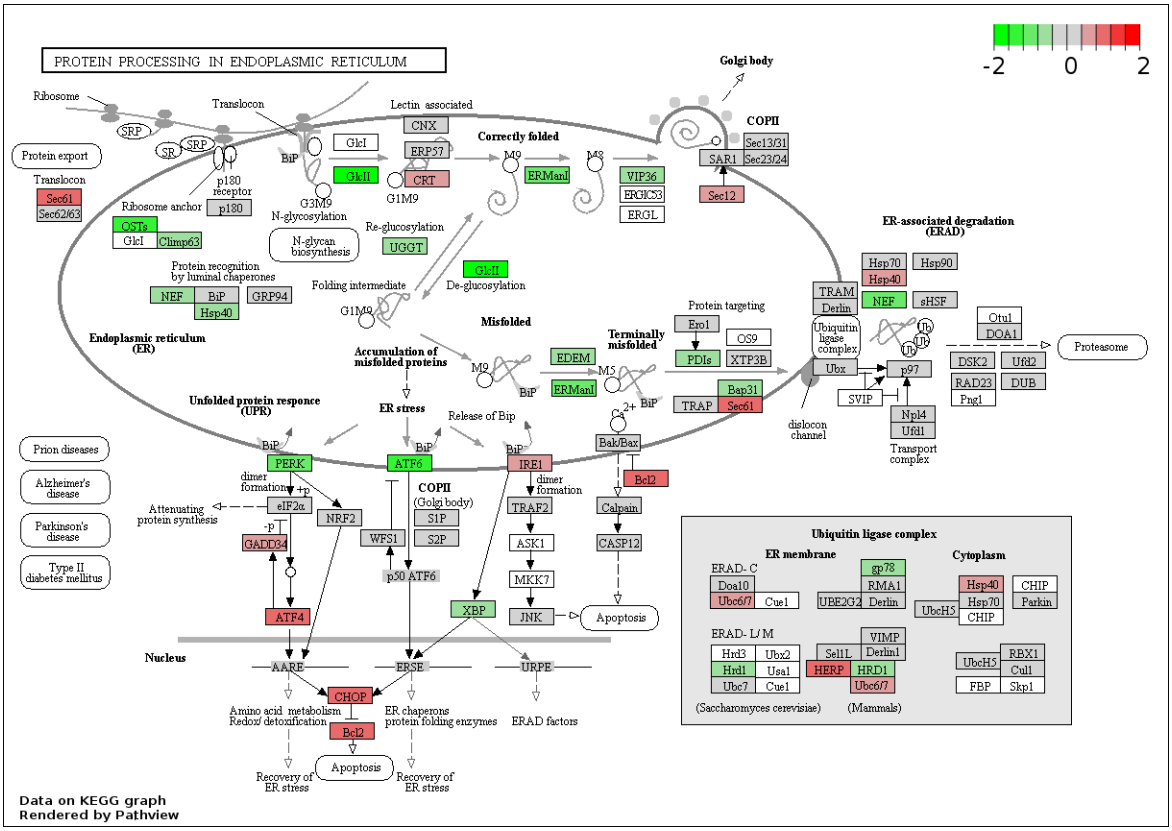


Figure 3.26: Continued on next page

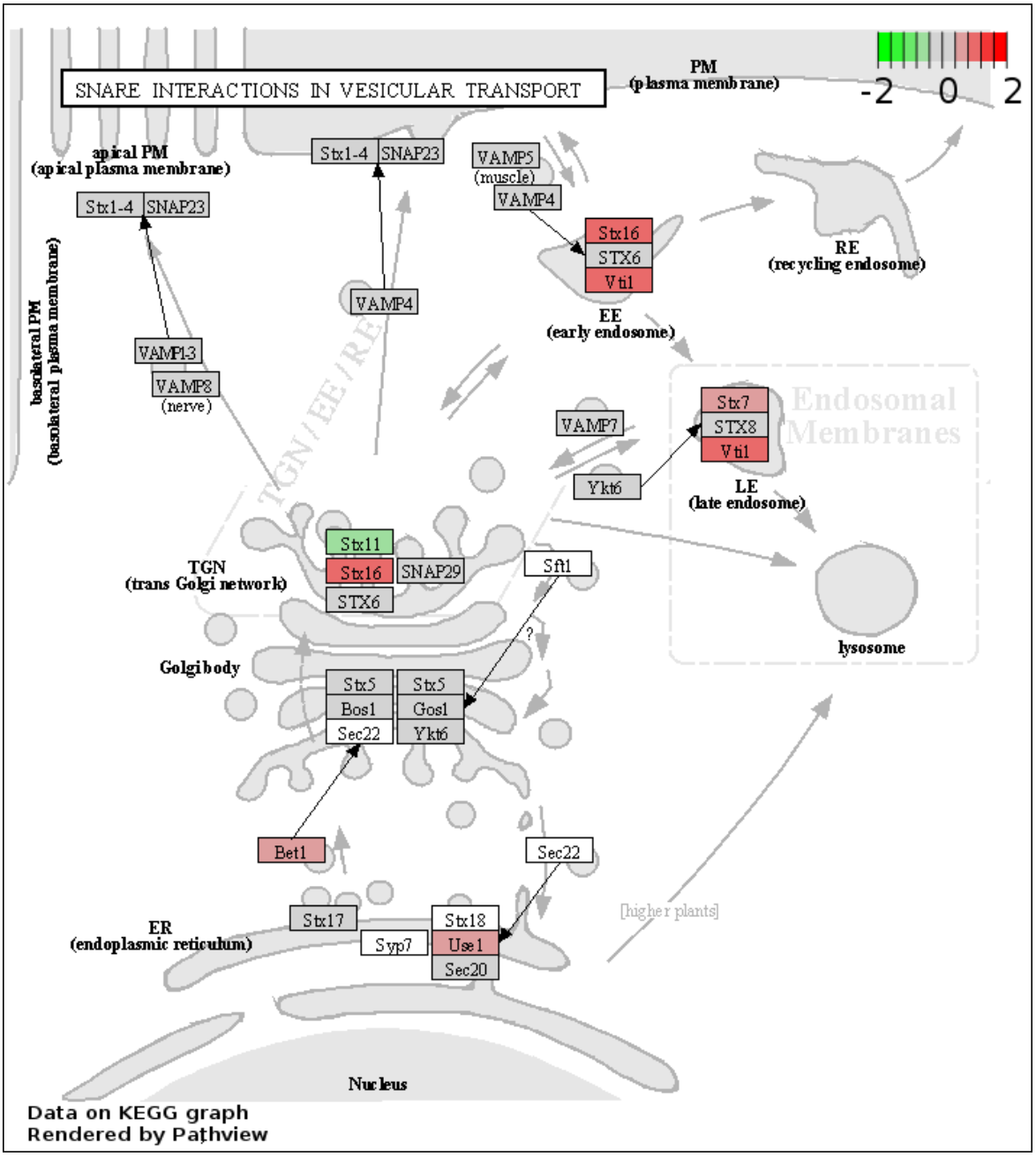
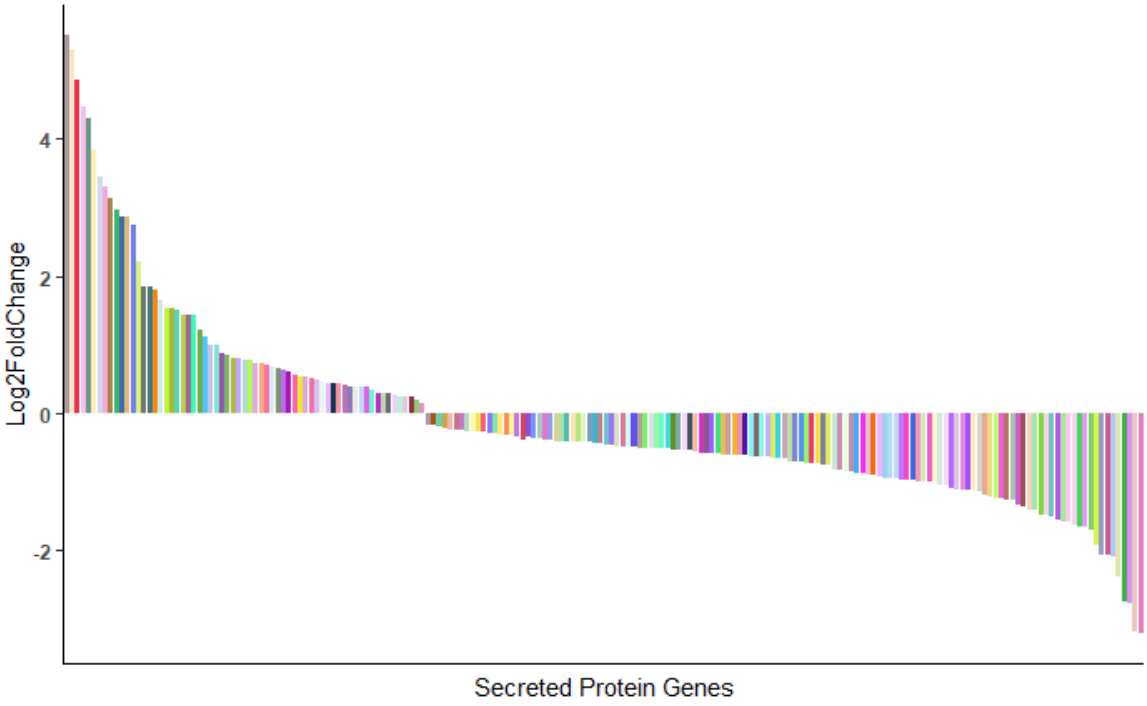


Figure 3.26: Pathways relating to Protein Processing for the producer versus non-producer on day 10. All genes that have been mapped with a fold change have a $p_{adj} < 0.05$ on day 10. There is very little difference between day 5 and day 10 in these pathways. There were some minor changes in genes related to inhibition of apoptosis in protein processing in the endoplasmic reticulum, which may indicate that the cell is trying and failing to auto-regulate itself on day 10 to say alive.

3.7.4 Day 10 Secreted Proteins

Figure 3.27 shows the likely secreted proteins expression on day 10. Unlike both day 5 and day 2, day 10 appears to be in the mid ground of both. It may indicate at this stage of culture, the variation between the producer and non-producer is reducing. This could be due to the producer cells trying to preserve themselves by trying to produce less protein, or it could also be due to the Mock being low on viability at this stage of culture and becoming stressed.



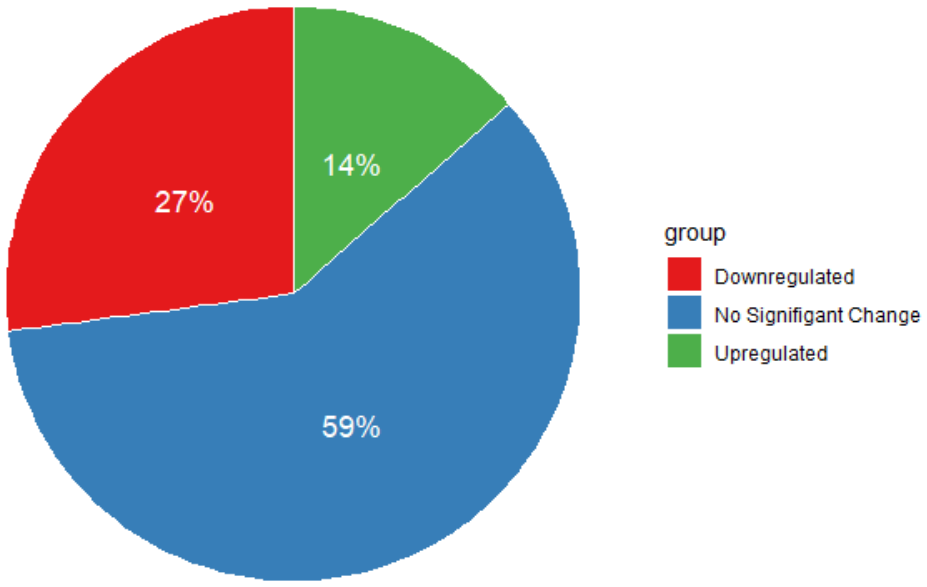


Figure 3.27: A comparison of Secreted Proteins for the producer versus non-producer on day 10. All genes that have been measured have a $p_{adj} < 0.05$ on day 10. Genes that are likely to be secreted show a trend of downregulation in the producers on day 10. Instead of showing a continued downregulation that was seen on day 5 instead, as with many of the pathways looked at on day 10 the overall downregulation of the pathways seems to have been reduced.

3.8 Conclusions from KEGG Pathway Analysis

The comparison of the producer versus non-producer on day 2, day 5 and day 10, has led to some interesting observations. One general observation is from day 2 at which point the clones are at their most similar. It can be seen that transcription has little difference but genes relating to pathways such as protein export and SNARE

interactions in vesicular transport appear to be upregulated. It also appears on day 2 that many of the genes associated with protein processing in the endoplasmic reticulum are downregulated compared to the non-producer, which was unexpected. It was expected that the clones producing an extremely complex large protein would have genes relating to protein processing upregulated as they would have enhanced ER function.

Genes related to protein export are upregulated, while genes that activate the UPR tend to be downregulated to a statistically significant degree. In contrast to this *CASP12* is upregulated in the producing cells which is a gene that is known to induce apoptosis by the caspase pathway due to the unfolded protein response and is triggered by an efflux of calcium in the endoplasmic reticulum (Nakagawa et al., 2000).

Day 5 showed many changes in the producer cells. The most interesting of these was the upregulation of the splicesome and ribosomal proteins. This, as expected, showed that pathways relating to translation were higher in the producer clones when protein was being produced. Genes associated with ribosomal biogenesis and function where also upregulated as expected with the producing cell lines having increased protein production over the non-producing clone. In contrast to day 2, *CASP12* had no significant difference from the non-producer at this stage and instead, *CALPAIN* is downregulated. This could indicate that from day 2 to day 5 the producer cells have depleted their endoplasmic reticulum calcium store for unknown reasons. It could be due to constant UPR induced stress, but this would need experimental testing to prove. It would also be worth testing to see if supplementing calcium to the producer cells would increase growth and antibody titre.

Day 10 was similar to day 5. There was little change between the two days. The only difference was the upregulation of pro-apoptotic genes such as *CHOP*. At this stage, the previously upregulated *CASP12* still has no upregulation, which indicates the level of ER calcium is okay and the ER is no longer pumping calcium into the cytosol. Interestingly, although the cells appear to be receiving stress signals and *BCL2* is upregulated, one of the leading indicators of stress *BIP* has no change in expression.

Overall, from the 3 days of comparisons, some observations can be made which could potentially help in future CHO cell engineering. The hypothesis derived from the analysis are as follows:

- Genes relating to protein synthesis should be increased in the producing clones and this should stay higher in general over the 3 time points compared to the non producing-clones.
- The higher amounts of protein synthesis in these producer clones, does not upregulate the ERAD pathway and instead the producer will have lower base gene expression of genes related to the ERAD.
- Genes related to the unfolded protein response were in the producer clones than in the non-producer, especially on day 2.

- Mitochondrial metabolism increased, potentially due to increased energy demand in the producing clones.
- The increased energy demand of the producer cells have caused a reduction in secreted protein genes and an increase in the lipolysis pathway which may indicate an increased energy demand.

3.9 Investigating Proposed Hypothesis and Potential Targets for Genetic Engineering

This section discusses the overall trends that were seen through the KEGG pathway analysis and analyse the proposed hypothesis to see if under closer inspection they hold true.

The analysis was performed by downloading the genes associated with certain functions from the Mouse Genome Informatics Gene Ontology tool and converting the differential expression data from CHO to Mouse orthologues. This allowed the analysis of all genes associated with specific functions to be sub set and compared. Each functional subset had the statistically significant genes plotted with their Log_2 Fold Change and a pie chart showing the percentage of statistically significant genes upregulated , downregulated and the % of genes that had no change.

3.9.1 Protein Synthesis Genes

Genes related to protein synthesis were expected to be upregulated in the producer clones at all 3 time points. The results of this analysis can be seen in Figure 3.28. On day 2 protein synthesis between the mock and producers shows little difference with slight overall upregulation. Day 5 shows a large change in this trend and the genes associated with protein synthesis are heavily upregulated. This is as expected, as it was seen when looking at the KEGG pathways. Genes relating to translation and ribosomal biogenesis were also upregulated at this stage. Day 10 shows a similar trend, although the amount of upregulation at this stage is reduced, which may indicate a slowing down of protein production.

The trend shows that overall in a producing cell line, genes related to protein synthesis are generally upregulated even on day 2. Albeit not by much. It appears the producer cells, when at their highest points of production, such as day 5 and day 10 have a much higher % of protein synthesis genes being actively transcribed. It could indicate that GS selection with an antibody may select for characteristics in the cell that allow the protein synthesis genes to be upregulated to an unnaturally high degree.

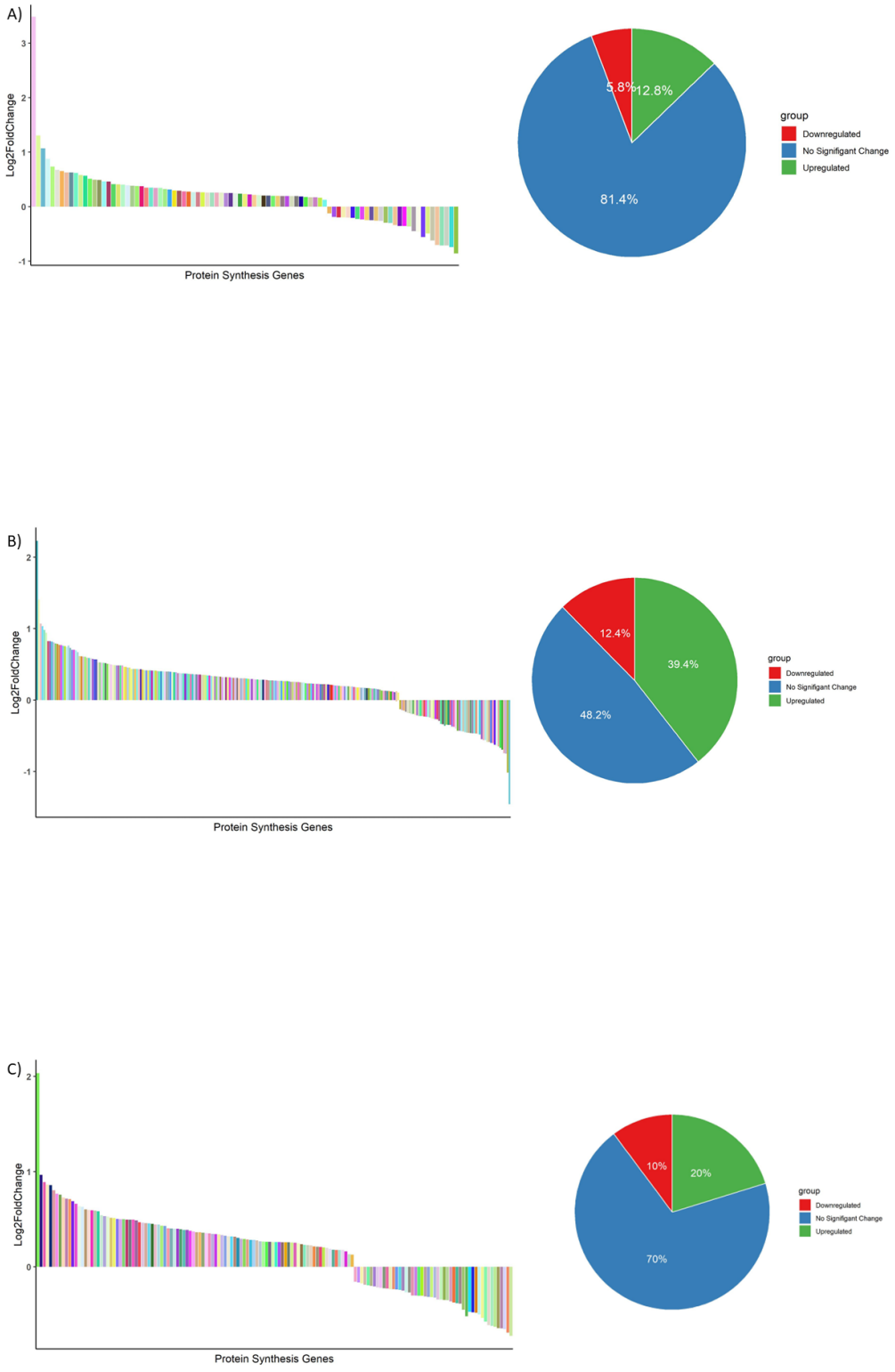


Figure 3.28: The up and downregulation of genes relating to protein synthesis on days 2, 5 and 10 for the producer versus non-producer comparison. All genes that have been measured have a $p_{adj} < 0.05$. Genes associated with protein synthesis are shown on A) day 2, B) day 5 and C) day 10. The comparison is the producer versus non-producing cells. Gene ontologies for the pathway were retrieved from the MGI database.

3.9.2 ERAD Genes

With the increased amount of protein in the ERAD pathway, it is safe to assume that the ERAD should be upregulated as the frequency of misfolded proteins should be increased. Interestingly, from an ERAD gene set of 99 genes, on day 2 the ERAD is quite heavily downregulated compared to the non-producing clone. This could indicate that the sensitivity of the ERAD pathway is basally downregulated in the producing clones and clonal selection criteria pick cells with a dysregulated ERAD, this can be seen in Figure 3.29. On day 5 the recombinant cells are producing more protein and as such, one would expect genes associated with the ERAD to be increased and they are substantially from 9% to 24.4%. However, the downregulation still outweighs upregulation, moving from 25% to 36% on day 5. The gene which is most heavily upregulated by 1.4 Log2FoldChange is *HERPUD1* which is interesting as it is thought to inhibit translation in response to ER stress. Still, it has also been linked to being required for efficient degradation of *CD3D* via the ERAD pathway.

Again, day 10 then looks like a midpoint between day 5 and day 2. The downregulation and upregulation mimic the trend seen on day 5 but with reduced effect, with only 29% of genes downregulated and 14% upregulated.

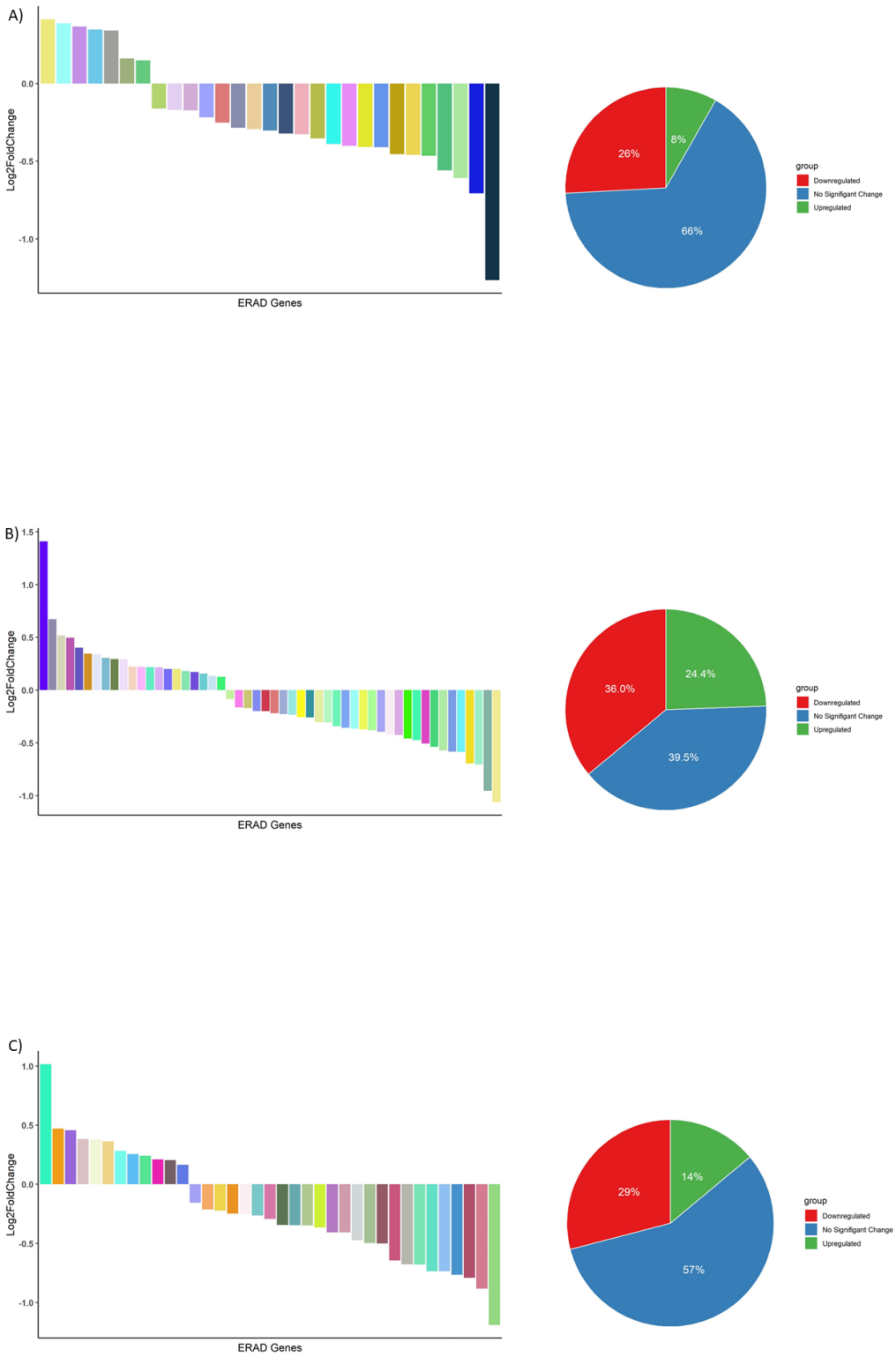


Figure 3.29: The up and downregulation of genes relating to the ERAD pathway on days 2, 5 and 10 for the producer versus non-producer comparison. All genes that have been measured have a $p_{adj} < 0.05$. Genes associated with the ERAD are shown on A) day 2, B) day 5 and C) day 10. The comparison is the producer versus non-producing cells.The comparison is the producer versus non-producing cells. Gene ontologies for the pathway were retrieved from the MGI database.

3.9.3 Unfolded Protein Response

Genes associated with the unfolded protein response (UPR) have been split into two subsets. These are the cellular response to misfolded proteins and the genes related to the UPR pathway.

The cellular response to misfolded proteins indicates how the cell reacts to the presence of proteins that are not folded correctly and in an overloaded endoplasmic reticulum, the chances of misfolded protein would increase considerably.

Figure 3.30 shows the differential gene expression for producer versus non-producer for the misfolded protein gene set. This gene set was only small, with 24 genes in it. But it does appear the response even to misfolded proteins is quite heavily downregulated on day 2, with 35% of genes being downregulated and only 4% being upregulated. Day 5 still shows significant downregulation, but genes such as *KLHL15* get upregulated to a maximum Log2 Fold Change of 0.27. This gene is thought to have a role in protein ubiquitination.

SDF2 is the most downregulated on both day 2 and day 5, although its downregulation increases slightly on day 5. The function of this gene is currently unknown. On day 10, when the cells have lost a lot of viability, the differences between the producer and non-producer retain a similar trend to what was seen on day 2. Indicating even when the cells have had their apoptotic pathways triggered, they still fundamentally have some genes up or downregulated over the non-producer. This again adds to the theory that GS selection with an antibody may fundamentally select for a cell that has a dysregulated response to the stresses of producing a recombinant antibody.

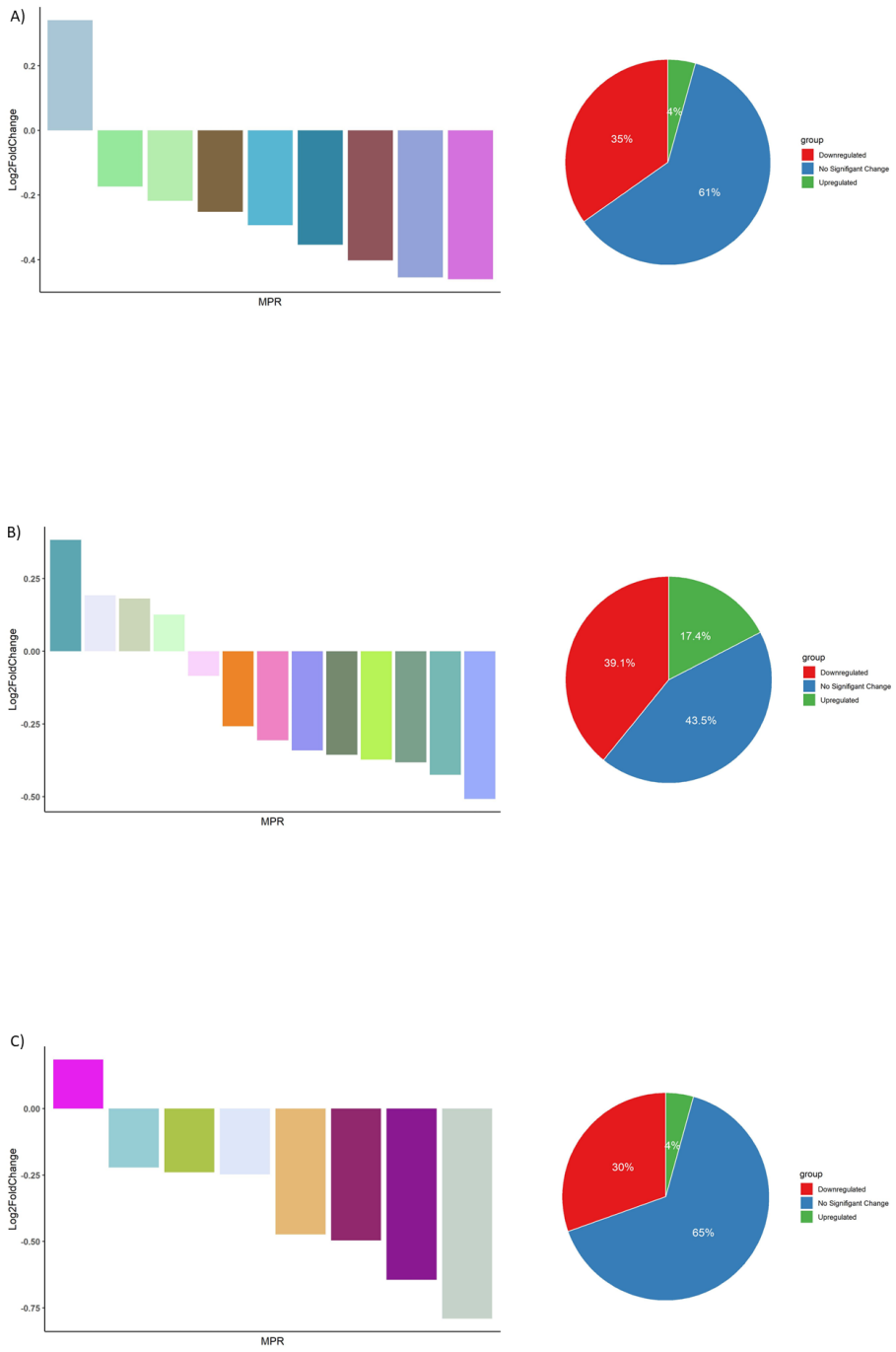


Figure 3.30: The up and downregulation of genes relating to the cellular response to misfolded protein on days 2, 5 and 10 for the producer versus non-producer comparison. All genes that have been measured have a $P_{adj} < 0.05$. Genes associated with the cellular response to misfolded proteins are shown on A) day 2, B) day 5 and C) day 10. The comparison is based on producer versus non-producer. The comparison is the producer versus non-producing cells. Gene ontologies for the pathway were retrieved from the MGI database. MPR stands for misfolded protein response.

The unfolded protein response did not follow the initial hypothesis that it would automatically be upregulated over the non-recombinant cell. Instead, it showed approx-

imately equal up and downregulation on day 2 with 14.3% downregulation and 13% upregulation. Upregulated genes of note are *PIDA6*, which is linked with inhibiting the aggregation of misfolded proteins and *BAX* which has been shown to be both pro and anti-apoptotic depending on its ratio with *BCL2*.

Day 5, as shown in Figure 3.31, looks more like the initial hypothesis expected. The unfolded protein response appears to be disproportionally upregulated compared to the non-recombinant cell line. Interestingly, *PDIA6* shows no significant difference on day 5. The most upregulated genes *DDIT3*, *PACRG*, *HERPUD1*, *ATF3* and *ATF4* all indicate endoplasmic reticulum stress and in particular, ATF4 has been shown in association with CHOP to induce ER-mediated cell death. Thus, the endoplasmic reticulum appears to be stressed, but protein folding/assembly machinery, as shown in the previous section, is downregulated. This could indicate that the ER is chronically stressed to the point of ceasing function. This contradicts the protein titre which shows increasing titres later into culture, even with a dysregulated ER. Interestingly the downregulated genes don't change much but instead the most downregulated genes *EIF2ak2* and *ATF6b* increase in downregulation from 0.75 to over -1. This is interesting as *ATF6B* is a transcription factor linked with inducing ER stress. As the cell produces more protein, it's even further downregulated, which is contradictory to the upregulated genes but may still indicate a dysregulated ER function in the recombinant cells. The downregulation of *HSPA2* was unexpected as it is linked with protection from ER stress. The trend on day 10 is similar to day 5. *ATF6b* is still the most heavily downregulated gene and *PACRG*, *ATF3* and *ATF4* are the most upregulated. *XBP1* is also downregulated on day 10 which was unexpected as it is thought to play a role in the recovery of ER stress but this may indicate a functional shutdown of the ER. On day 5 the gene had no significant difference between the producer and non-producer cell line.

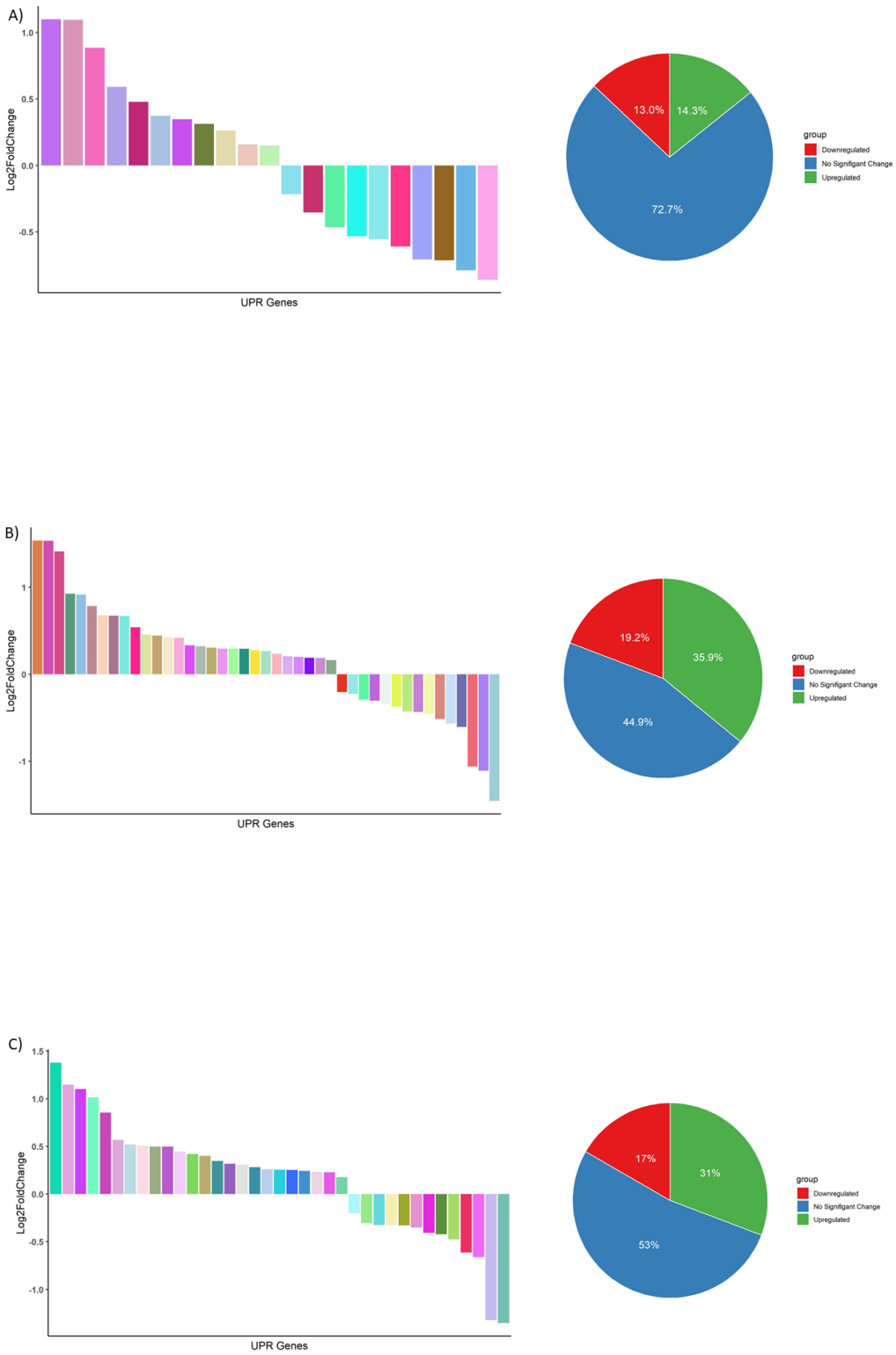


Figure 3.31: The up and downregulation of genes relating to the UPR on days 2, 5 and 10 for the producer versus non-producer comparison. All genes that have been measured have a $p_{adj} < 0.05$. Genes associated with the unfolded protein response are shown on A) day 2, B) day 5 and C) day 10. The comparison is based on producer versus non-producer. The comparison is the producer versus non-producing cells. Gene ontologies for the pathway were retrieved from the MGI database.

3.9.4 Oxidative Stress

Figure 3.32 shows the cellular differences in relation to oxidative stress genes. Day 2 shows only 24% total variation between the producer versus non-producer cell line. Day 5 shows a massive 57% difference meaning the recombinant antibody producing cells have much change in their oxidative stress pathway. On days 2 and 5 the amount of regulation appears to be equal. The *ALDH3B1* gene is constitutively downregulated in the recombinant cells, being the second most downregulated on day 2 and most downregulated on day 5.

This is interesting as it is a gene which allows detoxification of aldehydes generated by alcohol metabolism and may be linked to protection from oxidative stress. Another largely downregulated gene on both days is *PRDX3*, which has an antioxidant function. This may indicate that instead of having increased protection from oxidative stress. These recombinant cells have largely given up control of the oxidative environment within the cell. Although genes such as *ORX1* and *SESN2* which have a fold change of 0.8 and 0.6 on day 5 have also been implicated with resistance to oxidative damage. Overall from this data, no solid conclusion or hypothesis can be drawn as there is no clear indication of what is occurring.

As with many previous pathways, day 10 shows little difference to day 5. The most interesting change is the increase of the pro-survival gene *BCL2*, which increased from a fold change of 0.4 on day 5 to 1.16 on day 10. This indicates that although the cells are reducing in viability and the non-producing cells are also reducing in viability, the cells are fighting to try and survive.

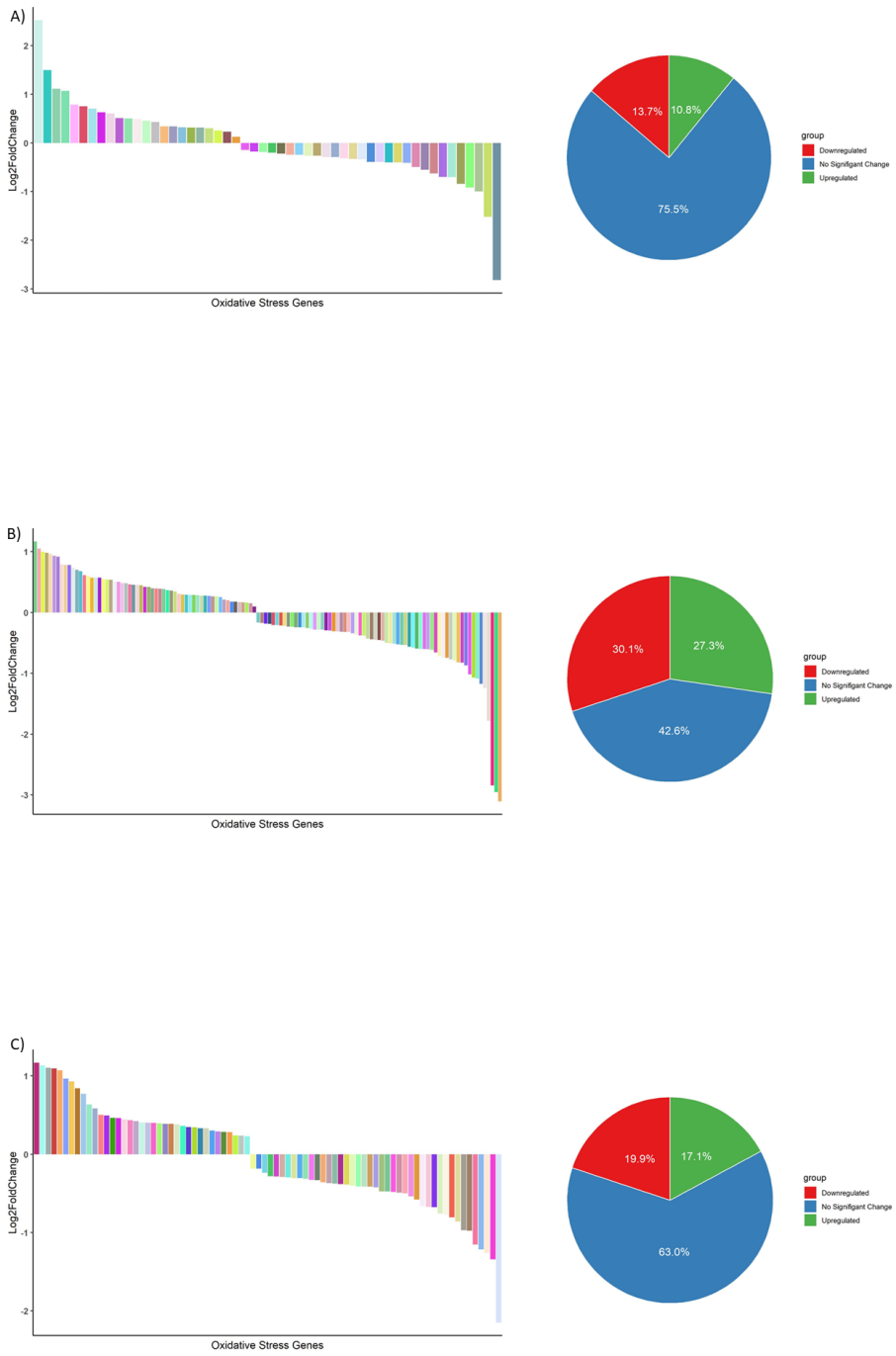


Figure 3.32: The up and downregulation of genes relating to oxidative stress on days 2, 5 and 10 for the producer versus non-producer comparison. All genes that have been measured have a $p_{adj} < 0.05$. Genes associated with oxidative stress are shown on A) day 2, B) day 5 and C) day 10. The comparison is based on producer versus non-producer. The up and downregulation appears to be relatively equal on days 2, 5 and 10. The comparison is the producer versus non-producing cells. Gene ontologies for the pathway were retrieved from the MGI database.

3.9.5 Mitochondrial Metabolism Genes

Producing a recombinant protein, in theory, would take up more energy than a cell producing no recombinant protein. For that reason, Figure 3.33 shows the differential expression on day 2, day 5 and day 10 for genes relating to mitochondrial metabolism.

Interestingly, against the previous hypothesis, there is not a large difference in oxidative phosphorylation between the producer and non-producing cell lines on day 2. On day 5 there are more significant genes both up and downregulated. For oxidative phosphorylation, more genes are downregulated than up. 26.7% of genes are downregulated and only 18.9% are upregulated. Interestingly on day 5 and day 10, many of the genes associated with ATP such as *ABCD1*, *ATPSCKM*, *ATP5J2*, *ATP5MG* and *ATP5PD* are downregulated. This could indicate a complete shutdown of the metabolism of the cells. There is also an increase in genes such as *MLXIPl*, which is linked with high glucose concentrations on day 5 and day 10. This is interesting as it may indicate that although glucose is present in the media the cells are unable to utilise it.

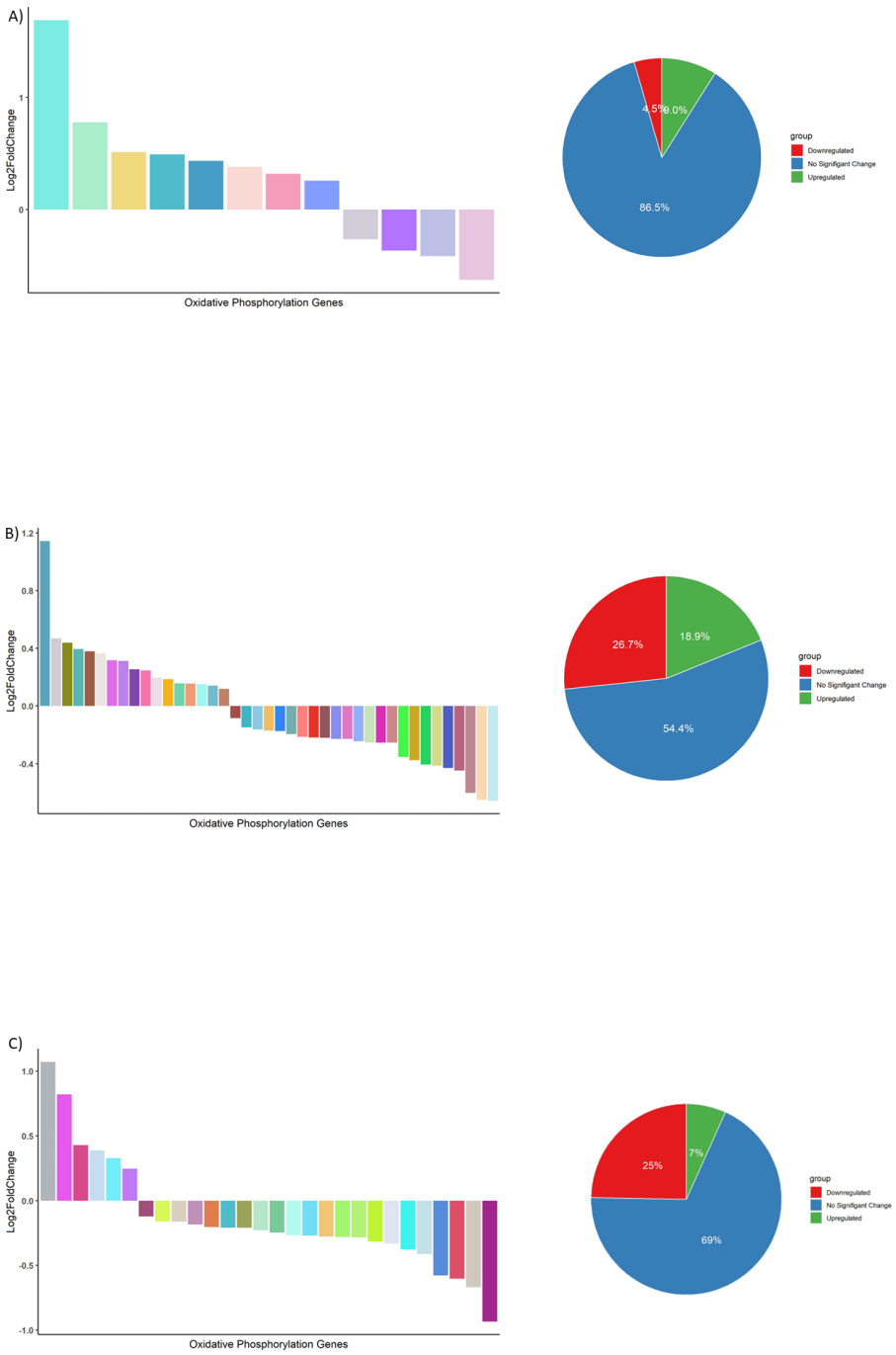


Figure 3.33: The up and downregulation of genes relating to mitochondrial metabolism on days 2, 5 and 10 for the producer versus non-producer comparison. All genes that have been measured have a $p_{adj} < 0.05$. Genes associated with mitochondrial metabolism are shown on A) day 2, B) day 5 and C) day 10. The comparison is from producer versus non-producer. The comparison is the producer versus non-producing cells. Gene ontologies for the pathway were retrieved from the MGI database.

3.9.6 Fatty Acid Metabolism

Figure 3.34 shows day 2 fatty acid metabolism with approximately a 28% difference in the producer versus non-producer. The number of genes that are upregulated and downregulated appear in equal amounts. Interestingly, one of the most downregulated genes for producing clones is *RARRES2* which encodes the Retinoic Acid Receptor. This usually functions in lipolysis.

Day 5 showed a large difference for the producer versus a non-producing clone. There is 33% of genes involved in fatty acid metabolism are downregulated versus 20% upregulated. The most upregulated gene is *EDN1* which is involved with angiogenesis, while *RARRES2* is still the most downregulated.

Day 10 shows little difference from day 5, the downregulation of *RARRES2* is conserved and no genes of note could be seen or trends with similar functions.

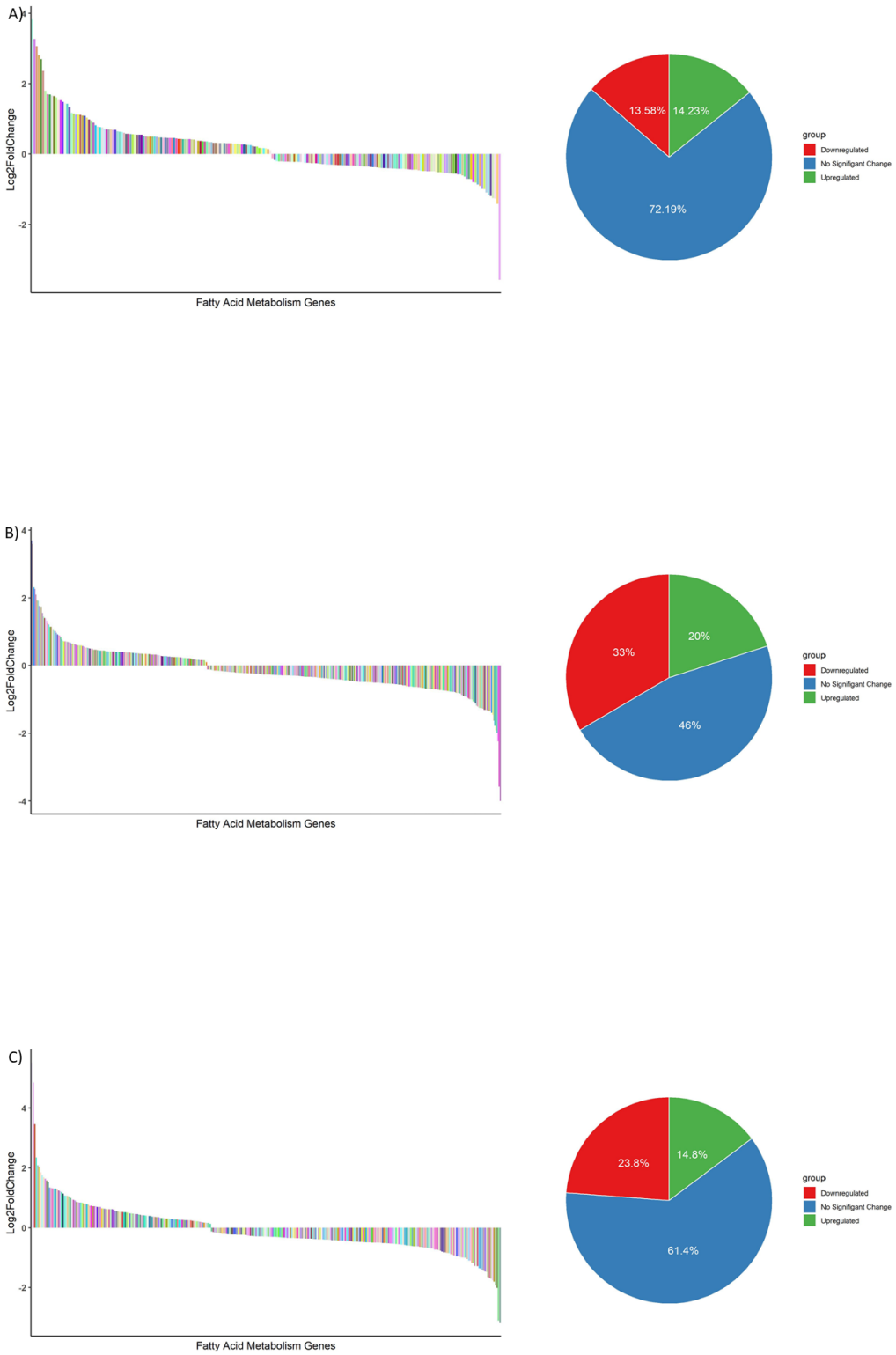


Figure 3.34: The up and downregulation of genes relating to fatty acid metabolism on days 2, 5 and 10 for the producer versus non-producer comparison. All genes that have been measured have a $p_{adj} < 0.05$. Genes associated with fatty acid metabolism are shown on A) day 2, B) day 5 and C) day 10. The comparison is from producer versus non-producer. The comparison is the producer versus non-producing cells. Gene ontologies for the pathway were retrieved from the MGI database.

3.10 Overall Hypothesis and potential targets

Pathways involved with protein production were analysed, from transcription to protein export. The two producer clones were pooled to compare the differences between a producer versus a non-producer. This gave interesting insights into how producing a recombinant protein changes the CHO cells on different days of culture. Some hypotheses were generated, as discussed in Section 3.8. The KEGG conclusions fundamentally showed that selection while producing a recombinant protein appears to select for a clone with dysregulated responses to misfolding, protein processing and degradation. It could potentially also select for more efficient translocation within the cell.

The resulting proposed hypothesis from the KEGG pathway analysis was investigated in greater detail. The genes associated with each pathway were taken and the overall trend of how many genes were statistically significantly up and downregulated was analysed. As was expected, genes relating to protein synthesis on day 2, day 5 and day 10 were upregulated over the non-producing clone, indicating increased utilisation of the pathway. This was also paired with everything to do with tRNA being upregulated, which makes sense, considering tRNA is an essential part of translation.

The ERAD pathway showed, as expected, fundamental downregulation in the producing cells. This was expected from the results of the KEGG pathway but was unexpected in terms of biological context. as it would help the cell produce more correctly folded protein if the ERAD was upregulated. It would also prevent the initiation of the UPR response. The downregulation of the ERAD, along with the KEGG pathways showing general downregulation of protein processing, may insinuate that the ER functions slower in the producing clone and there is less misfolded protein.

The unfolded protein response showed some fascinating results. The cellular response to unfolded protein is overall downregulated in the producing cells. In contrast, the actual genes associated with the unfolded protein response were shown to have upregulation, especially on day 5. This may indicate that the UPR itself is not dysregulated but the cell just ignores the UPR and misfolded proteins or perhaps has a higher threshold that has to be reached on day 5 to activate the apoptotic pathways in the producing clones.

Oxidative stress was looked at to answer the question of if the protein-producing cells are stressed or perhaps have a lower base level of genes expressed related to oxidative stress. The results from this section were unclear as there was equal up and down-regulation of genes on day 2, 5 and 10. Some of the genes upregulated in the producer cells were linked with survival and oxidative stress protection but there were no obvious trends in the types of genes up or downregulated.

Finally looking at the mitochondrial metabolism and fatty acid metabolism the aim was to see if the producing cells have a higher energy demand compared to their non-producing counterparts. Interestingly, genes related to both mitochondrial metabolism and fatty acid metabolism show downregulation which might indicate that the producer cells are using less energy. Overall, it seems that producer cells are generally

downregulated and maybe the slower growth, lower energy requirements etc help the cell produce more correctly folded protein.

Taking all of this into account and looking throughout the pathways, Table 3.2 shows potential genes that could be targeted for genetic engineering for CHO cells in the future.

Table 3.2: Genes found throughout the analysis that may have interesting outcomes if over-expressed or under-expressed for CHO cell engineering.

Gene	Function
PABP	Upregulated in producer clones. Protein binds to the poly A tails and promotes translation initiation.
eEF1A	Upregulated in producer clones. Responsible for enzymatic delivery of tRNA to ribosomes.
SRPR	Upregulated in producer clones. Part of the signal recognition particle receptor and is involved in translocation which is generally upregulated in producer clones.
SEC61	Helps to move polypeptides into the ER. Upregulated in producer cell lines.
SEC31	Function is not fully know but in yeast the Sec13 protein is required for vesicle biogenesis in the ER. Upregulate.
SEC23/24	Similar to Sec61, except it is found in the ribosome free space of the ER. Upregulate.
HSPA8	Multifaceted protein with many functions including cellular protection, folding, transport and it plays a role in protein quality control. This gene was not significantly upregulated on any day but on day 2 had a p adjusted value of 0.06. Upregulate
LSCS	Enzyme that delivers sulfur to partners for fe-S cluster assembly and tRNA. Upregulation may help tRNA synthesis.
TUM1	Involved in the modification of some tRNAs. Upregulation may be benifical.
FTH1	Involved in iron storage in eukaryotes. Upregulated in Producer clones.
CALR	Involved in maintaining the correct amount of calcium within the ER. Upregulation may further help the ER retain homeostasis.
CASP12	Gene can cause apoptosis in response to ER stress. Downregulation may lead to increased UPR threshold before apoptosis is triggered.
CANX	Involved with quality control of misfolded proteins. Upregulation may increase protein quality control.
PP2A	Growth suppressor. Downregulation could lead to increased biomass.
TCOF1	Upregulation of this gene may help in the production of tRNA.
REL1	Upregulation may help in increasing RNA editing in pre-mRNAs.
CALPAIN	Upregulation may again help calcium homoeostasis in the ER.
CHOP	Downregulation may inhibit apoptosis and induce a hypo oxidizing ER that may reduce abnormally high molecular weight complexes.
STX5	Downregulation may inhibit the process of autophagy/apoptosis.

Chapter 4

Finding New Synthetic Promoter Building Blocks

Overview

- This chapter discusses how a bioinformatic workflow was designed which allowed for quick and robust TFRE discovery.
- Section 4.2 discusses the first generation TFRE discovery pipeline, along with design considerations employed and the resulting outcomes.
- Sequence variant testing was investigated to assess if the inclusion of these variants would be beneficial for TFRE discovery. The results indicated small changes to a sequence have a significant impact on SEAP production. The results are discussed in Section 4.3.
- The second generation pipeline is discussed in Section 4.4, along with revised design considerations based on the outcomes from pipeline 1 and variant testing, resulting in the discovery of a new TFRE NFE2l2.
- The important aspects of TFRE discovery, along with considerations for automation are discussed in Section 4.5, with particular attention to potential future works for expanding TFRE discovery.

4.1 Introduction

Previous studies for synthetic promoter engineering have focused on using RNA-seq data to derive information from the CHO cells transcriptome (Johari et al., 2019; Brown and James, 2016). These works focused on sub-setting data manually and using small data type approaches to analyse the transcriptomic datasets.

Advancing on this methodology, there were two primary objectives. The first, was to create an automated system to analyse RNA-seq data, producing a final selection panel of TFREs to test. Through enhanced automation over previous works, the methodology of finding new TFREs could be expedited and potential human bias removed.

The second objective was to find new TFREs for the synthetic promoters to expand the available design space. This would be beneficial for IP purposes and to increase synthetic promoter design complexity.

4.2 TFRE Discovery Pipeline 1

Due to Genomatix containing a database of curated mouse promoters and mouse being taxonomically similar to CHO, it was hypothesised that using the mouse promoter database may be beneficial with little information lost. The use of this database would avoid the issue of no CHO promoter database existing. The alternative was to use the same protocol described in Johari et al. (2019).

To decide if mouse or CHO promoters should be used, promoters were analysed using Genomatix Matinspector to compare their similarity. The top 100 promoters, ranked by their TPM were taken forward for analysis. Initially, CHO promoters were defined as 1000bp upstream of the 5' UTR and 200bp downstream, totalling 1200bp. CHO with these regions had approximately three times fewer binding sites when analysed using Genomatix MatInspector. This indicated that the 1200bp regions used in previous studies may have excluded large portions of the typical promoter region. This was increased to a promoter length of 3600bp based on the amount of TFREs found to bind the promoter regions in CHO and mouse using Genomatix MatInspector. When 3600bp (3400bp upstream of the UTR and 200bp downstream) CHO promoters were analysed, a total of 80514 total TFRE binding sites were found in the top 100 genes promoter regions. Mouse had 96377 TFRE binding sites in the top 100 genes promoter regions. This indicates that on average the curated mouse promoter dataset has promoter lengths on average longer than 3600bp as the experimental mouse promoter regions on still had more observations then the 3600bp CHO promoter regions.
If mouse and CHO were similar, the curated mouse promoter database was readily accessible for the genes in the RNA-seq dataset and contained experimentally proven promoters. The CHO promoters had to be manually uploaded in blocks of 50-100 sequences and results were collated as Genomatix did not have the organism in its database. The use of the mouse promoter database would allow for less prior setup, but could have led to an increased rate of false positive TFREs being selected.

Table 4.1 shows how the TFRE frequencies were ranked as they occurred in CHO and mouse promoters once analysed in Genomatix Matinspector. There was a 22.52% difference between CHO and mouse across the top 100 promoters in relation to Genomatix matrices. If TFRE families were considered instead the difference decreased to 14.89%.

The percentage of each TFRE accounted for in the overall population of mouse and CHO TFREs was calculated by taking the observed/total and mutliplying by 100. The differences between CHO and Mouse in the percentages were then calculated by deducting the CHO and mouse TFRE values. Finally, to calculate the percentage difference totals, the absolute of these values were summed together. Based on these results 3600bp CHO promoters were carried forward for analysis. The code for this analysis is shown in Appendix B.7.

Table 4.1: Analysis of CHO and mouse promoters in Genomatix Matinspector showed a large difference in TFRE positions and overall percentage similarity. The table shows the matrix names in the first column, followed by the frequency of the TFREs in CHO and then mouse. The CHO and mouse TFRE percentages are the % that individual TFRE has made up of the total TFRE population. The Percentage difference is how big of a difference there was between CHO and mouse. Finally, the Percentage Difference Total is the absolute sum of the Percentage Difference Column. This showed a 22.52122% variation between CHO and mouse in this instance.

Matrix	Frequency in CHO	Position in CHO	Frequency in Mouse	Position in Mouse	CHO TFRE Percentage	Mouse TFRE Percentage	Percentage Difference	Percentage Difference Total
V$FOXP1_ES.01	628	1	473	1	0.780735	0.58806	0.192675	22.52122
V$GAGA.01	582	2	153	161	0.723548	0.190218	0.53333	22.52122
V$LMX1A.02	434	3	353	5	0.539553	0.438869	0.100684	22.52122
O$SPT15.01	433	4	238	41	0.538309	0.295895	0.242415	22.52122
O$PTATA.02	407	5	245	35	0.505986	0.304598	0.201388	22.52122
V$NKX61.01	322	6	288	18	0.400313	0.358058	0.042256	22.52122
V$ZBED4.02	322	7	351	6	0.400313	0.436383	0.036069	22.52122

Taking Table 4.1 into consideration pipeline 1 progressed using CHO promoter regions. The difference between mouse and CHO was to great. The 24,000 genes promoter sequences were acquired from ENSEMBL. The files were split into fasta files of approximately 50 sequences in number. To overcome the issue of manually uploading hundreds of separate fasta files a macro was created to automatically upload all of the fasta files and initiate the separate Matinspector analysis on each fasta file. The macro automated clicking a mouse in certain positions and moving through the split sequence files one by one. To ensure all sequence files had been analysed they were numbered in sequential order to check if any had been missed. The Matinspector tool was run with the settings of Core similarity = 1 and matrix similarity = Opt +0.01. This in general terms means the core binding sites of the TFRE had to be an exact match and the less influential parts of the binding were also more strict. All other settings were left as default.

The results were collated together using R and a novel analysis was carried out, measuring the abundance of TFRE sites in the promoter regions, the high expressing groups promoter regions were normalised against the average group. This was achieved through measuring the overall amount of TFRE binding sites in the promoter regions of genes that had a average TPM of over 1000 across the samples on day 2,5 and 10. Gene that were statistically significantly upregulated genes on day 5 were also analysed. Day 5 was taken in an attempt to find TFREs that may confer long term activity in culture. The number of observations was the number of overall TFREs reported to bind For instance, if 1000 TFREs bound 3 times each, the total observations accounted to 3000.

The same number of observations from the high expressing or upregulated genes was taken 5 times from the average group and then the average of these 5 samples was taken. This was an attempt to get an unbiased average of how often TFREs occur in the promoter regions of the CHO genome. The frequency in the high expressing group was deducted from the frequency of the average group to create a graph which shows the enrichment of certain TFREs in the high expressing or the day 5 group. The high expressing group was considered any gene with a TPM of over 1000 or a statistically significant upregulation on day 5.

Figure 4.1 shows what this normalisation technique yielded. This is in contrast to Johari et al. (2019), who looked at the presence or absence of TFREs. It was found that no TFREs could be determined by presence or absence. All of them appeared in some number, even when the high group of TFREs is not taken into account. The prevalence of insulator transcription factors such as v$GAGA.01 was interesting. This is one of the most abundant binding sites in high expressing genes and may indicate insulation from the genome is a key factor in high expressing genes. The code for producing this output is shown in Appendix B.4.

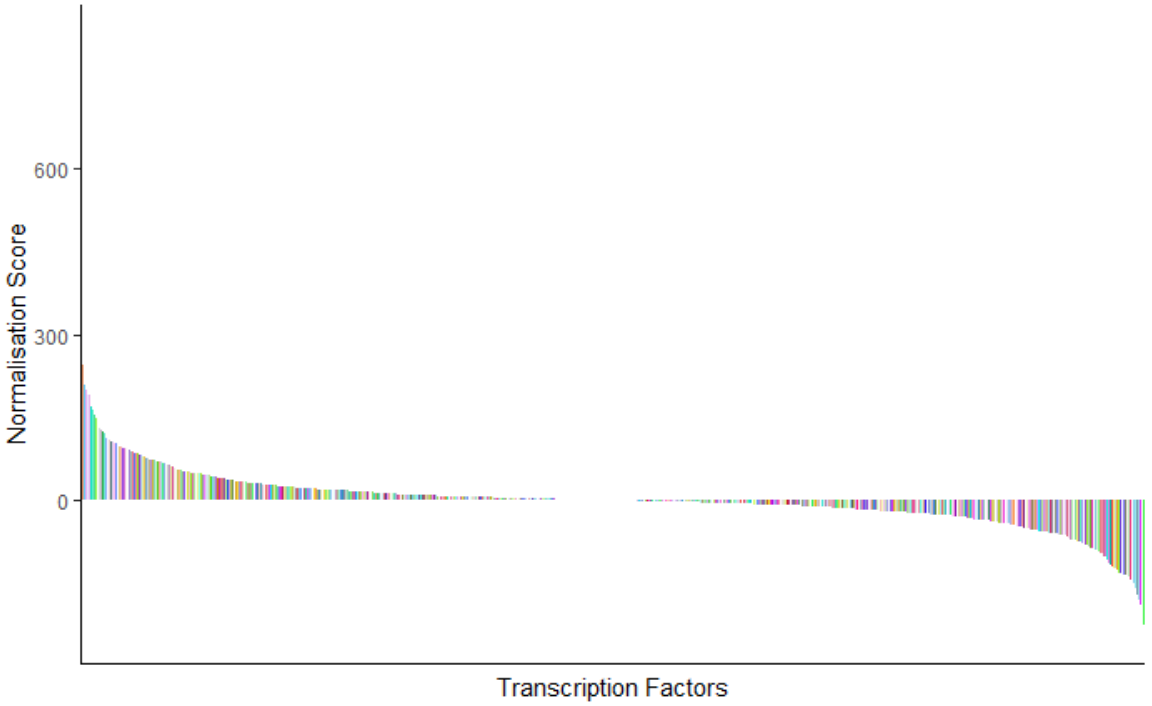


Figure 4.1: An analysis of the frequency of TFREs in promoter regions of high expressing CHO genes. The x-axis shows the different TFREs and the y-axis is the normalisation score which is calculated by taking the frequency a TFRE occurs in the high expressing group and deducting it from the average group. For instance, a normalisation score of +100 would indicate the TFRE is over-represented in the high expressing group.

Due to the large number of over-represented TFRE sequences found, extra manual filters were applied to reduce the selection panel size. These were as follows:

- The TFREs should have literature evidence of potentially positively affecting transcription.
- Families previously tested in Brown et al. (2014) or Johari et al. (2019) should not be tested.
- Avoid using sequences with extremely high similarity to those tested previously. At least a 2 base pair difference is required.

The literature screen and manual filters led to a final panel of 23 homotypic sequences to test. The name and sequence blocks used are shown in Table 4.2. The sequences which were tested are shown in Appendix A.2.

Table 4.2: TFREs used in TFRE discovery pipeline 1. The first column shows the matrix names. The second column shows the sequences used and the third column shows which analysis the sequence blocks originated from. TPM indicates the TFRE blocks were selected from the comparison of high expression genes promoter regions compared to the average and D5_versus_D2 is from the analysis of statistically significant promoter regions upregulated on day 5 of culture. This was done to find TFREs which conferred transcriptional activity and stable expression.

Name	Sequence Block	Analysis Origin
V$GAGA.01	GGGAGAGAGAGAGAGAGAGAGAGA	TPM
V$AHRARNT.03	AGTGCGTGGGAA	TPM
V$SMAD.01	TGTCTGGCT	TPM
V$NRF1.01	GCCGCGCATGCGCATC	TPM
V$E2F2.01	AAGGCGCGCGA	TPM
V$CSRNP1.01	AGAGT	TPM
V$NFIB.01	CCTGGCTCCGTGCCAGCT	TPM
V$CTCF.01	CTCCCCGGCCGCTAGGGGGCGGGC	TPM
V$EGR2.01	TTGCGTGGGCGT	TPM
V$MAZR.01	TGGGGGGGGGCCA	TPM
V$RREB1.01	CCCCAAACCACCCA	TPM
V$BACH2.01	CGTGAGTCATC	TPM
V$MEIS1A_HOXA9.01	TGACAGTTTTACGA	D5_versus_D2
V$TEAD4.01	CTGCATTCCTCA	D5_versus_D2
V$MIT.01	GAGATCATGTGATGA	D5_versus_D2
V$KLF2.01	AGGGGTGGGG	D5_versus_D2
V$NANOG.01	TACTCATTCATT	D5_versus_D2
V$HIC1.01	TTATGCCAACCTA	D5_versus_D2
V$YB1.01	CTGATTGGCCAA	D5_versus_D2
V$BARBIE.01	AGCTAAAGCAGGAGG	D5_versus_D2
V$GATA4.01	AGAGATAAGAT	D5_versus_D2
V$SOX6.01	TCCTTTGTCT	D5_versus_D2

Sequences were synthesised as described in Section 2.2.1 of Materials and Methods. Sequences had 6 repeats of the same TFRE block with what is called a spacer in between as described in Johari et al. (2019). Spacers are used to reduce the amount of non-specific binding in the homotypic promoter. The spacers "AA", "TA" and "TT" were tested, while spacers containing Guanine or Cytosine were avoided as most of the constructs already had high GC content. This was taken from Johari et al. (2019) and was an effort to reduce DNA synthesis costs. Table 4.3 shows the results of running these sequences through Genomatix MatInspector with settings of core similarity of "1" and matrix similarity of "Opt +0.01" The total number of matches for all sequences was then compared and the spacer which had the lowest total number of matches were used for all sequences to keep the experiment consistent.

Table 4.3: The in-silico testing of different 2bp spacers between TFREs indicated the spacer "TA" reduced non-specific binding for library 1. The first column shows the TFRE matrix names from Genomatix. The subsequent columns show the number of matches found when running each sequence and spacer combination through Genomatix Matinspector with settings of "Core = 1" and "matrix similarity = Opt 0.01".

Name	AA	AT	TA	TT
V$GAGA.01	65	60	60	55
V$AHRARNT.03	35	20	40	50
V$SMAD.01	11	7	11	17
V$NRF1.01	31	36	36	46
V$E2F2.01	38	28	43	38
V$CSRNP1.01	2	1	5	1
V$NFIB.01	21	21	26	21
V$CTCF.01	64	59	49	49
V$EGR2.01	56	21	21	16
V$MAZR.01	74	124	94	68
V$RREB1.01	33	33	23	33
V$BACH2.01	24	29	24	29
V$MEIS1A_HOXA9.01	64	44	39	39
V$TEAD4.01	15	24	9	24
V$MIT.01	23	38	13	28
V$KLF2.01	46	46	36	36
V$NANOG.01	108	88	43	48
V$HIC1.01	35	75	30	36
V$YB1.01	25	25	25	25
V$BARBIE.01	18	8	8	8
V$GATA4.01	31	16	17	21
V$SOX6.01	12	12	22	22
Total	831	815	674	710

A considerable issue with creating a homotypic promoter which has no secondary binding is that many TFREs have similar, if not identical binding sites across TFRE families. This spacer exercise, therefore, was an attempt to reduce non-specific binding. Without testing each sequence with CHIP-seq and TF-seq, it's impossible to truly understand what is binding to the sequences. For instance, Figure 4.2 A) shows numerous non-specific binding sites, indicating it's impossible to be sure V$AHRARNT.03 is causing increased SEAP production. While B) shows too few. In part B the combination of the core binding sites of the TFRE and the "TA" spacer led to only 4 of the wanted binding sites, instead of 6.

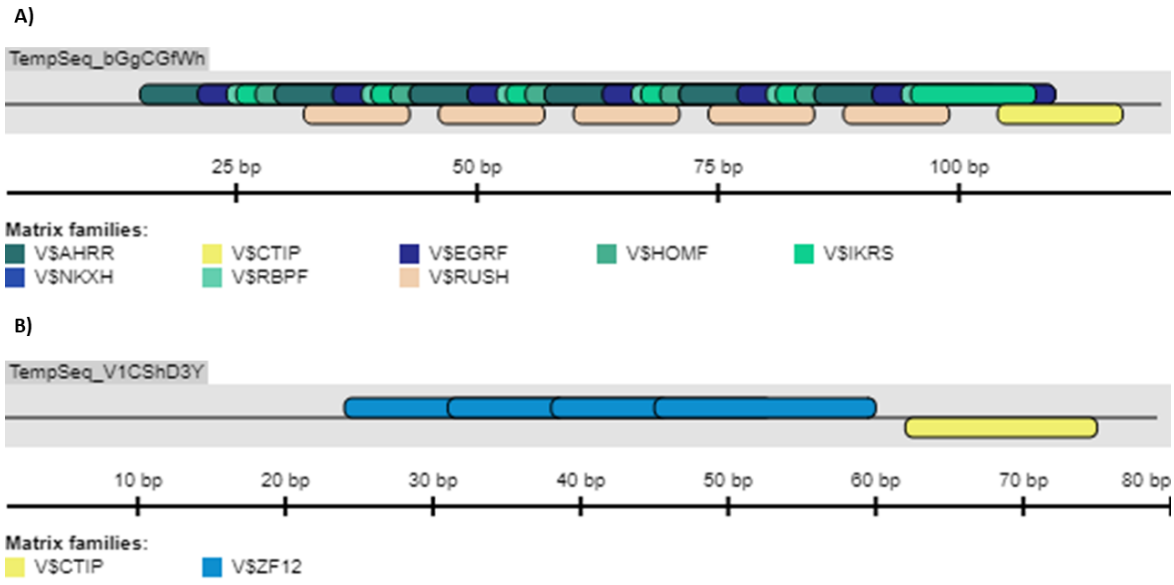


Figure 4.2: Spacer testing can only reduce non-specific binding to a certain degree. Results are from running the homotypic promoters through Genomatix Matinspector with settings of "Core = 1" and "matrix similarity = Opt 0.01". Part A) shows how unspecific a single homotypic can be. This sequence is V$AHRARNT.03 which has approximately 60 binding sites which were not designed. Part B) shows the extreme opposite end of the spectrum. Here the sequence for V$CSRNP1.01 has been analysed and as can be seen, it doesn't contain the six designed binding sites. This is due to a combination of the search criteria used on Genomatix Matinspector being too stringent and the core binding site, along with the spacer not matching MatInspectors flanking regions for this specific sequence.

Vectors were synthesized with Genewiz (Chelmsford, Massachusetts, United States). Genewiz inserted these 6 copy repeat sequences into pSEAP2-CMVCore created by Yusuf Johari. This vector contained a minimal CMV core promoter in front of the SEAP gene and contained the restriction sites KPNI and HINDIII for easy restriction digest cloning. The methods of transfections and assays are presented in Sections 2.1.6 and 2.3.1 of the Materials and Methods. Once the homotypic promoters arrived, they were transformed and midi prepped to ensure transfection grade DNA. All sequences were sequenced using Sanger sequencing at Genewiz after being midi prepped to ensure the promoter had not mutated.

Figure 4.3 shows the results of the SEAP assays performed. All data was normalised against the Full-length CMV (FCMV) promoter. This was the pSEAPs minimal CMV core promoter with the full CMV enhancer region and transfected at a DNA load of 400ng. Due to the small size differences of the promoters it was unneeded to account for copy number of plasmids as between the largest and smallest promoters there was only a 10% difference in DNA load. CMV was used as it is the industry standard promoter and would allow indication of the % contribution these TFREs could have compared to the full promoter. Due to the small size difference The results for the first TFRE discovery pipeline failed to achieve their goal, with only NRF1, MIT and BACH2 showing activity from the selection of TFREs tested. NFkB is the previously

highest activity TFRE found originally by Brown et al. (2015) and further confirmed by Johari et al. (2019). The new TFREs do contribute to expanding the design space of synthetic promoters as they give more weak building blocks to utilise and increase diversity, albeit their overall transcriptional output is no more than 25% of CMV.

As aforementioned in Section 4.1, extensive work has previously been carried out in CHO. The decision to not test families that have been previously tested, likely resulted in examining of families that either do not have a transcriptional activation function or do not contribute to that function a great deal in transient transfection. For instance, V$GAGA was found to be one of the most common TFRE binding sites in CHO, but when tested it showed no SEAP production. This could be due to it not functioning in transient mode but being an insulator element, as discussed in Srivastava et al. (2013). The function may only be seen in a stable context and may only affect the stability of the promoter region instead of the transcriptional activity. The over-representation of this factor upstream of genes with a high expression may indicate that instead of having very active promoter regions, the upstream region may inherently be extremely stable, holding the chromatin open much like UCOEs are thought to.

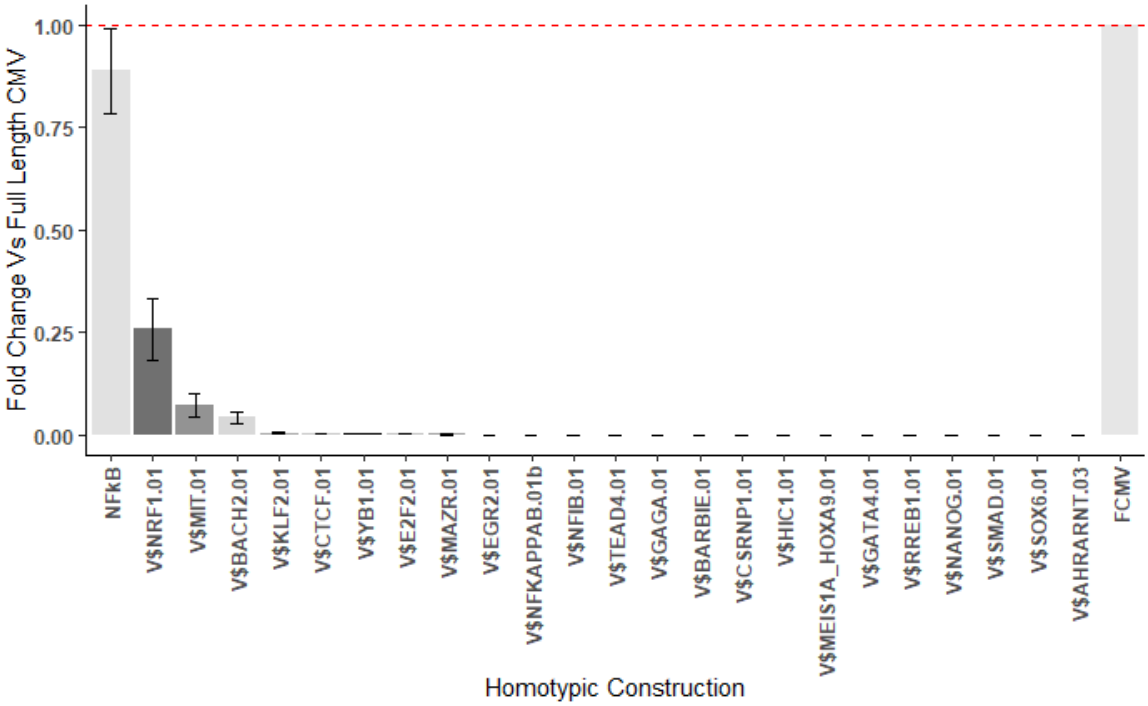


Figure 4.3: Homotypic testing of library 1. The x-axis shows the names of homotypic sequences tested and the y-axis shows the fold change versus the Full-length CMV. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest (n = 3 ± sd). NFkB is a comparison from a previous work by Brown et al. (2014) (NFKAPPAB.02). Overall, most of TFREs showed little to no transcriptional activity when producing SEAP.

4.3 Testing TFRE Sequence Variants

While pipeline one was ongoing, an investigation was carried out to assess if different variants of the same TFRE provide different activities. For this experiment, only the

NFkB family was tested as proof of concept. The hypothesis, as discussed in Section 1.4.3, is that small changes in a TFRE sequence can impact the overall transcriptional activity of that sequence. The sequence variants were obtained from the Genomatix database and also 3 synthetic binding sites were taken from Wong et al. (2011b) and tested. The normalisation for this section used the CMV core. This was due to this part of the work being performed initially in the project and mimicking the work done in Brown et al. (2015) and Johari et al. (2019).

In previously presented and subsequent data the CMV core normalisation was dropped as it required a different dilution factor for the SEAP assays compared to the rest of the samples. This caused inconsistencies in the data due to different dilution factors being required. As such it was decided that Full-length CMV would be a better normalisation construct for this work and allowed more uniform dilution factors in the subsequent work.

The selection of the NFkB variants was based on the data contained in the Genomatix Matbase tool. Each sequence was chosen based on the most common base in each position in the provided frequency tables shown in Genomatix Matbase. The synthetic sequences were taken from Wong et al. (2011a). Three high affinity constructs were chosen, along with one low affinity. The hypothesis was to check if affinity and transcriptional activity were linked. Table 4.4 shows the sequences tested in this experiment.

Table 4.4: Sequences used for the NFkB variant experiment. The first column is the name of the sequence and the second column is the sequences used.

Matrix	Sequence
V$CREL.01	GGGGCTTTCC
V$HIVEP1.01	TGGGGACTTTCCT
V$NFKAPPAB.01	GGGAATTCCC
V$NFKAPPAB.02	GGGGACTTTCCA
V$NFKAPPAB50.01	GGGGATTCCC
V$NFKAPPAB65.01	GGGAATTTCC
V$NFKAPPAB65.02	AGGGGATTTCCCAG
High Affinity1	GGGGAATTCCC
High Affinity2	GGAAATCCCCT
High Affinity3	GGGAAAGCCCC
Low Affinity1	CAGAAGATCCT

The results for the NFkB variant assay are shown in Figure 4.4. Small changes in the TFRE sequence of NFkB affect the transcriptional activity quite significantly. For instance, the only difference between HIVEP1.01 and NFKAPPAB.02 is a single base pair added to the beginning and a change of T to A at the end of HIVEP1.01 ("TGGGGACTTTCCT" versus "GGGGACTTTCCA"), which caused a 10 fold increase. The synthetic sequences indicate that strong binding affinity causes transcriptional

activity but does not correlate to the highest levels of expression. For instance, High Affinity 1 had an average z-score of 3.634, High Affinity 2 a score of 3.060 and finally, High affinity 3 had a z-score of 2.991. These sequences, never reached the activity of NFKAPPAB.02. The Low Affinity 1 sequence showed that if there no binding affinity then there is no transcriptional activity.

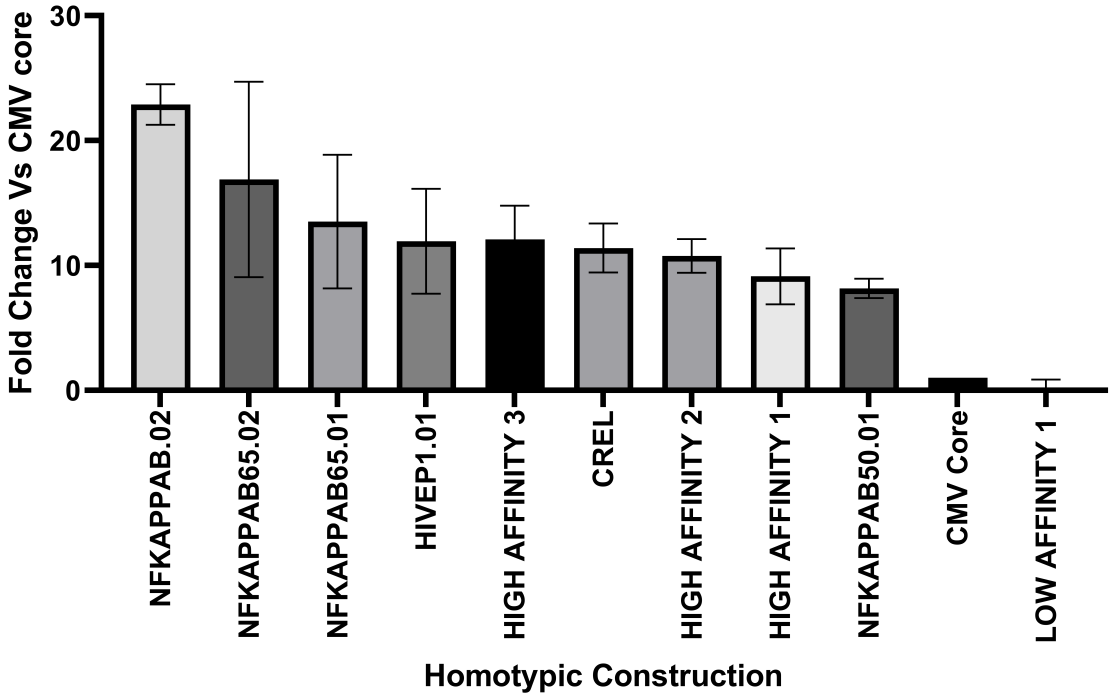


Figure 4.4: Small changes to the NFkB binding sequences can affect SEAP production. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest (n = 3 ± sd). The control is a minimal CMV Core promoter to which all data is normalised. The results show that even minimal changes in a sequence of as little as 2 bp affect the transcriptional activity. For instance, HIVEP1.01 and NFKAPPAB50.01 only have a 2 bp difference between them and NFKAPPAB.02 showed a result of 20 fold versus the CMV Core versus 10 fold for HIVEP1.01.

4.4 TFRE Discovery Pipeline 2

The underperformance of TFRE discovery pipeline 1, led to a complete revision. Instead of focusing on over-representation, a pipeline utilising a combination of over-representation and transcriptional activity was used. The promoter regions used were reduced, instead of 3600bp, only 1200bp regions were used. This was previously done in Brown et al. (2015) and Johari et al. (2019). It was theorised that the region closer to the promoter's core would contain more TFREs that serve as transcriptional activators.

The initial setup of this pipeline included creating a custom CHO background for Genomatix, to mimic the mouse database. A tool called Genomatix Overrepresented TFBS was used for the subsequent analysis. This tool allowed for two critical advantages in contrast to TFRE discovery pipeline 1. The first, unlike MatInspector, Overrepresented TFBS could accept all of the custom CHO background promoter sequences in one upload and secondly, it did an over-representation analysis itself using z-score. Overrepresented TFBS compared uploaded sequences to the mouse genome, mouse promoter regions and the custom CHO promoter background in one analysis. The only disadvantage was it reduced some of the automation, as Genomatix must be used on the web page and cannot be linked to R or Python to automate the process.

The over-representation performed in Genomatix Overrepresented TFBS results is similar to what was shown in Figure 4.1. The primary difference is the normalisation score. Instead, z-score was used to analyse how over-represented the sequence was. Anything with a z-score of over 2 or under -2 was considered statistically relevant compared to the sequences to which they were being compared to. The inference that Figure 4.5 and Figure 4.1 look similar, indicates the coded form of normalisation present in TFRE discovery pipeline 1 and over-representation used in this are very similar. Incidentally, the shorter promoter length used still provided V$GAGA as one of the most over-abundant TFRE in both analyses. Code for this analysis is shown in Appendix B.5.

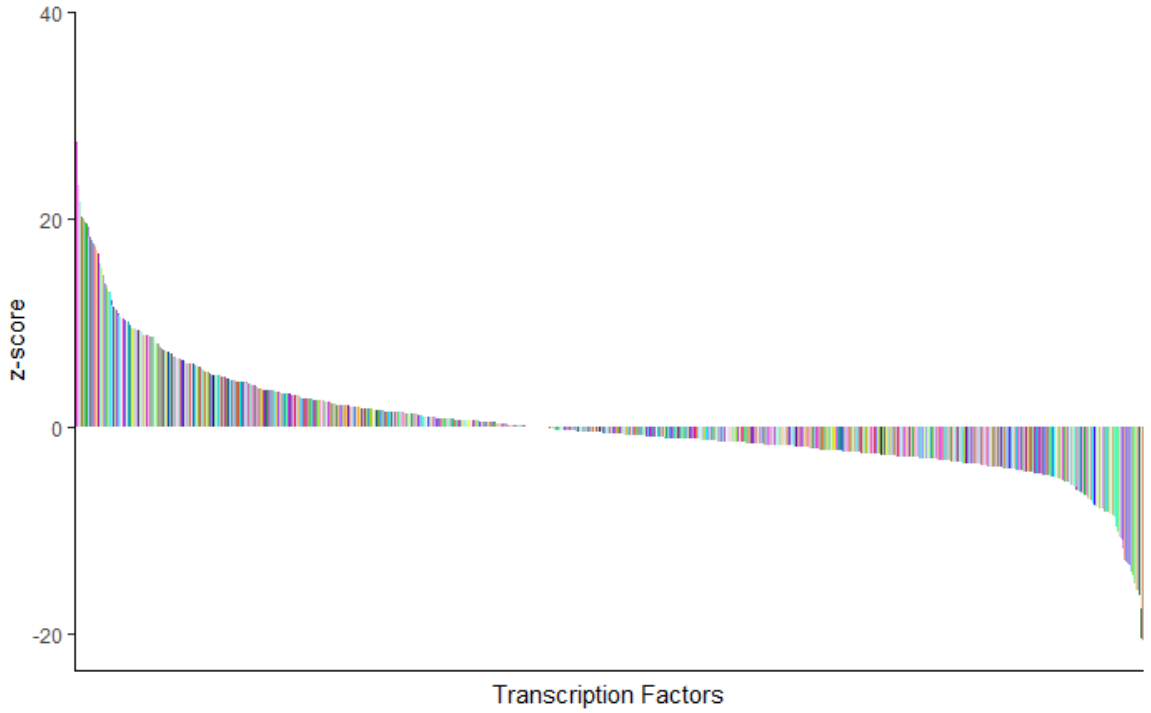


Figure 4.5: Analysis of the normalised frequency from Genomatix Overrepresented TFBS of TFREs in promoter regions of high expressing CHO genes. The z-score is on the y-axis and the TFRE sites are on the x-axis. The results are very similar to what was shown in pipeline 1. The use of z-score instead of normalisation score allows for the use of statistics to prove if the result is statistically relevant. For instance, a z-score of over 2 would be equivalent to a $p < 0.05$. If over-representation was being used alone in this pipeline it would allow for a screening panel cut-off point.

To quantify the expression of each transcription factor in the CHO genome the TPM metric was utilised. The average TPM across 3 biological replicates was graphed on day 2, day 5 and day 10 for transcription factor genes. The overall average of the gene TPM across days and clones was used for subsequent analysis. Although TPM shouldn't be used for differential expression, it was thought that an average was important as the overall expression on all days of cultures was required. Both the Mock and Clonal lines were included in this also to ensure the results would apply to a cell line which has not adapted to produce a specific product.

An extra filter was added to reduce the size of the selection panel. The MGI database was used and genes functionally annotated as "transcriptional activators" were taken, converted to mouse orthologues and merged with the RNA-seq files using ENSEMBL IDs. This was to minimize the amount of literature review needed when looking at the screening panel. The results shown in Figure 4.6 have been filtered in this way.

The results in Figure 4.6 reflect results obtained in Adam Brown and Yusuf Johari's work (Brown et al., 2015; Johari et al., 2019). For instance, *FOSL1, JUND, ATF4* and *RELA* have been found in previous works to be good transcriptional activators for synthetic promoters. The most notable of which is *RELA* which is the gene name for the transcription factor protein NFkB. This suggests if the expression of transcriptional activators alone was used, one could find transcription factor binding sites that provide high transcriptional activity for synthetic promoters construction. The results shown

in Figure 4.6 contrast those presented in Brown et al. (2017) in terms of transcription factors with high expression, but it may be due to naming conventions. For instance, E-box and MYC are the same transcription factor family but can have different naming conventions based on the data source. A major advantage this analysis provides by being fully bioinformatic based is that it's broader compared to what is shown in Brown et al. (2017).

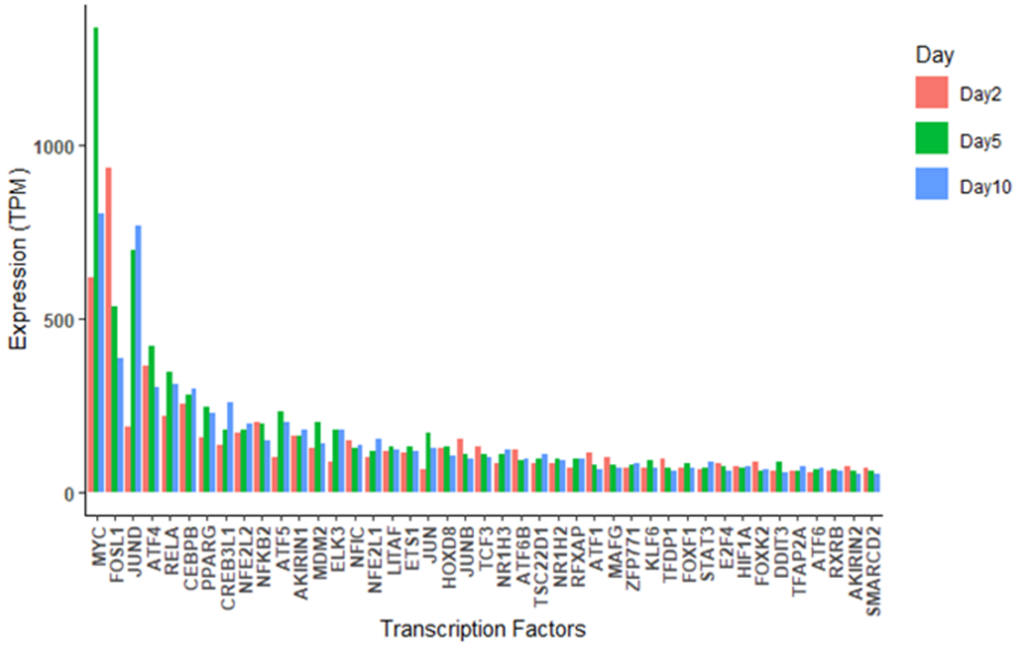


Figure 4.6: The expression of transcription factor genes were analysed across day 2, 5 and 10 of culture. These transcription factor genes have been filtered to show only transcriptional activators based on the MGI database. Only a few transcription factors have an extremely high TPM of over 100. Most transcription factors, even activators, have a TPM average value of less than 100. Based on empirical evidence provided by Brown et al. (2014) and Johari et al. (2019), it would be possible to simply screen this list and find building blocks for synthetic promoters.

The next step was to investigate the correlation between the z-score metric and the expression of transcription factors measured in TPM. This was to investigate if the expression of these transcription factors correlated with there over-representation. One hypothesis is that if genes have high expression due to the presence of a particular TFRE, the transcription factor that binds to said TFRE would also be highly expressed. Figure 4.7 demonstrates a disproving of this hypothesis and shows no correlation between high expression and abundance in high expressing promoter regions. The expression and over-representation have no correlation.

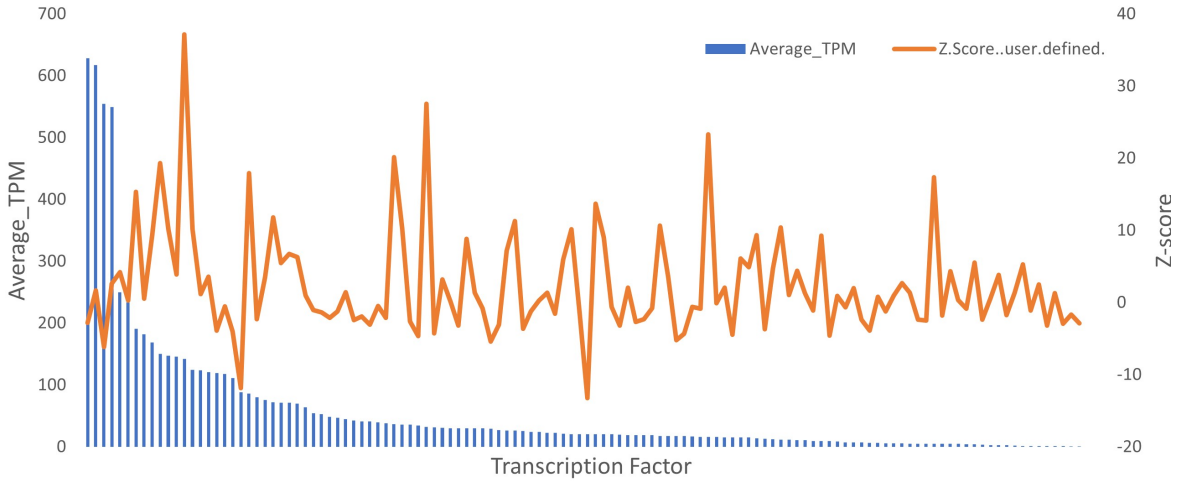


Figure 4.7: The relationship between over-representation (z-score) and expression (TPM) shows no direct correlation. This figure shows the average TPM on the left y-axis and z-score on the right y-axis. The transcription factor genes are graphed on the x-axis. When TPM is plotted along with z-score there is no correlation between the expression of the transcription factor and over-representation in high expressing genes. This was expected as Genomatix Matinspector detects many TFREs whose corresponding transcription factors are not expressed. This analysis was done using a filtered list of transcriptional activators, also used in Figure 4.6.

Following this, the z-score and TPM metric were combined by multiplying the two values together, as seen in Figure 4.8. This new metric emphasised both abundance and expression. By not putting any limit on the amount of over-representation or the expression, some transcription factors were found that would not have been tested when looking at either the expression or over-representation alone. Interestingly, in this analysis, some previously tested transcription factor elements (Section 4.2) appeared in the selection panel such as CTCF, NRF1 and BACH.

A drawback of this method compared to over-representation alone was the data loss due to the implementation of automation. For instance, one notable TFRE missing is NFkB. This is because of a disconnect between Genomatix annotations and the RNA-seq ENSEMBL annotations. For instance, in Genomatix NFkB is called NFkB, but in the RNA sequencing dataset, genes are listed as names such as RELA. Although a significant loss in terms of a historically great transcription factor for CHO synthetic promoters. The speed and breadth of the analysis outweigh the disadvantages, as it can fully be completed in as little as an hour from RNA-seq data to the screening panel.

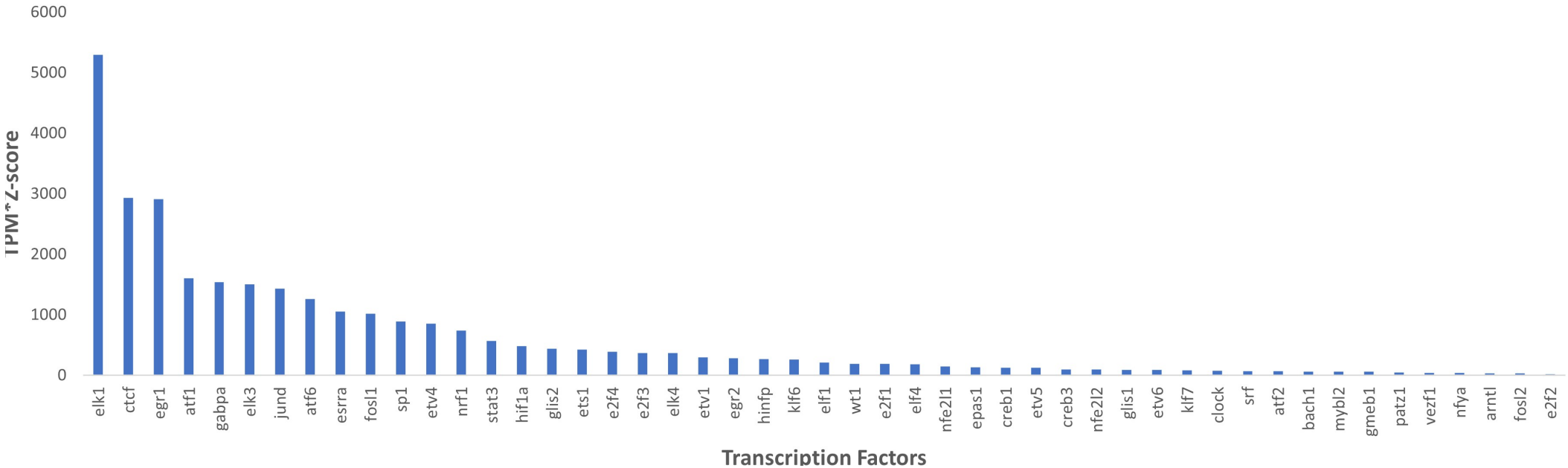


Figure 4.8: A new metric combining z-score and TPM to consider the abundance of TFREs and the expression of the transcription factors which bind to them. This metric is an effort to consider the abundance and expression of the TFREs and the transcription factors. One notable TFRE missing from this analysis is NFkB. Still, other transcription factors such as GABPA, JUND, NRF1 and SP1 appear very high in the analysis showing good predictive accuracy based on what has been seen in previous works (Johari et al., 2019; Brown and James, 2016) and the first homotypic screen. ELK1 has been previously tested and shown to be a good transcription factor, although it's called ETSF, which is the matrix family name (Brown et al., 2015).

From this panel, 37 sequences were chosen to bring forward. New rules were used to choose the TFREs to test. These were as follows;

- They must have an expression level of at least 20 TPM or over. To ensure that they have a general medium to high level expression.
- When the literature was checked, there needed to be evidence of transcriptional activation. This was an attempt to reduce false positives.
- Although the majority of TFREs came from selection by using a combination of TPM * z-score, if an interesting TFRE was observed either in the overrepresented or expression analysis, it was included. For instance, NM23 was included in the over-representation analysis as it had a z-score of 4.23 and a TPM of over 1000. It was excluded from the TFRE discovery pipeline 2 due to annotation differences between Genomatix and the gene IDs used in the RNA-seq data, much like NFkB.
- Families that were tested in previous works were included, their binding sites had to be at least 1 bp different. Any repetitive binding sites from previous works or in this analysis were excluded. This was done as it was shown in the previous chapter that small changes can have an effect on the transcriptional activity of the promoter.

Utilising the criteria set out above a subsection of sequences were chosen. The decision on what sequence variant should be used, if for instance "GABPA" appeared, was based on which version had the highest over-representation. The final selection of sequences is shown in Table 4.5. As described previously, based on previous data presented, it was decided that variants of sequences previously tested could be used and as such several members of some families are present. All sequences were synthesised with a "TA" spacer to ensure that the results were comparable to the data obtained in library 1. The sequences which were tested are shown in Appendix A.3.

Table 4.5: TFREs tested in TFRE Discovery Pipeline 2. The first column shows the matrix name as provided by Genomatix. The second column provides the sequence used.

Name	Sequence Used
V$NFE2L2.01	TGCTGAGTCAT
v$gabpb1.01	CCCGGAAGTGAC
V$GABPA.02v.1	CCGGAAGTGG
V$EGR1.04	GGGCGGGGGCGGGG
V$GABPA.02	ACCGGAAGT
V$CLOCK_BMAL1.01	GGGTCACGTG
V$ETV4.01	CCGGAAGT
USF1.02	GGTCACGTG
V$USF.04	GTCACGTGG
V$NRF1.01	CGCGCATGCGC
V$HIVEP1.01	GGGACTTTCC
v$atf2.01	TGACGTAA
V$ETS1.01	CAGGAAGTG
V$KLF6.01	GGGGGCGG
V$XBP1.01	GATGACGTG
V$PREB.01	CATCATCAGACACC
NF1.02	TGGCACCATGCCAAGA
V$MAFK.01	AGTCAGCATTTT
V$HSF2.01	GAACATT
V$ESRRA.02	AGGGGTCA
V$NM23.01	GGGTGGGGGGGGG
V$YY1.04	GCCGCCATCTTG
V$MAZ.04	GGGAGGGGG
V$NR2F6.01	GGTCAAAGGTC
V$HSF1.02	GAAGATTCGAGAACATTC
V$HSF1.04	TTCTGGAAGCTTCT
V$HPF1.01	AGGACAAAGGCCAGCC
V$HSF1.05	TTCCAGAA
V$ATF1.02	GTGACGT
V$HRE.03	ACGTGC
V$HRE.02	ACGTG
V$ZNF771.01	GCGCTAACCA
V$ATF6.02	TGACGTG
V$E2F6.01	GGCGGGA
V$RARG.01	TGACCTTTTG
V$BBX.01	TGAACGACGTTCA
V$RAR_RXR.03	GGGTCACAGAGAGTTCA

Figure 4.9 presents the overall results from pipeline 2. The results were superior to those previously seen in pipeline 1, with two new sequences NFE2l2 and GABPB1.01, being discovered which had no statistically significant difference from NFkB. When compared to NFkB using an unpaired two-tailed t-test, they had p-values of 0.83 and 0.90, respectively. This indicated better activity than other alternative TFREs found in Brown et al. (2014, 2015) and matched the activities shown in Johari et al. (2019). Unexpectedly, NFE2l2 had previously not been tested and in a homotypic scenario, matched the activity of both NFkB and GABPB, indicating a strong new TFRE had been found.

Compared to pipeline one, a total of 7 new TFREs with a fold change of 0.2 or over versus Full-length CMV were found. This pipeline found a higher deal of transcriptionally active TFREs, although out of the 36 tested, 9 showed activity. Revealing with this system of discovering new TFREs, empirical testing is still essential to assess if the TFRE is functional. However, the success rate has increased from library 1 where 3 out of 23 were functional.

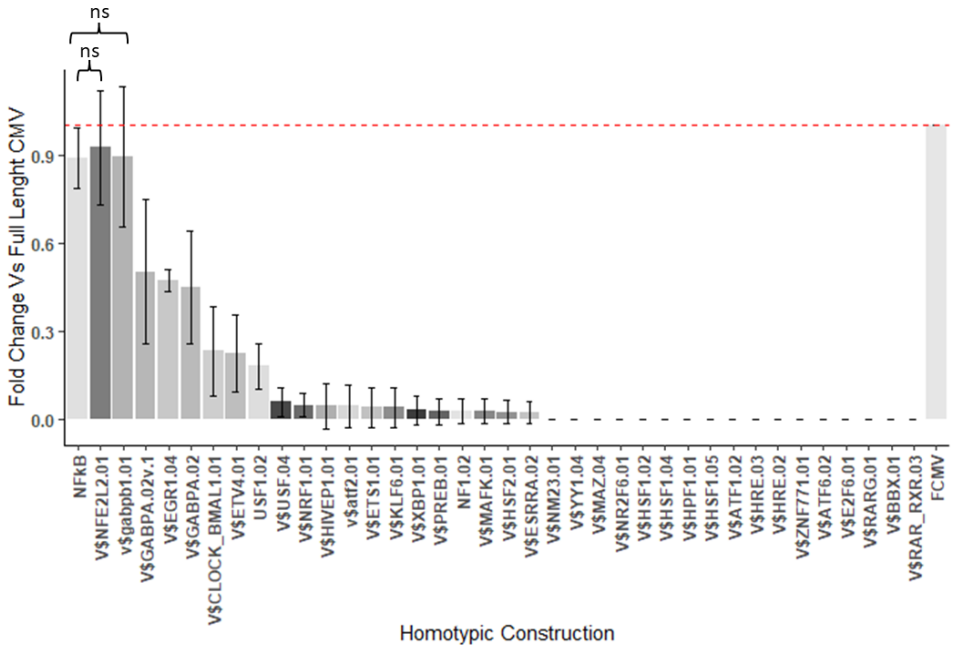


Figure 4.9: TFRE Discovery Pipeline 2 found 10 new active TFRE sites. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest (n = 3 ± sd). NFKB is NFKAPPAB.02 from Section 4.3. The results indicate pipeline generation 2 was more successful than the previous generation. The new TFREs NFE2l2 and gabp1.01 matched the activity of NFKB with no statistically significant difference. Some newly found TFREs, such as USF1.02 showed slight activity in some replicates and not others, leading to error bars which fall below zero. Overall, pipeline 2 appears to be successful in automating the TFRE identification process and finding new TFREs to expand the design space of synthetic promoters. The statistics presented on the graph indicate the following p-values based on a unpaired two-tailed t-test: * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$ and ns = not significant.

4.5 Conclusions and What's Next?

Overall, this study sought to create a bioinformatic pipeline for the discovery of new TFREs, while implementing strategies to reduce manual screening.

The TFRE discovery pipeline 1 resulted in disappointing outcomes as it relied on over-representation and novel sequences alone. The benefits of this pipeline are its potential ability to identify unique stability structures upstream of promoters, such as the "V$GAGA" motif. This may indicate that by looking at the over-representation of TFRE binding sites and also comparing the regions to lower expressing genes promoter regions, one may be able to identify structural elements which confer stability or have other useful functions not related to transcriptional activation.

Although presented outside of chronological order to simplify understanding, the NFkB variation experiment was performed in parallel with TFRE Discovery Pipeline 1. The experiment was crucial in verifying a misstep taken in pipeline 1. The exclusion of sequence variants. This was crucial for pipeline 2, because as seen in Figure 4.4 small changes to the reported TFRE sequence can drastically affect the activity when tested in homotypic contexts. This is further verified in the results for TFRE discovery pipeline 2 in which sequence variations were allowed.

TFRE Discovery Pipeline 2 aimed to take into consideration the expression of the transcription factors which bind to the TFREs. As expected the over-representation of TFRE binding sites in active promoter regions and expression of the transcription factor don't correlate but by creating a new metric of z-score multiplied by TPM of the gene a new metric was created which hoped to account for how often a TFRE occurs and the expression of the transcription factor which binds to said TFRE.

The major drawback of pipeline 2 was information loss. For instance, using this pipeline, NFkB would have been overlooked, although TFREs with an equivalent activity that had previously not been discovered would not have been found. This was due to inconsistent annotation between different data sources that were used.

Pipeline 2 has significant advantages over pipeline 1 and what was used in Brown and James (2016); Brown et al. (2017) and Johari et al. (2019). The first is automation. Once the pipeline had been developed and the background sequences extracted, the whole pipeline could be run in as little as 30 minutes. It also provided the advantage of taking gene expression in the form of TPM into account, the number of false positives is reduced. This is because TFREs that appear over-represented but are not expressed in CHO are downranked in the screening panel using the new metric.

This work has expanded the design space by finding new building blocks, but there is still much scope to further the potential designs. For instance, in this work TFREs are looked at as individual blocks which bind to the DNA and cause a function. Still, it is well documented that transcription factors do not act as individual proteins but bind in cooperation to cause different functions. For instance, GABPA and GABPB bind cooperatively before binding to DNA to initiate their function Jia et al. (2020). This is the same for NFKB and many other transcription factors.

The other area of potential expansion would be through binding studies. By discovering what binds to synthetic sequences through CHIP-seq or TF-seq. It would provide an essential understanding of the binding kinetics and give an indication of the mechanisms occurring, allowing more informed design for the next generation of building blocks. Currently, as shown by the work presented in this Chapter, a TFRE is just a binding site to which a transcription factor should bind and initiate a specific function. Although in creating homotypic promoters, the aim is to test what the strength of that TFRE is, one cannot say that it is the actual transcription factor binding to these sequences. At most, all that can be determined is that this TFRE sequence provides this activity. It's not necessarily the NFKB or GABPB transcription factors.

Overall, this work has fulfilled the initial aim of providing bioinformatic discovery of new TFREs for synthetic promoter design. This work was done in parallel with Chapter 5 and as such reference will be made back to this chapter to explain when the results discussed in this section intercede.

Chapter 5

Unidirectional Synthetic Promoters

Overview

- The utilisation of homotypic TFREs to create heterotypic unidirectional synthetic promoters will be discussed in this chapter.
- To ensure Merck had the deliverables it requested from the project, unidirectional promoters were created by combining the knowledge from previous literature with novel promoter designs. This is discussed in Section 5.2.
- Unidirectional library 2 is shown in Section 5.3, which incorporates the newly found TFRE NRF1 into the synthetic promoter building block repository and examines how this affects the overall activity of the synthetic promoters.
- The cumulative result of the promoter research is in Section 5.4, which depicts how the inclusion of all the new design blocks affected SEAP production. It also led to the creation of synthetic promoters, which relied on none of the previous literature.

5.1 Introduction

Even the strongest building blocks such as NFkB can only match the activity of CMV but not overcome it in a homotypic context. The TFRE building blocks must be combined to achieve higher activity then that of full-length CMV. This can be seen in previous works such as Brown et al. (2015); Brown and James (2016); Brown et al. (2017) and Johari et al. (2019).

The need for diversity in synthetic promoters to achieve higher activity than that of full-length CMV is expected, as naturally, promoters contain many different TFREs that contribute to both activity and regulation. Previous studies have tried to use mathematical modelling to predict what impact each TFRE will have on the general activity, such as Johari et al. (2019) using design of experiments (DOE). Still, in each instance, this is only applicable to the research itself and not useful in practice unless the exact experiment was to be performed again, with the exact same cell line. The ideology behind the study here is that previous literature TFREs can be grouped into providing high activity or low activity and when combined can overcome the activity of FCMV. For this reason, this study does not try to model the data as it is performed in CHOS due to IP constraints. Diagrams of all constructs are supplied in Appendix A.8 if visual interpretation is required.

Due to the industrial nature of this project, the primary aim was to provide synthetic promoters that Merck could use in their industrial systems. For this reason, the first library aimed at just using previous literature to provide promoters which provide a varying amount of activity and are transcriptionally more active than full-length CMV.

5.2 Unidirectional Library 1

As aforementioned, the first library of synthetic promoters was built solely on previous knowledge. This was due to the requirement of deliverables for Merck. The works in this chapter were performed in parallel with works in chapters 3 and 4, and as such the information discovered was unavailable in the first library.

Figure 5.1 shows the general thought process for library 1. The idea was to take all of the homotypics founds in Brown et al. (2014) and Johari et al. (2019) to create promoters with a 2bp spacer in-between each element. All sequences for this section are contained in Appendix A.4. Initial extra design considerations for the first library are as follows:

- Randomize the TFRE position, unless intentional design was to be implemented.
- Vary the amount of each TFREs in each promoter.
- Attempt to increase the complexity by increasing the number and combination of TFREs included in promoters.

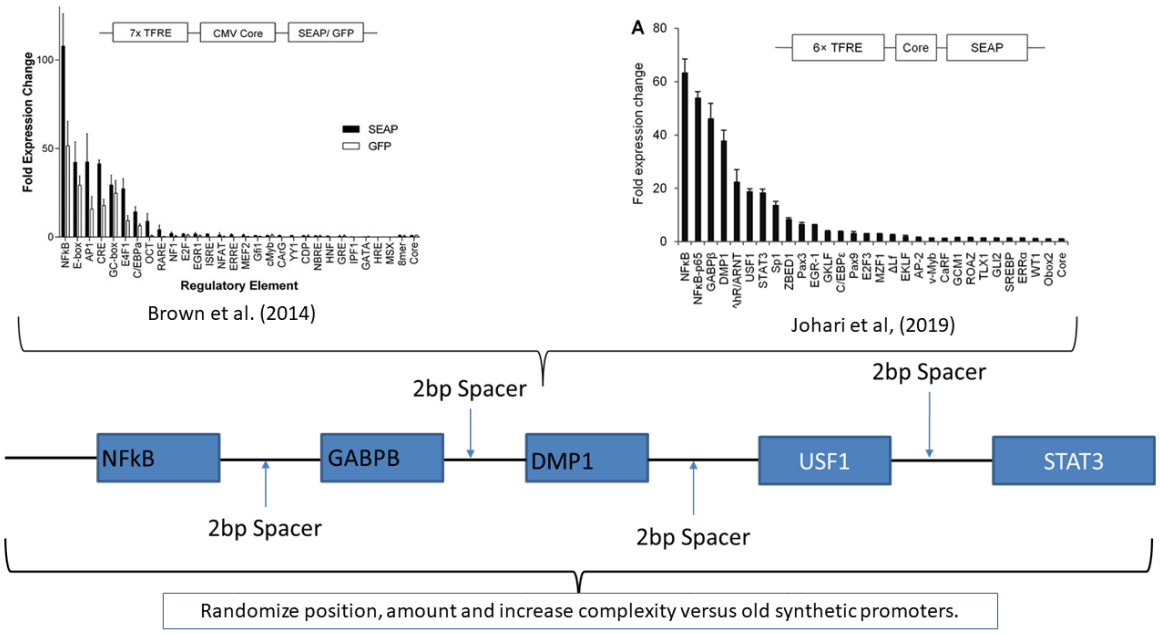


Figure 5.1: Promoter design workflow for library 1. The information found in Brown et al. (2014) and Johari et al. (2019) was combined to create heterotypic promoters. A 2bp spacer was placed between each TFRE block to try to reduce non-specific binding. The blocks in black are high activity and in white are lower activity.

The composition of all the promoters tested in this library are shown in Table 5.1. Like the homotypic section, the spacer was kept the same in each sequence. The spacer was chosen based on having the least number of matches running through Genomatix Matinspector. The actual TFREs used were chosen based on Table 1.2, with a mix of high and low activity TFREs being chosen in a effort to increase the heterotypic promoter diversity. The spacer for all sequences was "AA".

The design decisions considered for the first two sequences in Table 5.1 initially tried to look at the effect of transcription factors that are thought to confer stability, such as ZBED4 and SP1 (Mokhonov et al., 2012). These sequences are "Heterotypic High Transcription Promoter" and "Heterotypic High Transcription Promoter without stability".

Next, the use of repressive TFREs was tested in sequences "Promoter with 5' repressor" and "Promoter with 3' repressor". As mentioned in the literature review, many transcription factors have been thought to be bidirectional. By placing repressing sequences on the 5' end of DNA, the unidirectional transcriptional activity may be increased by funnelling the RNA Poll II in one direction as described in Häkkinen et al. (2011) and Zanotto et al. (2009). The idea behind this is that there is a wall preventing transcription in the anti-sense direction sending more RNA Poll II in the sense direction

The next two sequences, "Really High TFRE" and "Remixed High TFRE" were trying to max out the amount of TFREs put into a promoter with no consideration of the length. The remixed version attempted to see if the same composition but in a different order would lead to different results. The "5' weighted" and "3' weighted" were an attempt to check if putting the more transcriptionally active TFREs closer or further away from the core affected SEAP production/transcriptional activity to any degree.

Finally, "Heterotypic TFRE Compliment Prom 1" and "Heterotypic TFRE Reverse Compliment Prom 1" are variations of "Heterotypic High Transcription Promoter" and aimed to look at if the compliment of TFRE sequences and the reverse complement of TFRE sequences affected the transcriptional activity. This was important to assess for future works involving bidirectional promoters. All sequences tested are shown in Appendix A.4.

Table 5.1: TFRE composition of each of the sequences for library 1. The first column shows the names of the sequences, followed by the length and the amount of each TFRE the sequence is composed of. The final column (Matches) shows the number of resulting matches when run through Genomatix Matinspector with settings of Core = 1 and Matrix Similarity = Opt +0.01.

Name	Length	NFkB	GABP beta	DMP1	ARE	AhR/ARNT	HRE	AARE	Sp1	ZBED4	YY1	Matches
Heterotypic High Transcription Promoter	396	6	6	4	4	4	4	4	3	3	0	104
Heterotypic High Transcription Promoter without stability	340	6	6	4	4	4	4	4	0	0	0	61
Promoter with 5' repressor	373	6	6	4	4	4	4	4	0	0	3	71
Promoter with 3' repressor	373	6	6	4	4	4	4	4	0	0	3	68
Really High TFRE	577	7	7	5	5	5	5	5	5	0	0	145
Remixed High TFRE	577	7	7	5	5	5	5	5	5	0	0	138
5' Weighted Promoter	577	7	7	5	5	5	5	5	5	0	0	138
3' Weighted Promoter	577	7	7	5	5	5	5	5	5	0	0	141
Heterotypic TFRE Compliment Prom 1	396	6	6	4	4	4	4	4	3	3	0	74
Heterotypic TFRE Reverse Compliment Prom 1	396	6	6	4	4	4	4	4	3	3	0	146

The transfection methods and SEAP assay are described in Section 2.1.6 and 2.3.1 respectively . Another comparison was added in the case of BM1 and BM2. These sequences were the top promoters taken from Brown et al. (2014) (BM1) and Johari et al. (2019) (BM2). The addition of these sequences was to benchmark how well the new sequences were performing compared to what had previously been created in literature.

The results are shown in Figure 5.2. Combining TFREs from previous literature results in the creation of promoters with the equivalent activity of the previous benchmarks BM1 and BM2. An unpaired two-tailed t-test was performed and showed BM1 and BM2 had no significant difference from the top promoter "3' weighted promoter", but all three promoters showed significant differences when compared to the control.

The results showed some interesting observations. The first is that the strongest promoter overall was the "3' weighted promoter" which had an average fold change of 1.7 fold higher expression than the FCMV control. Interestingly, although this promoter's composition is the same as "5' weighted promoter" the average activity differs by 0.4 fold when normalised to FCMV. This is also a statistically significant difference when compared using an unpaired two-tailed t-test. This may indicate that having strong transcriptional activators closer to the core promoter region allows the transcription factors to have a greater effect. For instance, NFkB is thought to reduce proximal pausing and perhaps having more of these transcription factors binding near the core promotes this function (Core and Adelman, 2019).

As also shown in Johari et al. (2019), the presence of SP1 does not affect the transcriptional activity, along with ZBED4. Remixing the promoters, as shown by "Really High TFRE" and "Remixed High TFRE" show that, as discussed in Brown et al. (2014), the position of the TFREs generally does not seem to matter for synthetic promoter composition when not heavily weighted in one direction or the other. This is only in contexts when TFREs are evenly distributed throughout the promoter.

The hypothesis that using a repressor on the 5' end of the promoter to increase unidirectional activity was disproved. The presence of the repressors at the 5' end of the promoter reduced the activity of the "Promoter with 5' repressor" compared to the "Heterotypic High Transcription Promoter without stability" by 0.3 fold when normalised against FCMV, although not to a statistically significant degree ($p = 0.3$). It was found that repressors, even if strategically placed, appear to reduce the transcriptional activity of synthetic promoters.

Finally, the inclusion of promoters with their TFRE sequences changed to the reverse complement or complement of the sequence indicated that the complement of TFREs leads to no transcriptional activity with no SEAP being detected in the assay. The reverse complement of the DNA acts the same as the original sequences with "Heterotypic High Transcription Promoter" having a fold change of 1.49 versus its counterpart "Heterotypic TFRE Reverse Compliment Prom 1", which was 1.43.

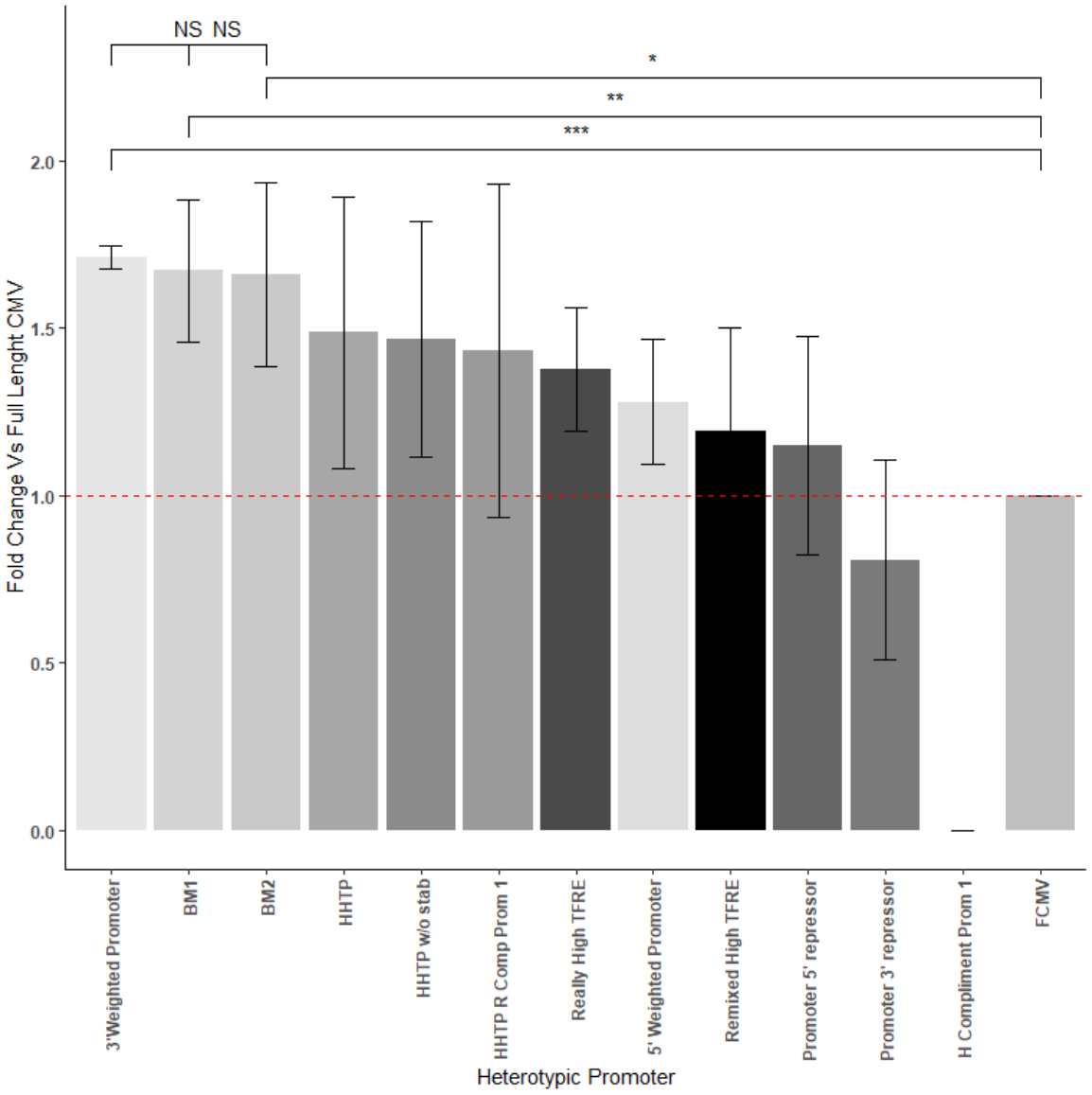


Figure 5.2: Analysis of the SEAP production of unidirectional promoter library 1 normalised to FCMV. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest (n = 3 ± sd). HHTP stands for Heterotypic High Transcription Promoter. Promoters with H in front stand for heterotypic. The results show the successful creation of sequences with higher activity than the control and indicate TFREs found in literature can be combined to rival the activity of the benchmark sequences BM1 and BM2. It appears 3' weighting may be beneficial for transcriptional activity and the reverse complement of TFREs acts the same as the normal sequence. The statistics presented on the graph indicate the following p-values based on a unpaired two-tailed t-test: * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$ and NS = not significant.

5.2.1 Conclusions on Library 1

Overall, the first library fulfilled the objective of creating a library of synthetic promoters with varying activity and providing higher activity than FCMV. The major observations from this library regarding informed design are as follows:

- 3' weighting of strong TFREs may be beneficial for transcriptional activity.
- The presence of stability elements doesn't appear to affect transcriptional activity in transient culture.
- Remixing the promoters without informed design, appears to not affect transcriptional activity.
- Funnelling the transcriptional activity using repressors doesn't appear to confer any advantages and may reduce transcriptional activity.

5.3 Unidirectional Library 2

This library incorporated the information from TFRE Discovery Pipeline 1, which included the transcription factor regulatory element NRF1 to replace SP1 and ZBED4. From literature, the GC box was also included to increase diversity further. The library also aimed to look at some concepts, such as the exclusion of spacers and varying promoter lengths.

Table 5.2 shows the promoters created for this instance of testing. The first sequences named "Balanced Promoter" were designed to have an even distribution of high and low activity TFREs throughout the sequences. "Balanced Promoter without Spacers" has removed the "AA" spacer between each TFRE to see how that affects activity.

The "3' weighted promoter with 5' repressor" was a further attempt to see if using repressors on one end of a sequence could affect transcriptional activity in the 3' direction. The reason for this was to see if it could further enhance the activity of the best promoter from the previous library and also see if the use of repressors could be applied to bidirectional promoters.

The next sequence, "Reduced Promoter_without_Spacer" was derived from the balanced promoter but with fewer binding sites to see if having the longer promoter length benefits the transcriptional activity. This was tested with and without a spacer.

To look at the effect of promoter length the promoters "100bp Promoter single copies", "200bp Promoter two copies" and "300bp Promoter" were created. This was an attempt to determine the required minimum length to match FCMV with current synthetic promoter methodologies.

Lastly, the "Super_Promoter" was created that was an attempt to create a promoter that would have unrivalled transcriptional activity with no consideration for the length of the promoter. This sequence had every TFRE available and at the maximum number of 6 used. This was also tested with and without spacers. The sequences tested in this library are shown in Appendix A.5.

Table 5.2: TFRE composition of each of the sequences for library 2. The first column shows the names of the sequences, followed by the length and the amount of each TFRE the sequence is composed of. The final column shows the amount of resulting matches when run through Genomatix Matinspector with settings of Core = 1 and Matrix Similarity = Opt +0.01.

Name	Length	NFkB	GABP beta	DMP1	ARE	AhR/ARNT	HRE	AARE	NRF1	GC BOX	Matches
Balanced Promoter	488	6	6	4	4	4	4	2	5	0	100
Balanced Promoter without Spacers	420	6	6	4	4	4	4	2	5	0	119
3' Weighted Promoter with 5' Repressor	521	6	6	4	4	4	4	2	5	0	106
Reduced Promoter	267	3	3	3	2	2	2	1	3	0	57
Reduced Promoter_without_Spacer	231	3	3	3	2	2	2	1	3	0	60
100bp Promoter single copies	108	1	1	1	1	1	1	1	1	0	18
200bp Promoter two copies	218	2	2	2	2	2	2	2	2	0	42
300bp Promoter	328	3	3	3	3	3	3	3	3	0	66
Super_Promoter	564	6	6	4	4	4	4	2	5	4	120
Super_Promoter_without_Spacers	488	6	6	4	4	4	4	2	5	4	140

Figure 5.3 shows the activity of these synthetic promoters against FCMV. The first observation from this library was that the highest average activity was only 1.5 times greater than the control, compared to the previous library, which had a maximum activity of 1.7 fold. This may indicate that the inclusion of NRF1 and the GC-box has reduced activity overall in the library.

The effect of spacers can drastically impact the activity of a sequence. Out of the 3 sequences that were tested with and without spacers, 2 increased significantly and one decreased. The "Super_Promoter" sequence saw the most drastic effect with a large increase in SEAP production when the spacers were removed.

The sequences without spacers that showed higher activity than those with spacers were put into Genomatix Overrepresented TFBS tool and normalised against the sequences with spacers. This showed that the sequences that lacked spacers had an increased abundance of the CREB and HIFF TFRE families. When this was broken down to matrices, it instead showed an over-abundance of STAT1 and STAT4. CREB has been shown to be abundant in CMV Brown et al. (2015).

A comparison was also made of the "Balanced Promoter" with spacers versus the other weaker sequences with spacers, but this showed no over or under abundance of note when families and matrices were analysed. This may indicate that the TFREs such as CREB, HIFF and perhaps even the STAT transcription factors have had binding sites unintentionally created by taking out the spacer in the stronger sequences and thus providing increased diversity which has increased the activity in two of the three instances.

The promoters of varying lengths showed doubling a promoter's composition is not directly equated to its activity. What can be seen is that the 100bp promoter does not have much activity with a activity of only 28% of the FCMV. The 200bp promoter then more than doubles the transcriptional activity with a value of 0.86 and finally, the 300bp promoter has a value of 1.34.

This showed that promoter length does matter for transcriptional activity. The activity from 100bp to 200bp more than doubled and when increased to 300bp, led to a 58% increase in activity for a length increase of only 50%.

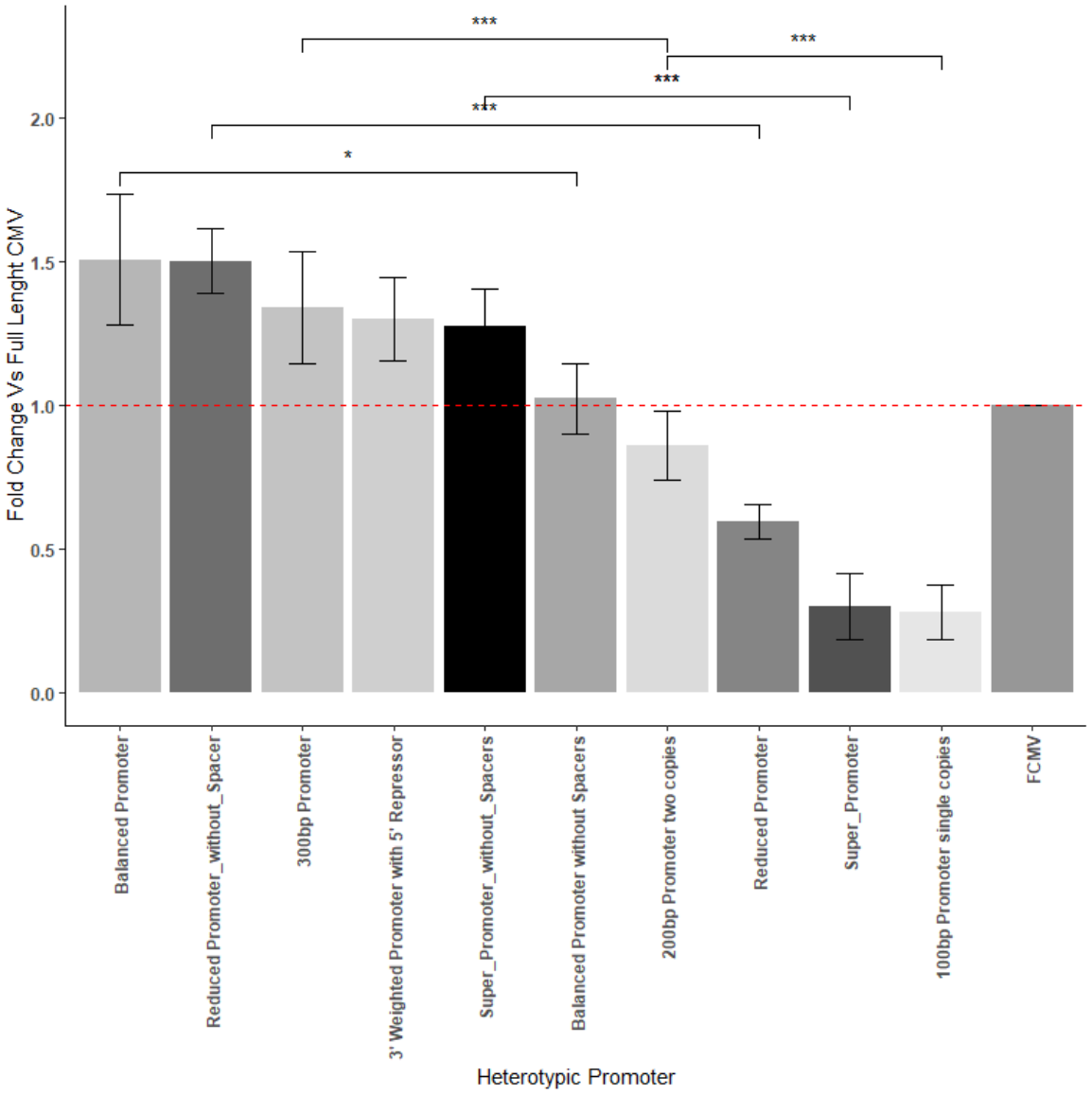


Figure 5.3: The inclusion of NRF1 has led to an overall decrease in maximum activity compared to what was seen in library 1. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest (n = 3 ± sd). The main points investigated in this library were the effect of spacers, promoter lengths and another attempt at the use of repressors. The statistics for the main points above are plotted on this graph. The statistics presented on the graph indicate the following p-values based on a unpaired two-tailed t-test: * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$ and NS = not significant.

5.3.1 Conclusions on Library 2

This library looked at expanding the design space of heterotypic promoters by trying to understand the variables of length and spacers. The most important observation from this library is that the inclusion of the new TFRE NRF1 did not increase the activity of the promoters in a heterotypic context. This suggests that TFREs tested in a heterotypic context may alter functions in unexpected ways due to transcription factor-transcription factor interactions or also having the potential to bind other non-canonical binding sites, thus leading to different activities (Biswas and Chan, 2010).

This library also shows that promoters under 200bp struggle to overcome the activity

of full-length CMV, although "Reduced Promoter_without_Spacer" actually reached a normalised value of 1.5 times full-length CMV, which was just as active as the "Balanced Promoter" in which it shared the same composition but just had half the amount of TFREs contained within it.

This is why predicting sequence activity using equations is so difficult. If one were to take the "Balanced Promoter" and the "Reduced Promoter_without_Spacer" it would be expected that they have an outcome that is not similar based on the work of (Brown et al., 2014) and (Johari et al., 2019). This is not the case as there are underlying mechanisms that are currently not understood. The difference between the sequences is a 200bp length difference and the GC content of the "Reduced Promoter_without_Spacer" is 60% versus 50%.

Lastly, from this library, it can be further clarified that using repressors on the 5' end of sequences can be detrimental to their unidirectional activity, as was seen with the "3' Weighted Promoter with 5' Repressor". The sequence went from having an approximate fold change of 1.7 fold to approximately 1.3 fold with the addition of repressors on the 5' end.

In summary, the results have led to the following considerations for synthetic promoter design:

- The inclusion or exclusion of spacers can have either a positive or negative effect on transcriptional activity. It is generally better not to include them to reduce promoter length and synthesis cost, but for informational purposes, it is worth trying both.
- Promoter lengths appear to affect transcriptional activity, although sequences over 200bp in length can rival or overcome the activity of FCMV.
- Further emphasised that using repressors to try and funnel transcription does not work in this context.

5.4 Unidirectional Library 3

At this point in the design flow, all of the information from the TFRE Discovery Pipelines had been collated. This library incorporated the new potential sequences found, such as NFE2l2, EGR1 and Clock Bmal. There were 3 general objectives of this library. The first was to see how the new homotypic sequences act in a heterotypic setting. The second was to see if changing the weak sequences in previous promoters to new stronger ones increased the transcriptional activity. Finally, to see if promoters built from entirely new sequences could match or even beat the activity of previously tested synthetic promoters.

Table 5.3 shows the sequences that were created for library 3 testing. The section design system 1 focused on merging the previous literature with newly derived sequences to create promoters containing new and old sequences. The only addition to this section, compared to the last section, is NFE2l2 which was found to be the best completely new TFRE sequence.

The first sequence, "NFE2l2_3' Weighted Promoter" was an attempt to see if using NFE2l2 as the 3' weighted TFRE along with sequences such as NFkB could be the same or greater activity be achieved when compared to FCMV. This was tested with and without a spacer as in the previous section, it was seen that either could be beneficial.

The "Even_Promoter" and "Even_Promoter wo spacer" are a general dispersion of TFREs across a whole promoter sequence to see what activities a more heterotypic promoter may achieve when the new building block is included.

The NFkB and "NFE2l2 Absent Promoter" attempt to see how NFE2l2 acts in a heterotypic context. All the NFkB sites in the "NFkB Absent Promoter" have been replaced with NFE2l2 and vice versa in the "NFE2l2 Absent Promoter".

The "5' homotypic NFE2l2 on 100bp Promoter" was an attempt to see what would happen if you put the homotypic tested previously in front of the relatively weak 100bp promoter tested in the previous section.

Design system 2, took a different approach. In this section, the promoters "Balanced Promoter" and "Reduced Promoter_without_Spacer" were taken from library 2. The binding sites were replaced as follows, Ahr/ARNT was replaced with NFE2l2. AARE was replaced with EGR1 and NRF1 was replaced with NFE2l2. This was done to check if substituting what is considered weak sequences in the synthetic promoters with stronger ones made a difference to the activity previously measured.

The final section was design system 3. These promoters were created with the simple objective of creating promoters containing no TFREs from literature. The idea to vary the length was to try and titrate the activity of the new synthetic promoters. The sequences used in this library are shown in Appendix A.6.

Table 5.3: Sequence information and composition for library 3.

Name (Design System 1)	**Length**	**NFkB**	**GABP beta**	**DMP1**	**ARE**	**AhR/ARNT**	**HRE**	**AARE**	**NRF1**	**GC BOX**	**YY1**	**NFE2l2**
NFEL2_3' Weighted_Promoter	377	4	4	2	2	2	2	2	2	2	1	4
Even_Promoter	460	3	3	3	3	3	3	3	3	3	3	3
NFkB Absent Promoter	329	0	4	2	2	2	2	2	2	2	1	4
NFE2l2 Absent Promoter	321	4	4	2	2	2	2	2	2	2	1	0
NFEL2_3' Weighted_Promoter without spacer	325	4	4	2	2	2	2	2	2	2	1	4
5' homotypic NFE2l2 on 100bp Promoter	204	1	1	1	1	1	1	1	1	0	0	6
Even_Promoter wo spacer	396	3	3	3	3	3	3	3	3	3	3	3
Name (Design System 2)	**Length**	**NFkB**	**GABP beta**	**DMP1**	**ARE**	**NFE2l2**	**HRE**	**EGR1**	**CLOCK_BMAL**			
Balanced_Promoter(Weak Sequences Replaced with new ones	468	6	6	4	4	4	4	2	5			
Reduced_Promoter_Without_Spacer (Weak Sequences Replced with new ones)	218	3	3	3	2	2	2	1	3			
Name (Design System 3)	Length	NFE2l2	v$gabpb1.01	EGR1	V$CLOCK_BMAL1.01	V$ETV4.01						
New_Sequence_Promoter_200bp	208	4	4	2	2	2						
New_Sequence_Promoter_300bp	358	6	6	4	4	4						
New_Sequence_Promoter_Max	477	7	7	6	6	6						

Figure 5.4 shows the results for heterotypic library 3. What can be seen in this library is that the new TFREs being included have led to the creation of sequences which had higher maximum activity than the previous two libraries. The sequence "Reduced_Promoter_Without_Spacer (Weak Sequences Replaced with new ones)" had an average activity of 1.8 fold versus CMV, while the best promoter in library 1 had an activity of 1.7 fold versus CMV. The difference between the two promoters, although is not statistically significant ($p = 0.27$). This indicates that the inclusion of new TFREs in this library appears to have increased the maximum capacity of the sequences tested.

By replacing some of the weaker sequences in two promoters from library 2 (Design System 2), activity has risen from approximately 1.5 fold versus CMV to 1.8 fold and 1.55 fold. The "Reduced_Promoter_Without_Spacer" from library 2 to this new version in library 3 has significantly increased ($p < 0.05$), although the "Balanced Promoter" has not.

Interestingly in library 3 the promoters that contain spacers and don't contain spacers show no statistically significant difference between them. The "NFE2l2_3'_Weighted _Promoter" with spacers and without spacers has a p-value of 0.36. The "Even_Promoter" with and without spacer has a p-value of 0.29. This throws doubt on if spacers are beneficial overall or if using them is a poor use of resources. Likely, what is occurring in this instance, compared to library 2 is no new beneficial or negative TFREs are being created by their inclusion or exclusion.

The NFE2l2 and NFKB absent promoters show that based on the average fold change, the NFE2l2 absent promoter (The promoter with NFkB sites within it) has higher activity than the NFkB absent promoter. The difference however is not statistically significant ($p = 0.0523$).

The "NFE2l2 Absent Promoter" provides a 0.37 fold change over the "NFkB Absent Promoter" promoter. The "NFE2l2 Absent Promoter" has a statistically significant difference compared to FCMV ($p < 0.001$) while the "NFkB Absent Promoter" did not ($p = 0.58$). This indicates that although when considered in a homotypic context, the two TFREs showed rivalling activity, in a heterotypic context, it can be suggested that NFkB provides greater activity to the heterotypic promoter than NFE2l2. This can be further seen in the improved sequences from library 2 where adding the new sequences, along with historic sequences such as NFkB has led to the maximum fold change being reached.

The completely new sequences (Design System 3), denoted by New Sequence in front of their names, showed overall poor activity. Interestingly, the 200bp promoter performed better than the 300bp version, which is surprising. Also, the max version, which is 477bp in length, has no significant difference from full-length CMV. This suggests that the new building blocks that have been found cannot fully support transcription alone to overcome the activity of CMV, but can at least match CMV with reduced diversity.

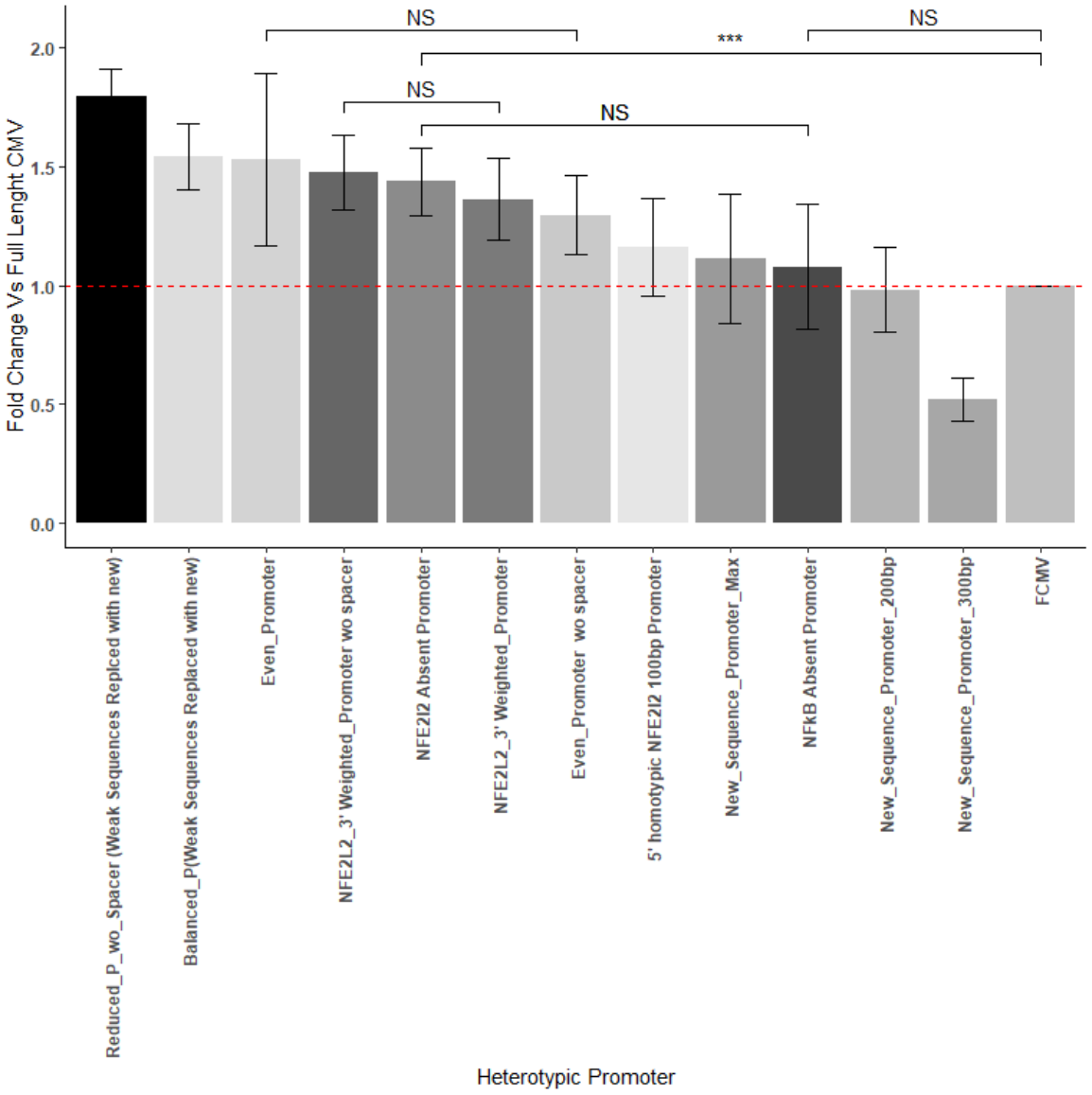


Figure 5.4: The combination of literature-derived TFREs and the new TFREs has led to a promoter with the greatest activity of the 3 libraries tested. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest (n = 3 ± sd). Sequences with a combination of new and old TFREs show high activity. The sequences that are only composed of completely new sequences are not as active, just matching the activity of FCMV. The statistics presented on the graph indicate the following p-values based on a unpaired two-tailed t-test: * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$ and NS = not significant.

5.4.1 Conclusions on library 3

This library's main goal was to incorporate the newly found TFREs into the synthetic promoter design space. The results show the inclusion of the newly found TFREs can further enhance the activity of promoters and has led to the highest average fold change being achieved for a promoter.

The attempt to create entirely new sequences showed that reduced diversity of the TFREs in a heterotypic context cannot match the transcriptional activity provided by mixing literature-derived TFREs with the newly defined TFREs. The new sequence

promoters also showed an interesting trend of the 200bp promoter being stronger than the 300bp promoter.

Finally, spacers were again included in this library to check if there was any general trend. Unlike the previous library, there was no difference between the promoters that had spacers and those that didn't. This suggests that the spacers effect is neither beneficial nor disadvantageous overall.

Overall, the results in this section has shown 3 key design considerations. These are:

- The inclusion of NFE2l2 can benefit transcriptional activity.
- Using just the newly discovered sequences is not enough to overcome the transcriptional activity of FCMV.
- The use of spacers appears to be inconsequential. Either including them or excluding them can have a positive or negative effect.

5.5 The Expansion of Synthetic Promoter Designs

This study's overall objective was to test the current design space of synthetic promoters and see if there is still room for expansion or if there are any novel design considerations that can be considered in future works.

Library 1 focused on the novel use of literature-derived TFREs and seeing if including stability elements, weighing the promoters and using repressors could aid in achieving higher transcriptional activity than previously seen. What was found was that 3' weighting can potentially help encourage high transcriptional activity but only provided the same transcriptional output as the two benchmark sequences.

Library 2 tried to incorporate the newly found TFREs while also testing other design considerations. The newly found TFRE NRF1 was included, but this seemingly reduced the overall activity of library 2. The library also showed that depending on composition, sequences as small as 200bp-300bp are enough to achieve a 1.5 fold change versus FCMV.

Library 3 ultimately tried to incorporate the newly found TFREs in various ways to see if they could aid in future synthetic promoter construction. The inclusion of these TFREs can be beneficial but when completely new sequences are used which contain none of the previously derived TFREs the activity could not overcome full-length CMV. The addition of NFE2l2 to previously tested sequences increased the activity of the promoters.

This work has expanded the design space for synthetic promoters by providing some informative designs based on weighting and no longer needing to use spacers in synthetic promoters. It has also aided designs by contributing to the incorporation of newly found TFREs into a heterotypic design which, when tested in stable cultures may have a greater effect than what was seen in transient cultures.

The design space in future works could be further expanded by large-scale synthetic promoter studies. To truly understand the mechanisms or be able to engineer synthetic promoters to a greater degree in the future, a larger scale study is needed. Perhaps the most application-based way to do this would be to design hundreds if not thousands of synthetic promoters and transfect all of them into CHO cells along with GFP. They could then be sorted using FACS and the promoter sequences discovered using an RNA-seq tag. By doing this, a machine learning algorithm such as XGBoost could be used to decipher the patterns within the promoter sequences and build a model that could more accurately predict the activity before testing.

This could be further expanded on to use deep learning. A convolutional neural network could be used to pick up patterns in the promoters and discover non-expected regulatory regions in the sequences. This, albeit, would require much more data. The future of synthetic biology will be the exact prediction of what sequences do or what function they may have prior to laboratory testing. Currently, the only company making headway in this area to a truly meaningful degree is Deepmind which has come out with a model which can accurately predict the effect of single nucleotide polymorphisms on gene expression in humans (Avsec et al., 2021).

Chapter 6

Applying Synthetic Unidirectional and Bidirectional Promoters to the Production of a Recombinant Antibody

Overview

- This study applies learnings from the previous chapters to create a library of unidirectional and bidirectional constructs expressing an antibody.
- The unidirectional golden gate design is summarised in Section 6.2.1. This system allowed the quick creation of any promoter combination required for experimentation.
- The design decisions behind the promoters selected for the unidirectional constructs are discussed in Sections 6.2.2, 6.2.3 and 6.2.4.
- As no pre-screening occurred for the bidirectional promoters, it was tested directly in an antibody-producing context. The thought behind the design of the sequences and how bioinformatics was used to leverage the designs are discussed in Section 6.3.
- Lastly, some real system data will be addressed from Merck's onboarding process. This data was performed by the team at Merck and was an attempt to see how the constructs perform in their real-world stable expression systems. This data is currently being collated and is not fully available for the submission of this work.

6.1 Introduction

The previous chapters have set out how to expand synthetic promoters' design space and try to create new variants of synthetic promoters but have not dealt with the production of a recombinant antibody.

Producing an antibody is much more complicated than just creating a synthetic promoter with higher expression than FCMV. Dynamics such as promoter, promoter interference, translational bottlenecks and in stable cultures, epigenetic silencing come into play. For that reason, it was essential to test how these newly created synthetic promoters could be used to create an antibody called avelumab. The sequence of avelumab was provided by Julien Douet at Merck and was optimised for CHO. One important note is that the antibody sequences used for the RNA-seq analysis and the avelumab antibody are different. The sequences of the RNA-seq antibody were not allowed to be shared for synthesis in Sheffield.

The literature is scarce on using synthetic promoters to create vectors expressing an antibody and as such, the dynamics are currently unknown. It is thought that excess light chain is beneficial for protein production (Schlatter et al., 2005) and synthetic promoters give an excellent tool for alternating the amount of heavy and light chain to be produced.

The other important factor when considering vector construction is vector size. As discussed in the literature review (Section 1.5.1), many of the promoters that drive the core genes in the genome are bidirectional. The idea that synthetic bidirectional promoters could be created was with the hopes that this could lead to smaller vectors that don't suffer from the drawback of promoter, promoter interference.

The screening process has been undertaken in transient systems and as such, to asses the constructs in a real world application, it was important that the constructs were tested in a situation as close as possible to a stable system. For that reason, all of the transient vectors contained the glutamine synthetase (GS) marker and SV40 promoter. The most interesting question within the stable context is if GS selection ultimately reverts the changes that are seen in the transient culture and as such no stable synthetic constructs are better than FCMV.

6.2 Unidirectional Antibody Library

6.2.1 Designing the Golden Gate System

Due to the cloning strategy used in the previous libraries, all sequences shared the same restriction sites of KPNI and HINDIII. Also, with the nature of creating an antibody, at least two chains are needed. The heavy chain and the light chain. Expressing an antibody construct takes at least two promoters. For this reason, a golden gate system was decided on as it would easily allow the mix and matching of many different promoters.

The first goal of the system was to involve no-site directed mutagenesis for the promoter regions. The reason for this is due to the repetitive nature of the promoters. The PCR can become error-prone. Golden gate avoids these issues as it is a ligation-based system. Each golden gate vector was sequenced with 7-8 overlapping primers to ensure there were no mutations in the sequence using Sanger sequencing.

To create a system that is flexible enough to create any required construct, a unique golden gate system was created, which used two delivery vectors called "Light Chain Promoter" and "Heavy Chain Promoter". Figure 6.1 shows the vector for the light chain promoter. These destination vectors had the required overhangs to ligate in the heavy chain promoter region or light chain promoter region of the penultimate golden gate vector. The required promoters were restriction digested and ligated into these vectors to be used in the golden gate mixture. The vector already had a CMV core and the golden gate sites at the 5' and 3' ends before the promoter region and after the CMV core.

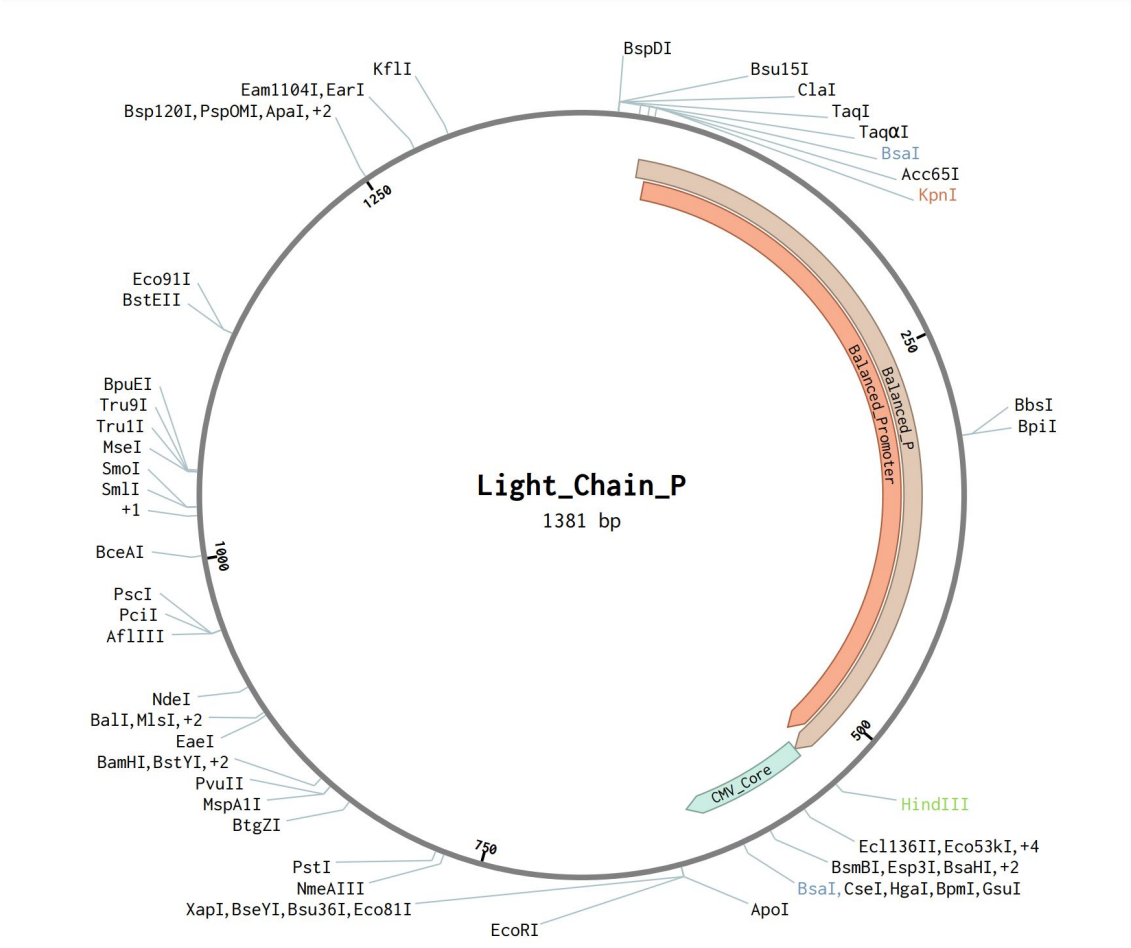


Figure 6.1: Light Chain promoter destination golden gate vector. This figure shows the light chain promoter vector for the unidirectional golden gate system. The restriction sites KpnI and HindIII are shown. The golden gate sites are depicted by the BSAI sites at the 5' and 3' end of the sequences. The Light Chain promoter vector has overhangs which ligate the promoter in front of the light chain during the golden gate reaction.

Apart from the novel destination vectors, the rest of the golden gate system was designed using Benchling and NEB GoldenGate. Table 6.1 shows the overhangs that were used for the unidirectional system. The genetic parts were synthesised with the needed golden gate overlaps between them so that they could be used as soon as they arrived. The heavy chain and light chain sequences of avelumab were provided by Merck and synthesised at Genewiz. Very little DNA is needed for the golden gate reaction (75ng)and as such, the lyophilised DNA sent from Genewiz was adequate for many reactions.

The final design consideration included an SV40 promoter and GS selection marker. The reason for this was to account for promoter interactions, even in the transient culture, but also to ensure the vectors would be ready for stable transfections as soon as the transient screens were completed.

Table 6.1: Overhangs and parts used for the unidirectional golden gate system.

Upstream	Downstream	Bases	Reverse Complement
Linear_Backbone	Light_Chain_P	GATA	TATC
Light_Chain_P	LC_Gene	TCGC	GCGA
LC_Gene	HC_Promoter	ATTC	GAAT
HC_Promoter	HC_Gene	AAGA	TCTT
HC_Gene	SV40_Linear	TAGT	ACTA
SV40_Linear	SM_Gene	TGTT	AACA
SM_Gene	Linear_Backbone	CGTG	CACG

6.2.2 Information from the Bioinformatic Analysis

To asses how to create the unidirectional promoter combinations the first analysis that was done was to look at the RNA-seq data. As the heavy and light chains were added to the CHO PICR gtf and fa files, the expression of both could be measured. By looking at this, it was hoped that it could indicate that more heavy or light chain is required for protein production in the clones.

The first sample to be looked at was the Mock to get an idea of GS expression throughout the culture. The results can be seen in Figure 6.2. As can be seen, the Heavy chain and Light chain have minimal expression as expected. Although it was not zero, the TPM of both heavy and light chains was above 50 in both instances. This may indicate that the cells already contain sequences very similar to the heavy and light chain sequences provided by Merck.

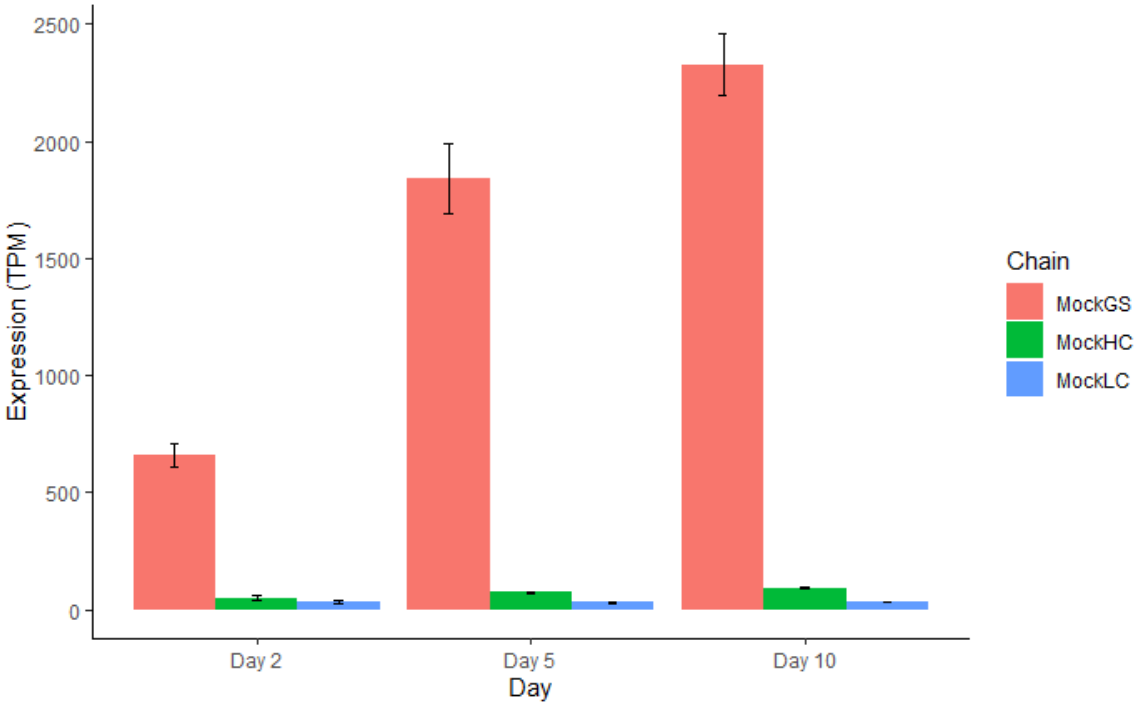


Figure 6.2: Expression of the heavy chain, light chain and GS in the Mock Clone. The y-axis shows expression measured in TPM and the x-axis shows the sample day. What can be seen is that the GS expression is relatively high and throughout culture, it increases in expression. The heavy chain and light chain were expected to show no expression. The error bars of standard deviation from n=3 biological replicates.

Clone 3 was producing a recombinant antibody that Merck uses for process optimisation. Figure 6.3 shows the data obtained for this. What can be taken from this clone is it appears that over day 2, day 5 and day 10, there is an excess of light chain over heavy chain. However, when the average TPM across the days is compared there is no statistical significance ($p = 0.5$), but it may suggest that this clone is producing an excess of light chain, especially on day 2 and day 10. Clone 3, according to Merck is an unstable Clone with a tendency to lose expression and cannot be considered alone.

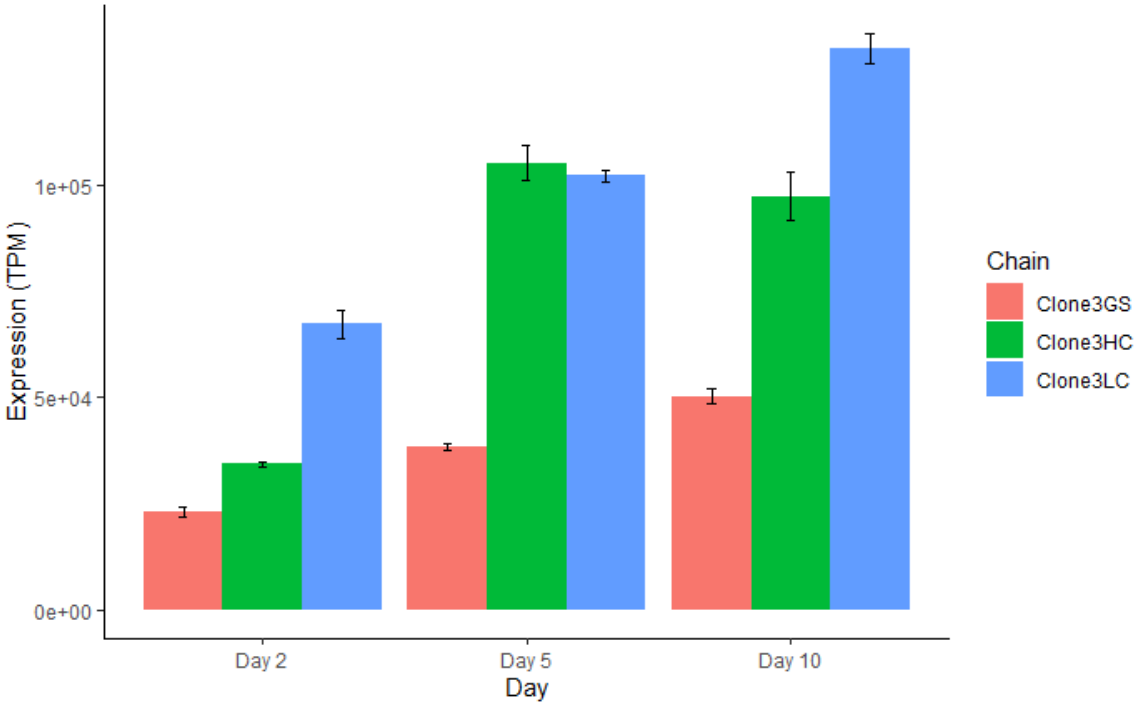


Figure 6.3: Expression of the heavy chain, light chain and GS in Clone 3. The y-axis shows expression measured in TPM and the x-axis shows the sample day. The error bars are standard deviations with n = 3 biological replicates. What can be seen in general is that the gene for light chain expression has a higher expression each day compared to the heavy chain.

Looking toward Clone 9, which is a highly stable producer, what can be seen is a reverse in the trend. Here the heavy chain has a higher average expression on day 5 and day 10. This is not statistically significant (p= 0.75). The data in comparison to Clone 3 may show that it is worth trying constructs which use excess heavy chain. Although this goes against most previous literature (Carrara et al., 2021) there is literature that suggests excess heavy chain can be good for transient production (Li et al., 2007).

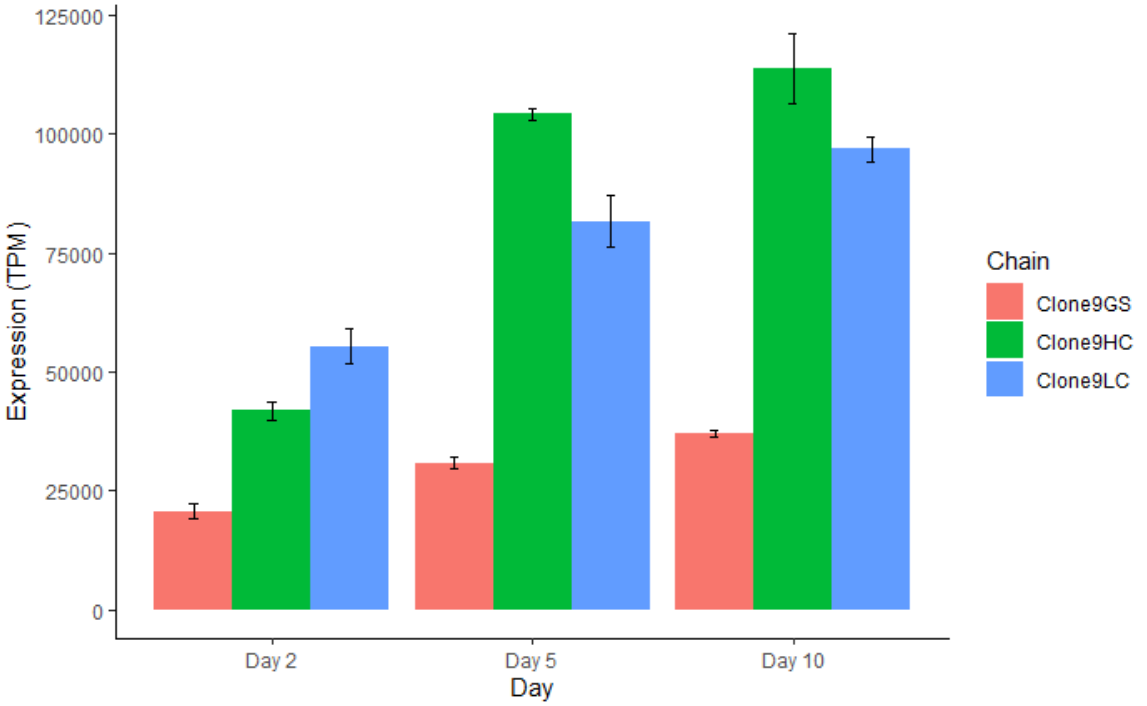


Figure 6.4: Expression of the heavy chain, light chain and GS in Clone 3. The y-axis shows expression measured in TPM and the x-axis shows the sample day. The error bars are standard deviations with n = 3 biological replicates. The data shows a higher expression of heavy chain on day 5 and day 10. On day 2 the reverse trend is seen and the light chain is in excess.

Looking at the above results, the idea of heavy or light chain excess is unclear and for that reason, both would be tested to see if excess light or heavy chain matters for recombinant protein production. One of the more interesting observations from this section was the difference in GS expression between the Mock, Clone 3 and Clone 9. Although the Mock underwent the same selection process only had a GS expression level measured in TPM of approximately 1500 TPM. In Clone 3 and Clone 9 these values are 37,000 and 30,000, respectively.

This could indicate that although they underwent the same selection process, the Mock, for some reason, produces much less glutamine synthetase. This could be due to the Mock not having the added pressure of making a recombinant antibody and having fewer selection criteria post-selection. For instance, Clone 3 and Clone 9 would have been picked for high productivity, but this could not have been done for the Mock.

6.2.3 Library 1 Design and Results

In terms of design, there were 2 main objectives. The first was to assess how synthetic promoters act in a similar fashion to how CMV is currently used. The next was to evaluate if using different promoters on the heavy and light chain could impact the titre produced from the vectors.

As mentioned previously, the vector was assembled using the golden gate cloning and all of the unidirectional promoters were also tested in the presence of SV40. Figure 6.5 shows an example vector used. The construct went light chain promoter - light chain - heavy chain promoter - heavy chain - SV40 promoter and the GS gene. Although GS is not required for transient transfections, its promoters presence is crucial as it will affect the synthetic promoter through promoter interaction or cellular resource usage.

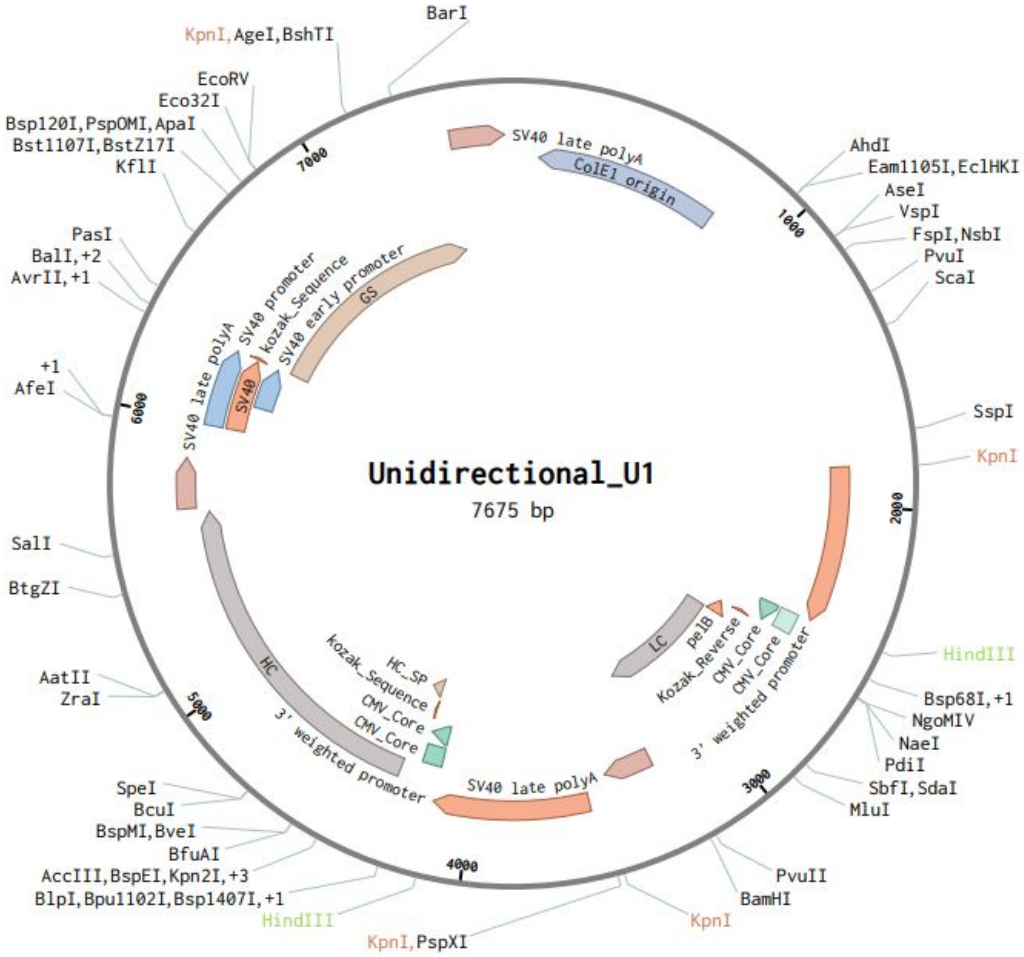


Figure 6.5: Fully constructed golden gate vector for construct U1. This construct had the 3' weighted promoter on both the heavy and light chains. The construct also had an SV40 in front of the GS. The figure is taken from Benchling.

The unidirectional library was designed to test both the same promoter and a mix of promoters on the heavy and light chains. Constructs U1 to U7 as shown in Table 6.2 were chosen based on picking different strength promoters and testing them on both the heavy and light chain. Construct U8, U9 and U11 were an attempt to produce excess heavy chain compared to light chain. Constructs U10 and U12 were created

to test whether the excess light chain made a difference in expression. At this point in testing with an antibody, only heterotypic library 2 was completed and promoters including NFE2l2 such as the "Reduced_Promoter_Without_Spacer (Weak Sequences Replaced with new ones)" were not included in this library.

Table 6.2: Constructs created for unidirectional antibody library 1. This table shows the code name of each construct, along with the promoter allocation. The first promoter is the heavy chain and the second is the light chain. If only one promoter is listed, it is used on both chains.

Construct Name	Promoter Allocation(HC/LC)
U1	3' Weighted Promoter
U2	Reduced_Promoter_Without_Spacer
U3	300bp Promoter
U4	Balanced Promoter
U5	Balanced Promoter without spacer
U6	Super_Promoter
U7	100bp Promoter single copies
U8	3' Weighted Promoter + 300bp Promoter
U9	3' Weighted Promoter + Super_Promoter
U10	Super_Promoter + 3' Weighted Promoter
U11	3' Weighted Promoter + Balanced Promoter without spacer
U12	Balanced Promoter without spacer + 3' Weighted Promoter
U13	FCMV + FCMV

Sequences were transfected and assayed as described in Sections 2.1.6 and 2.3.2 of the Materials and Methods. The only difference was that after supernatant extraction, the cells were spun down for ddPCR to investigate the constructs' heavy chain light chain ratio.

Figure 6.6 shows the results obtained for library 1. What can be seen is that in the majority of cases the average expression of the synthetic promoters is higher than CMV. All constructs except U4 and U2 have a $p < 0.05$ when tested with an unpaired two-tailed t-test. As expected U7, which was two 100bp promoters has lower activity, than CMV, which was anticipated as the promoter, when tested using SEAP showed much lower activity than CMV.

Overall, the highest result obtained was a value of 2.52 fold higher than the CMV control with a p-value of less than 0.001. Other constructs such as U11, U12 and U10 all achieved a fold change of over two also. Interestingly, it appears from the data that having two different synthetic promoters appears to provide a greater increase in titre than using the same promoter on both heavy and light chain. U3 achieved the highest results for a construct with the same promoter on both chains and was 1.95 fold higher than the control. This was surprising as it was not the highest activity promoter when tested in the previous SEAP experiments. When compared to the highest construct U8 there was a statistically significant difference ($p < 0.05$)

Comparing U8 to U9 ($p < 0.05$) it appears that the inclusion of a much weaker promoter on the light chain caused a loss in protein titre. To test what would happen if promoter were just swapped from the heavy chain to the light chain construct U9, U10, U11 and U12 were created. Surprisingly changing the promoters around had no statistically significant effect on the fold change in either comparison (U9 versus U10 and U11 versus U12). What this may indicate is that the promoters are acting much differently than how they performed during the heterotypic SEAP screen and their transcriptional output is now much higher or lower.

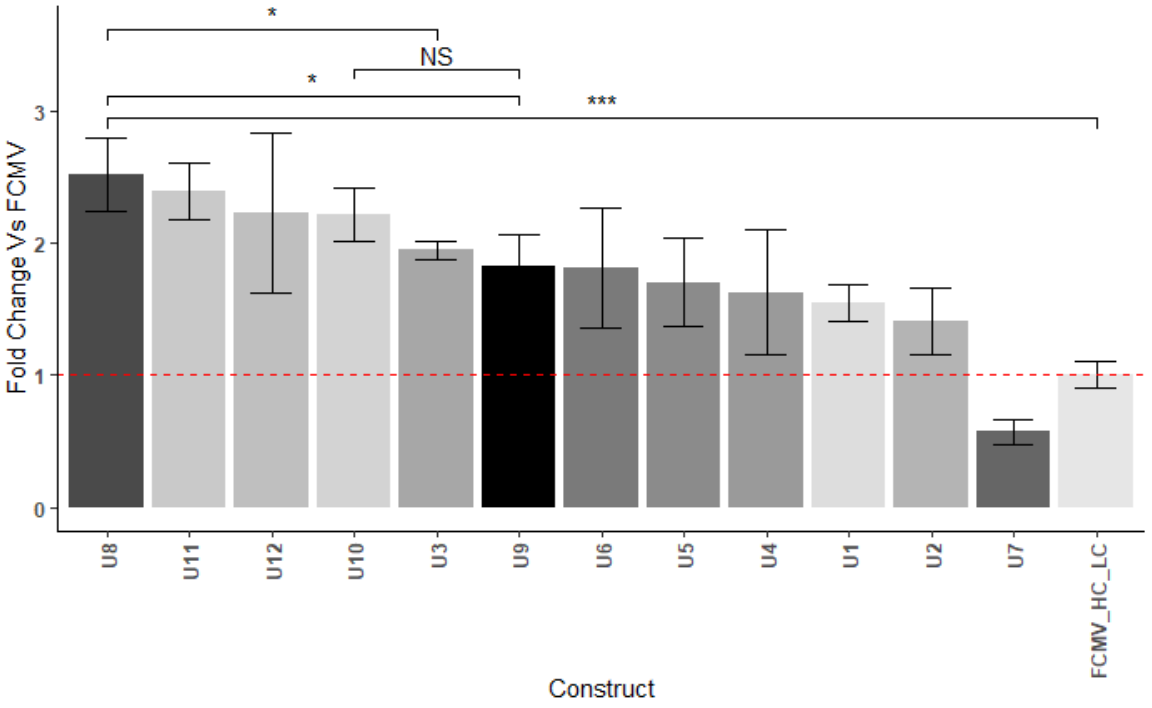


Figure 6.6: The combination of promoters is beneficial in unidirectional antibody library 1 normalised to FCMV. The results above were obtained from testing the constructs using a heavy chain and light chain of the Avelumab antibody. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest ($n = 3 \pm$ sd). The results indicate, in general, that the synthetic promoters outperform the control in transient expression in CHO cells. The results also show that the constructs that have achieved the highest fold change have a different synthetic promoter on the heavy and light chains. The statistical testing was an unpaired two-tailed t-test. The statistics presented on the graph indicate the following p-values based on a unpaired two-tailed t-test: * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$ and NS = not significant.

To get a better idea of what is occurring for each construct ddPCR was carried out. For the unidirectional library U8, U11, U12, U10, U3, U9, U6 and U1 were tested. Figure 6.7 shows the heavy chain light chain ratio (HC/LC) (This is the amount of heavy chain divided by light chain). What can be seen is that overall it appears in this instance that there is little correlation between the ratio of heavy and light chain production and the overall titre. Neither having an excess of heavy chain or a ratio of 1:1 appears to affect the fold change versus FCMV.

One interesting observation from this data is that no matter the promoter selection, there is always an excess of heavy chain. Even when using two CMV promoters,

the HC/LC ratio was still 1.47:1. The only instance where this did not occur was U12 which had a ratio of 0.99:1. One possible explanation for this is that inefficient termination is occurring. The RNA Pol II is running on from the light chain and begins transcribing the next gene (Proudfoot, 2016), or perhaps the presence of RNA Pol II at the termination region of the light chain gene allows the heavy chain promoter to steal these RNA polymerases and switch through the transcription cycle more efficiently.

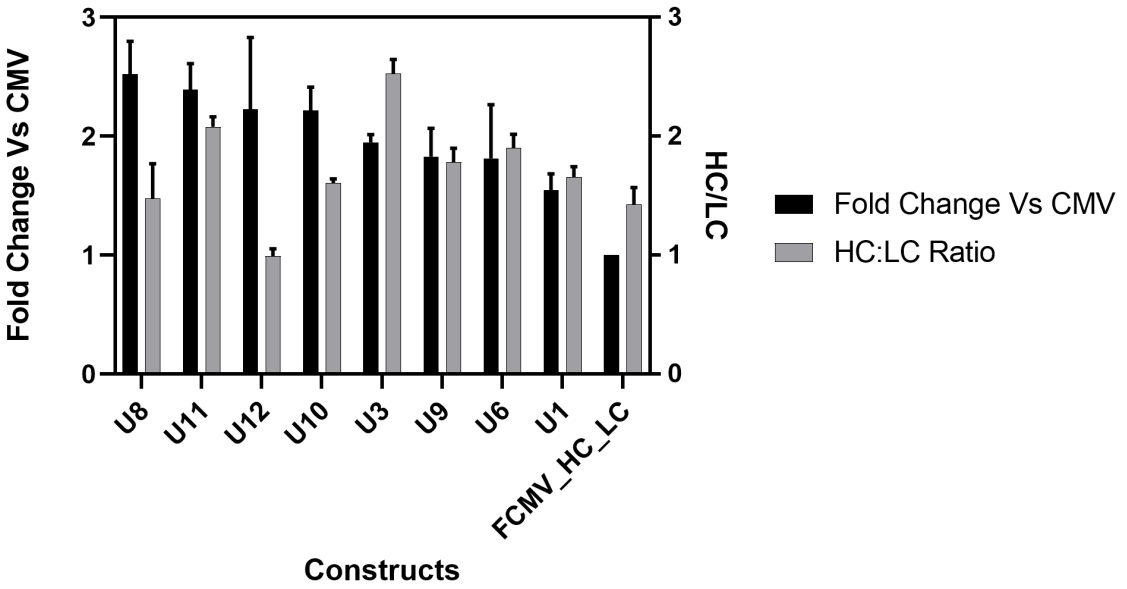


Figure 6.7: An investigation of the HC/LC ratio between a selection of promoter constructs. The fold change shown is normalised to the CMV control, as shown in Figure 6.6, with the heavy chain/light chain ratio. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest (n = 3 ± sd). There is no correlation between the HC/LC ratio and the fold change achieved. The attempts to change the ratio of HC/LC ultimately appear to have failed as can be seen in U11 and U12, along with U9 and U10. The heavy chain is always at least at a 1:1 ratio. This may indicate inefficient termination is occurring.

6.2.4 Library 2 Design and Results

Due to heterotypic library 3 being incomplete when the unidirectional library was first constructed, some promising promoter sequences were left out. For this reason, another library was created which attempted to improve the best constructs from the previous library by substituting the newly found promoters, which contained the newly found NFE2l2 TFRE.

Table 6.3 shows the sequence combinations used in library 2. They are modifications of the U8 construct from the previous section. U14 was an attempt to see if replacing the 3' weighted promoter with a similar strength promoter would increase expression. U15 was a copy of the U8 construct but used sequences completely from library 3. U16 was just changing the light chain promoter and U17 was the same but attempting to increase light chain expression. Finally, U18 was just another variant using the 300bp promoter on the light chain and the new Even_Promoter on the heavy chain.

Table 6.3: Constructs created for unidirectional antibody library 2. This table shows the code name of each construct, along with the promoter allocation. The first promoter is the heavy chain and the second is the light chain. If only one promoter is listed, it is used on both chains.

Construct	Promoter Allocation (HC/LC)
U14	Reduced_Promoter_Without_Spacer (Weak Sequences Replaced with new ones) + 300bp Promoter
U15	Reduced_Promoter_Without_Spacer (Weak Sequences Replaced with new ones) + Even_Promoter
U16	3' weighted Promoter + Even_Promoter
U17	3' weighted Promoter + Reduced_Promoter_Without_Spacer (Weak Sequences Replaced with new ones)
U18	Even_Promoter + 300bp Promoter

The results from library 2 showed no improvement over the previous library, as shown in Figure 6.8. The highest result achieved was 2.13 fold higher than the CMV control, which was significant($p < 0.05$). What this data shows is that by changing the promoters, even those which were deemed equivalent when tested in a heterotypic context using SEAP, it is hard to predict what the outcome will be. Even changing the light chain promoter as can be seen in U16 and U17 can dramatically affect the final titre of the construct. Maxing out the heavy chain and light chain with the two best promoters in the heterotypic library (U15) only showed an average activity of 1.73 times the control. This data suggests that although the library was designed with the idea of varying the strengths of the heavy and light chain promoters, it is still worth testing random combinations as the outcome is hard to predict. This can be seen when U8 and U14 are compared. The heavy chain promoter was changed like for like and the average fold change versus the control dropped from 2.5 fold to 2.1 fold.

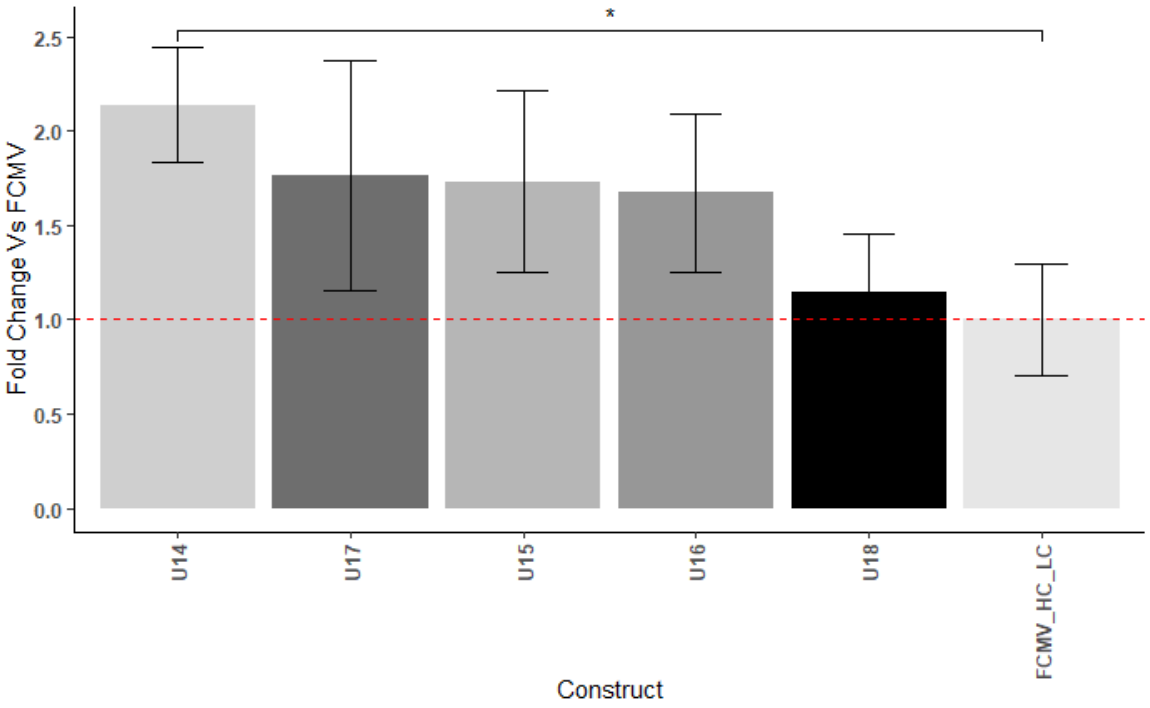


Figure 6.8: Creating variations of the U8 constructs diminished IgG production in unidirectional antibody library 2 normalised to FCMV. The results above were obtained from testing the constructs using a heavy chain and light chain of the Avelumab antibody. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest (n = 3 ± sd). The results show that replacing either the heavy chain or light chain promoters with another equivalent promoter does not retain the activity of the initial construct. The statistics presented on the graph indicate the following p-values based on a unpaired two-tailed t-test: * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$ and NS = not significant.

6.2.5 Observations from the Unidirectional Antibody Library

The two main objectives of this library have been achieved. The first was to see how the synthetic promoters perform in a context similar to CMV. In general, it appears that even using synthetic promoters in a similar fashion of having both the same promoter on the heavy and light chain works better than the traditional CMV system it was tested against.

Unfortunately, it is not the case to use the "best" promoter on both the heavy and light chains. What can be seen from Figure 6.6 is that the best promoter when tested on both the heavy and light chain did not do as well as using the 300bp promoter on both the heavy and light chain. This could be due to some promoters containing the same TFREs which are contained in SV40, or it could even be due to the inclusion of extra TFREs present in the heterotypic library 2 promoters.

For industry, it could be suggested that using the same promoter on the heavy and light chain provides simplicity. However, for maximum optimisation in this work, a combination of promoters appears beneficial for the highest protein titre. This is beneficial for industry partners hoping to create multiple different molecules as between different recombinant molecules, fusion molecules and even antibody drug conjugates it is vital to alter the expression of different chains to ensure optimal expression.

This application could be achieved by a considerable amount of additional work to the clonal selection process. Still, a workaround would be to add a genetic tag to each construct and make every possible combination from the library for a targeted molecule, be that 2, 3 or even 4 chains that need expression. From there, a pooled transfection would occur where a plasmid mix is created, which contains all the possible constructs. Once selection is applied and cells are selected for high productivity, the barcode could be sequenced to find out which construct provided the highest amount of expression.

The unidirectional library 2 was an attempt to incorporate the newly found promoters generated during unidirectional library 3. These sequences had new TFREs contained within them, such as NFE2l2. The idea was to take the U8 construct and try and improve it. Surprisingly, the results were worse than the original construct. This library suggests that it is impossible to design the antibody expressing constructs based on data from expressing SEAP. The ddPCR data also suggests this. The intended designs, such as excess heavy chain or light chain did not work; nearly every construct had more heavy chain no matter what promoters were used. Library 2, especially construct U14 and U16 showed that swapping promoters, like for like can negatively impact the final titre compared to the CMV control.

Overall, the results have fulfilled the objectives of creating unidirectional synthetic constructs which produce more antibody than the CMV control in a transient system. It has also given more context to what should be considered when using synthetic promoters in this context, such as randomly combining the promoters is still the best approach to get an optimal outcome and the expression measured from the SEAP assays does not necessarily correlate with the titre that will be achieved by that promoter

when tested with an antibody.

6.3 Bidirectional Library

Bidirectional promoters have only one enhancer region but transcribe in both directions. These could provide many benefits for industry, especially in the application of synthetic promoters as they would allow the avoidance of potential promoter-promoter interference and also it could be used for the rapid screening of the optimal heavy and light chain ratios for molecules as described in Vogl et al. (2018).

To create the synthetic promoters for the antibody screen, two approaches were taken. The first was an attempt to utilise bioinformatic data to decipher what is naturally in bidirectional promoters and the second was attempting to use unidirectional design in a bidirectional context.

6.3.1 Bidirectional Golden Gate System

Due to the need for a golden gate fragment that contained two CMV cores, one on the sense and the other on the anti-sense a stuffer sequence had to be placed in-between them to allow the sequence to be sequenced once synthesized and confirm it is correct. For this reason and also to reduce the workload, it was decided that the bidirectional golden gate vector, would be assembled once and then the bidirectional promoters would be ligated in between the CMV cores using XbaI and HindIII. Figure 6.9 shows a plasmid map for the bidirectional golden gate system. This system has the promoter titled BD1 in between the two CMV cores.

Unlike the previous section, sequencing of the vector was more difficult in this circumstance. All of the libraries' promoters were sequence verified using Sanger sequencing, but for one sequence (BD5), which was only 171bp in length, the secondary structure prevented sequencing as the CMV cores were too close together. This issue was also seen when the initial synthesis of the gene fragments was being performed and as such, a stuffer sequence was placed between the CMV cores to reduce the secondary structure. This worked for the gene synthesis and all of the longer promoters had no issue being sequence-checked. BD5 was carried forward without a fully verified sequence.

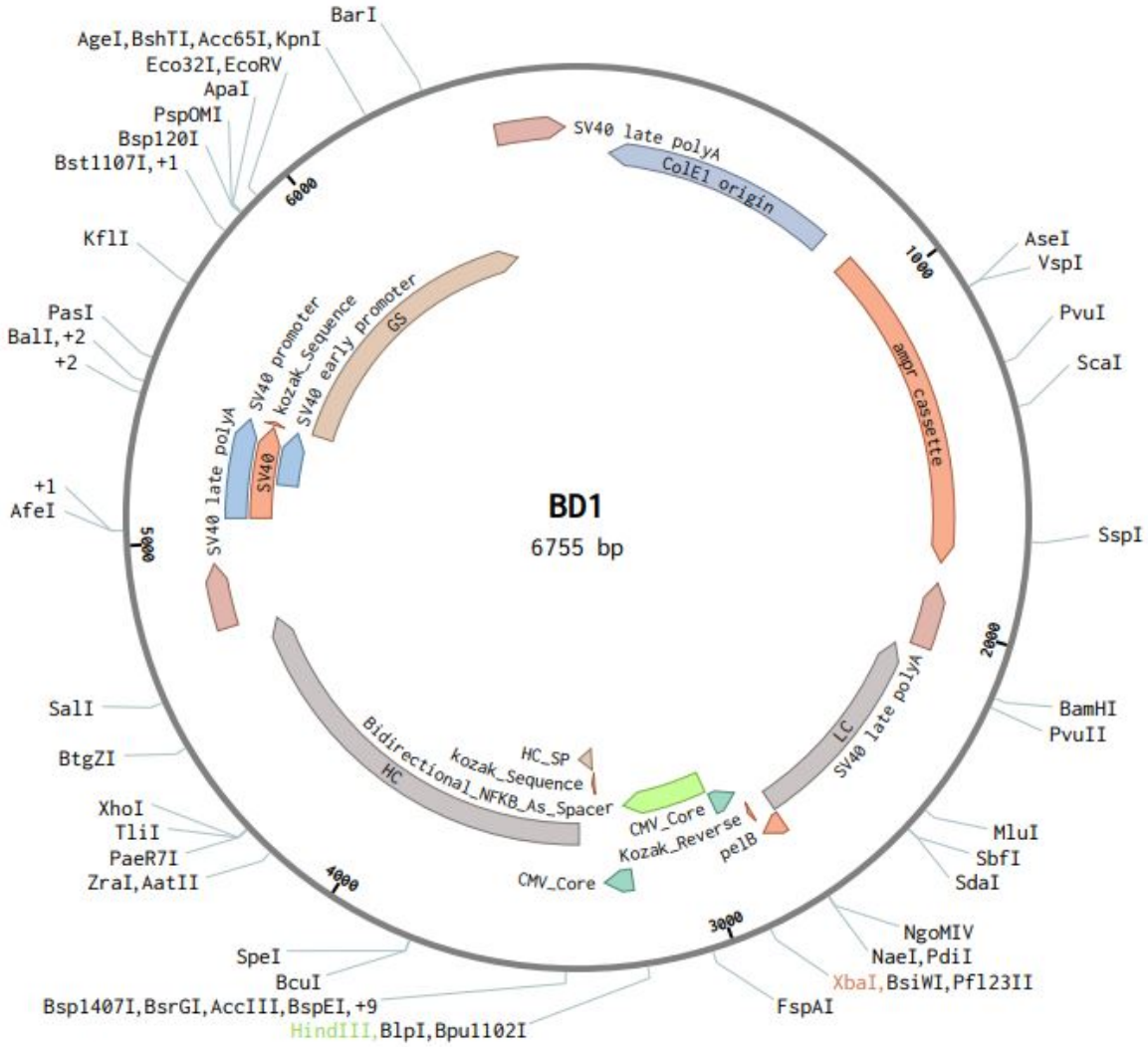


Figure 6.9: Fully constructed golden gate vector for construct BD1. Map of the BD1 construct showing the promoter between the two CMV cores. The restriction sites used for ligation are XbaI and HindIII.

Table 6.4 shows the overhangs that were used for the initial golden gate assembly for the bidirectional vector. This was then followed by restriction digest and ligation using XbaI and HindIII. The weakest part of this strategy was the transformation efficiency was extremely low. After ligation, only one or two colonies would grow, even when using high-efficiency cells.

Table 6.4: Overhangs and constructs used for bidirectional golden gate system.

Upstream	Downstream	Bases	Reverse Complement
Linear_Backbone	BLC_Gene	GATA	TATC
BLC_Gene	CMVCx2	TCGC	GCGA
CMVCx2	HC_Gene	AAGA	TCTT
HC_Gene	SV40_Linear	TAGT	ACTA
SV40_Linear	SM_Gene	TGTT	AACA
SM_Gene	Linear_Backbone	CGTG	CACG

6.3.2 Bioinformatics for Bidirectional Promoters

To find out what was abundant in CHO bidirectional promoters a dataset was taken from Trinklein et al. (2004) which aimed to find bidirectional promoters in the human genome and converted to CHO orthologues. The sequences were then extracted from the Ensembl database and analysed in the same way as was performed for discovering homotypic building blocks. Unlike the homotypic analysis, there was more weight put on literature and previous knowledge of what works in CHO cells as there was no quick screening process for the bidirectional promoters. Figure 6.10 shows the general workflow for this.

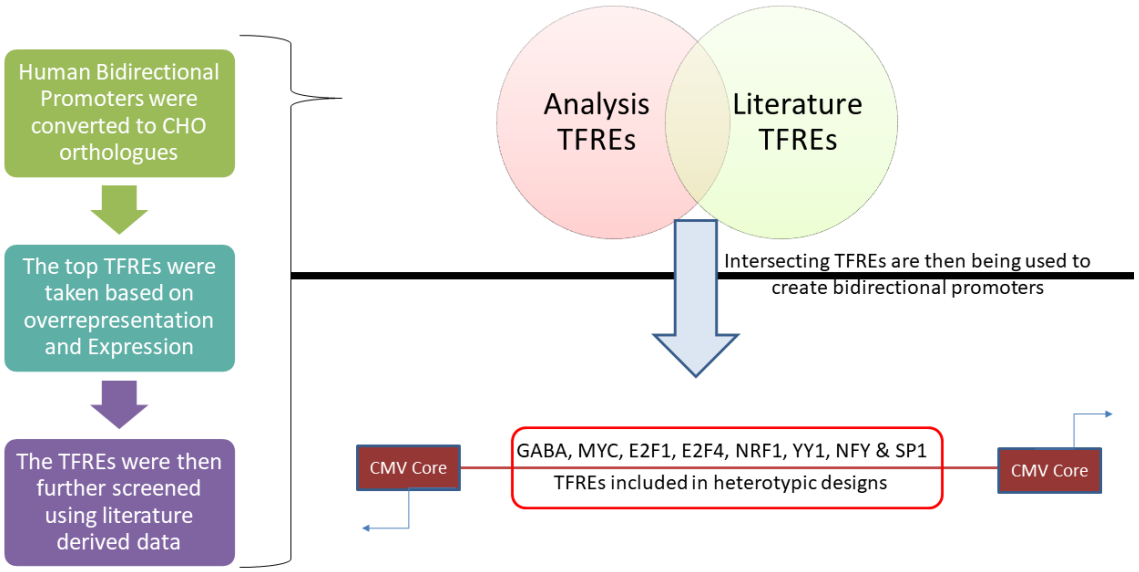


Figure 6.10: The workflow used to identify elements of importance for bidirectional promoter construction in the CHO genome. The initial data set was taken from Trinklein et al. (2004) and converted from Human gene IDs to CHO Ensembl IDs. 1200bp from the gene start sites were then taken and analysed for both over-abundance versus the general CHO background and the expression of the transcription factors that bind to these TFREs.

Figure 6.11 shows this analysis's general output achieved from the z-score*TPM metric. As discussed in the section on bidirectional promoters in Section 1.5, several TFREs have already been found to be over-abundant in human bidirectional promoters. These are "GABPA, MYC, E2F1,E2F4,NRF1,NF-Y and SP1" (Yang and Elnitski, 2008). The results shown below are in agreement with these findings and not just show that they are overabundant but also expressed in the CHO genome. GABA, MYC(ATF6), E2F1, NRF1 and SP1 all appear in the top 60 TFREs from this analysis. The analysis results are shown in terms of their matrices, so there are duplicate values in the dataset and many matrices are actually from the same family. The highest valued TFREs from this analysis were elk3 and etf4, which are both members of the ETSF family and have been previously shown to be transcriptionally active in CHO cells. The code for this analysis is shown in Appendix B.6.

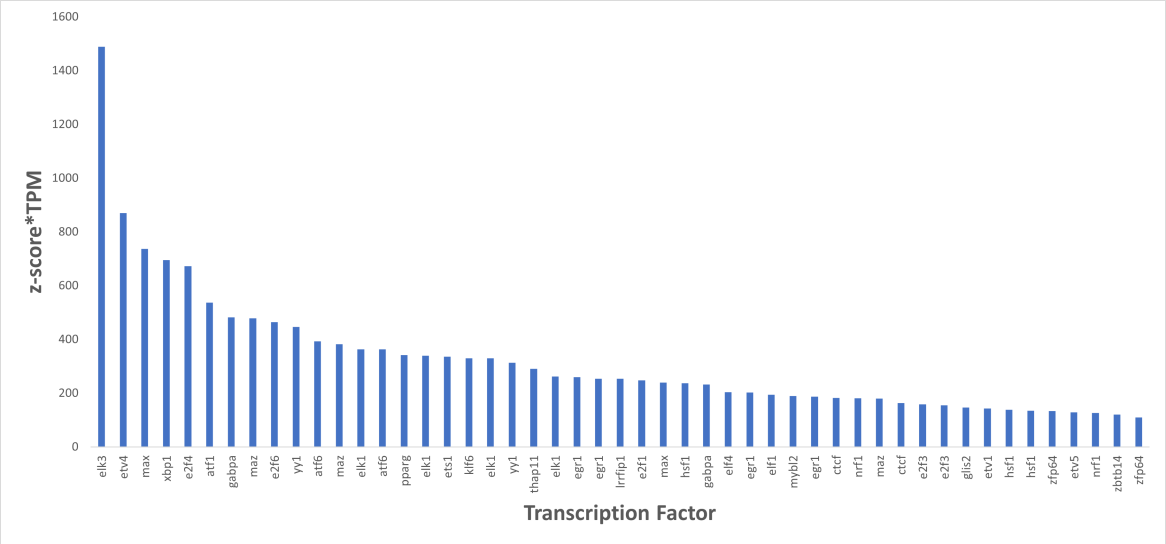


Figure 6.11: Results of the z-score * TPM metric for the analysis of bidirectional promoters in the CHO cell genome. The x-axis shows the transcription factors gene name and the y-axis shows the metric of z-score*tpm. Many of the TFREs mentioned in the literature appeared in this analysis, further increasing the confidence this analysis provided.

The results presented in this section provided insight into TFRE sites which may provide bidirectionality in the bidirectional promoters. This could be important as if these promoters are expressing both the heavy and light chain it will be important to be able to create libraries with a range of activities in the sense and antisense direction. If transcription factors indeed are directional and bidirectional it will change how they will need to be designed as the unidirectional design rules have never considered the anti-sense direction.

6.3.3 Library 1 Design and Results

The bidirectional library investigated if the concepts used for unidirectional design work for bidirectional promoters and if the informed design from the bioinformatic analysis can create working bidirectional promoters. For this reason, the library was split into two. The first section focuses on building promoters the same way unidirectional are built and the second section will try and incorporate some TFREs found in the bioinformatic analysis. All bidirectional promoters are listed in Appendix A.7.

Table 6.5 shows the sequence construction for each sequence in this library. B1 and B2 were just normal unidirectional promoters that had NFkB put in as a reverse complement and complement in-between each TFRE to see if putting TFRE sites on the antisense strand affected the transcriptional activity of the promoter.

B3 and B4 are just old promoters tested in the bidirectional sense. B3 is the Balanced Promoter and B4 is the Super Promoter without spacers. The idea behind this was to see how promoters designed with unidirectional transcription performed in the bidirectional sense.

B5 was the first sequence with informed bioinformatic design behind it. This sequence

was testing what would happen if you put the minimum amount of TFREs in. B6 was then this same promoter but with increased NFkB (to increase unidirectional transcription) and GABPA (to increase bidirectional transcription). Increased amounts of NFE2l2 were also included as its also known as NRF2. NRF2 and 1 are from the same family, so they may be expected to serve similar functions.

B7 was the same as B6 but just with increased amounts of SP1. B8 was then the same as BD7 but with increased amounts of CTCF to see if this affected the transcriptional activity. Finally, B9 was an attempt to try and put as much of everything in that was under 600bp. The idea behind this was to try and max out transcription and bidirectionality. All sequences used in this experiment are listed in Appendix A.7.

Table 6.5: Table showing the TFRE composition of each of the bidirectional constructs. The first section is shown in the first 4 rows and these are the bidirectional promoters generated using unidirectional methodologies. The New Approach Design is the bioinformatically informed section.

Name	Length	NFkB	GABP beta	DMP1	ARE	AhR/ARNT	HRE	AARE	NRF1	GC BOX	YY1	NFE2l2			
Bidirectional Using NFkB as Spacer Reverse Complement	302	0	3	1	1	1	1	1	1	1	1	3			
Bidirectional Using NFkB as Spacer Complement	302	0	3	1	1	1	1	1	1	1	1	3			
Bidirectional Balanced Promoter	488	6	6	4	4	4	4	2	5	0	0	0			
Bidirectional Super Prom W/O spacer	488	6	6	4	4	4	4	2	5	4	0	0			
New Approach Design	**Length**	**NFkB**	**GABP beta**	**DMP1**	**ARE**	**AhR/ARNT**	**HRE**	**AARE**	**NRF1**	**GC BOX**	**YY1**	**NFE2l2**	**V$CTCF.04**	**V$EGR1.01**	**SP1**
One Block Of Each TFRE	171	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NFKB NFE2l2 GABPA Skewed Promoter	276	4	4	1	1	1	1	1	1	1	1	4	1	1	1
Increased SP1 Promoter	306	4	4	1	1	1	1	1	1	1	1	4	1	1	3
Increased SP1 and CTCF Promoter Bi	334	4	4	1	1	1	1	1	1	1	1	4	3	1	3
Complicated Promoter With Elevated Everything	548	4	4	3	3	3	3	3	3	3	3	4	3	3	3

Table 6.6 shows the promoter names and the construct names they were given for easy graphing and the discussion above. One construct to note is B10. This is the full CMV enhancer region put between the two CMV cores to see how it functions compared to its unidirectional counterpart.

Table 6.6: Bidirectional construct names and the promoters they contain.

Construct Name	Promoter Sequence
B1	Bidirectional Using NFkB as Spacer Reverse Complement
B2	Bidirectional Using NFkB as Spacer Complement
B3	Bidirectional Balanced Promoter
B4	Bidirectional Super Prom W/O Spacer
B5	One Block OF Each TFRE
B6	NFKB NFE2l2 GABPA Skewed Promoter
B7	Increased SP1 Promoter
B8	Increased SP1 and CTCF Promoter
B9	Complicated Promoter With Elevated Everything
B10	FCMV in Bidirectional Format

The results from the first synthetic bidirectional promoter library are shown in Figure 6.13. The most notable finding from this experiment was B1 which achieved a fold change of 1.6 fold higher than the CMV control ($p < 0.05$). This is interesting as it shows that using the NFkB TFRE as the reverse complement causes expression. This is likely because the NFkB TFRE initiates transcription of the light chain in this context. The use of the complement sequence of NFkB shows the opposite trend and this sequence's expression is only 0.36 fold of the CMV control.

B10 has no statistically significant difference from the unidirectional FCMV_HC_LC control. This is interesting as it shows using one CMV promoter in a bidirectional context is the same as having a separate CMV promoter for the heavy chain and light chain. Although, this may be explained by the CMV promoter having a natural bidirectional nature (Romero-Santacreu et al., 2010). Alone, this is exciting as it indicates the need for the current system of using two CMV promoters could be converted to a bidirectional system using CMV and show similar expression, in a transient system. Although, this has been shown previously (Andersen et al., 2011).

The bioinformatically informed promoters showed some promise. B6, the first bioinformatically designed sequence that could be verified showed no statistical difference compared to CMV ($p = 0.35$). This shows that implementing TFREs used in the unidirectional design and those found in the bioinformatic analysis can lead to successful bidirectional promoters being created. Increasing the amount of blocks of SP1 and CTCF reduced the amount of expression, with B7 having a fold change of 0.6 and B8 having a fold change of 0.4 compared to the control. The increase in expression from B5 to B6 may suggest that NFkB, GABPA or NFE2l2 may increase bidirectional transcription. However, this needs to be confirmed as it may be that the sequence of

B5 is incorrect, or perhaps the secondary structure of the construct is too strong to transcribe.

B3 and B4 show that the application of unidirectional promoters designed in library 2 has very poor bidirectional performance with a fold change of 0.55 and 0.21 achieved. This unfortunately indicates that the application of unidirectional synthetic promoters, as they are currently constructed can not be directly applied to creating bidirectional promoters. Likely the anti-sense strand must be considered in future designs.

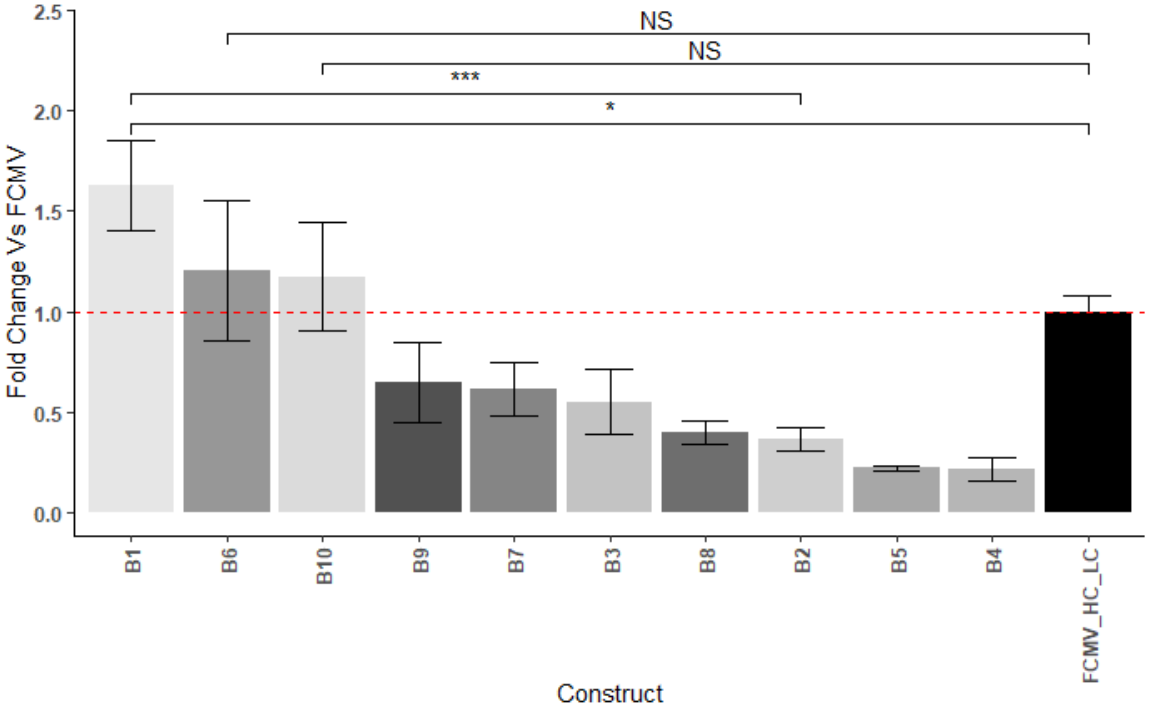


Figure 6.12: The successful creation of bidirectional promoters for the production of IgG in bidirectional antibody library 1 normalised to FCMV. This graph depicts the results for the first library of bidirectional promoters. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest (n = 3 ± sd). The results show B1 achieved a higher fold change than the FCMV control. One bioinformatically designed sequence achieved equal activity to the control and CMV itself in a bidirectional context also achieved a similar activity level. The statistics presented on the graph show * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$ and NS = not significant.

To check if the bidirectional promoters failed due to an inefficient ratio of heavy and light chain being used they were also tested using ddPCR. Overall, this result is completely inconclusive. The results of B1 and B2 show nearly the same HC:LC ratio and even when there is an outlier like B10, which had a HC:LC ratio of 14:1, it still has a similar final titre to the control.

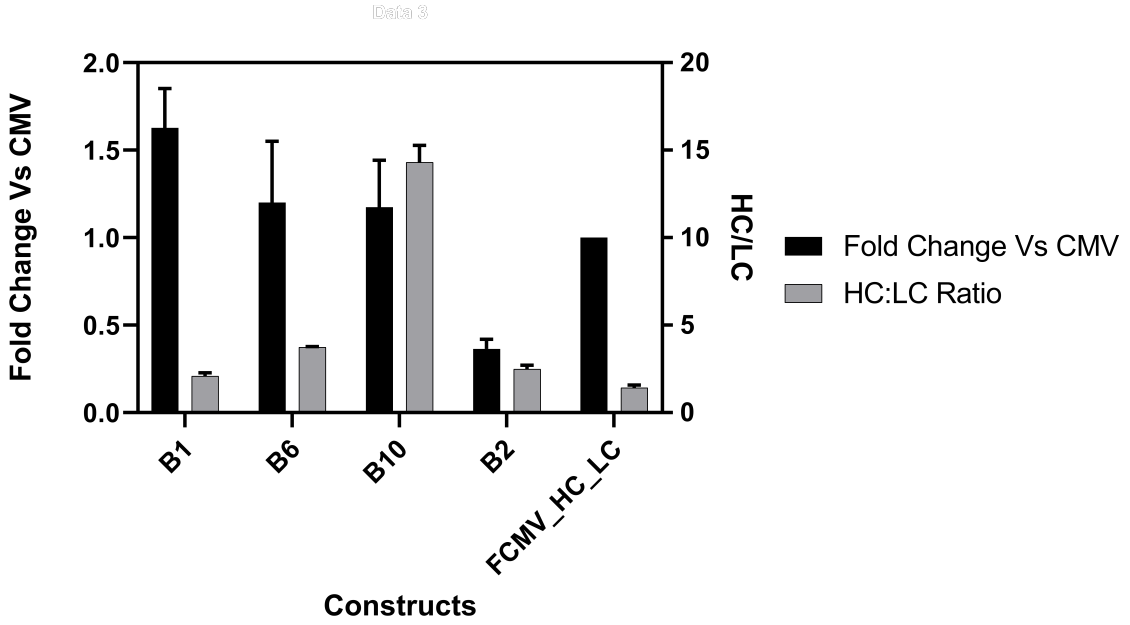


Figure 6.13: An investigation of the HC/LC ratio between a selection of bidirectional promoter constructs. The left y-axis shows the fold change versus FCMV, the right y-axis shows the HC:LC ratio and the x-axis shows the construct names. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest (n = 3 ± sd). The results show a similar trend to what was seen in the unidirectional results. The HC:LC ratio doesn't show any trend and even a 14:1 outlier ratio still produced a similar final fold change compared to the control.

This section presents the first attempt to produce a synthetic bidirectional promoter that could be used for recombinant antibody production. The investigations performed in this library showed it is possible, although the exact mechanisms that cause the bidirectionality are unclear. Likely, it is a mix of bidirectional TFREs and transcription factors on the reverse strand of DNA, but more work is needed to confirm this. Overall, the aim of this study has been achieved and unexpectedly, a bidirectional promoter has been created, which had a higher final titre than the control. Lastly, it was found that FCMV is also very good in the bidirectional context.

6.3.4 Library 2 Design and Results

This section discusses a proof of concept. The idea was to take the best unidirectional construct, which was U8 in this case and see if it could be converted into a bidirectional promoter. Figure 6.15 shows a diagram of this workflow. The idea was to take the 3' weighted promoter and use this on the sense strand of DNA. Then the 300bp promoter was taken and turned into the reverse complement. This allowed the design of only the sense strand of DNA while placing the binding sites in the correct orientation for the 300bp promoter on the anti-sense strand, which simplified the design.

Wherever the 3' weighted promoter initially had spacers, a TFRE from the new reverse complement 300bp promoter was placed. As the 300bp promoter was shorter than the 3' weighted promoter when the TFREs ran out, the original aa spacer was once again used.

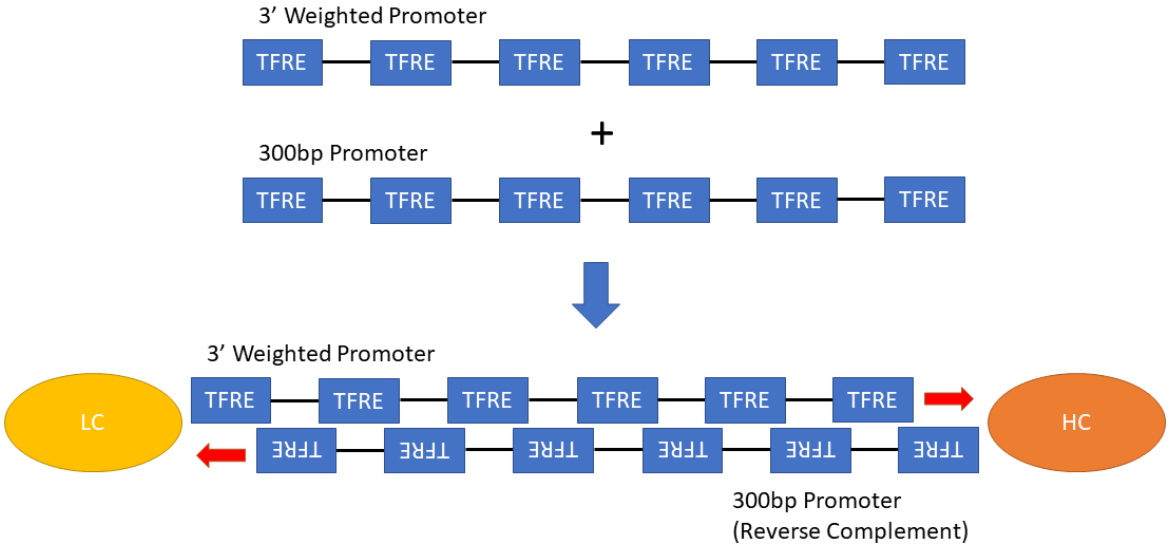


Figure 6.14: Workflow explaining how the new bidirectional promoter for library 2 was created. The idea was to take the U8 construct and reform it into a bidirectional promoter. The 3' weighted promoter was on the sense strand and the 300bp promoter was on the antisense strand. The idea was to express the heavy chain (HC) using the 3' weighted promoter and the light chain (LC) promoter using the 300bp promoter. The promoter length was ultimately 803bp in length. The longest sequence synthesised to control transcription.

Figure 6.15 shows the results for B11. As can be seen, the B11 promoter achieved a high fold change of 1.8 fold versus the control. Unfortunately, due to an outlier in replicate 3 (1.08 fold change versus CMV), there was no statistical significance between the construct and the control. The outlier was unable to be removed according to the Grubbs test. The result was unable to be repeated due to time pressures with the project. It will be carried forward to stable testing at Merck to see how it functions in a stable system. Although not statistically significant, the higher average versus B1 in the previous section (1.8 versus 1.6) suggests that this combining of promoters works to create a bidirectional promoter.

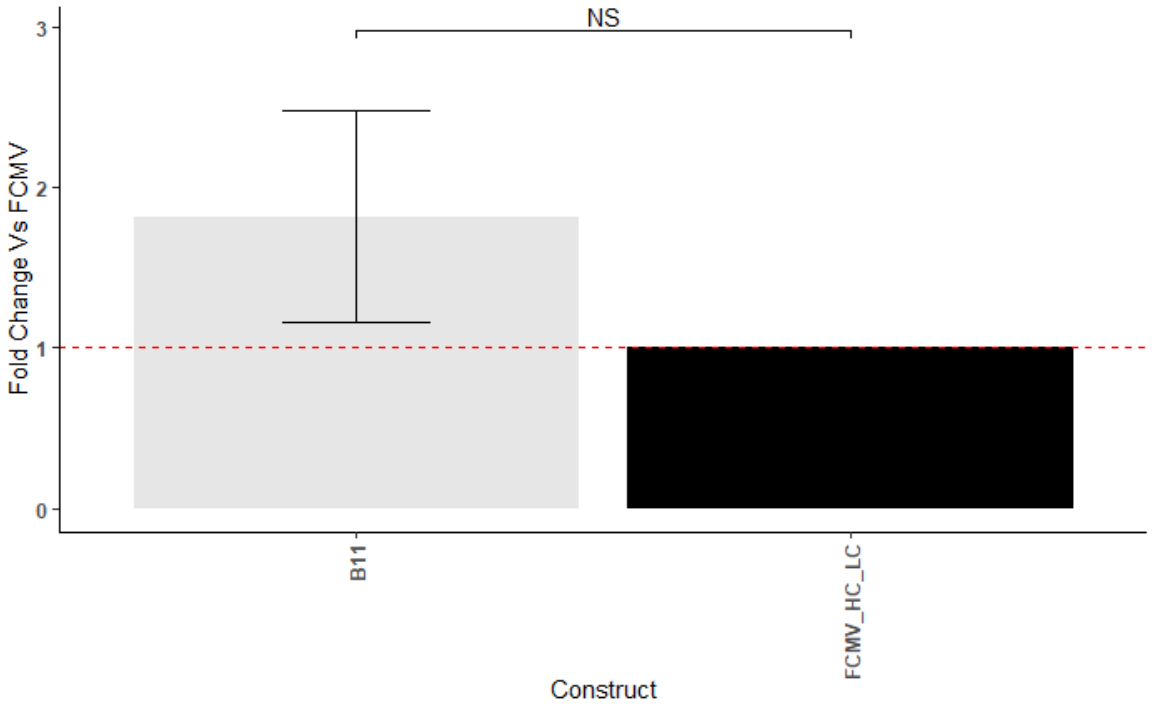


Figure 6.15: The B11 construct has improved the IgG production of bidirectional antibody library 2 normalised to FCMV. Transfections were performed using electroporation and cells were seeded in 24 shallow well plates for 4 days prior to supernatant harvest (n = 3 ± sd). The results show promise for the B11 construct with an average fold change of 1.8 fold, but unfortunately, due to variability, this is not statistically significant. The statistics presented on the graph indicate the following p-values based on a unpaired two-tailed t-test: * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$ and NS = not significant.

6.3.5 Observations from the Bidirectional Library

In conclusion to this section, the major aim has been achieved of creating a fully designed synthetic promoter. To the authors' knowledge, this is the first time this has been done from scratch without utilising promoters taken from nature.

The first section used the bioinformatic pipeline created in Chapter 4 to identify potential TFREs present in CHO bidirectional promoter regions which could be used for synthetic promoter design and agreed with literature. The interesting note in this is that many of the transcription factors identified were also previously tested and also used in the previous work of this thesis. However, when previously designed unidirectional promoters were tested in a bidirectional context, they lacked sufficient titre to be useful. One promoter, B6 which was created using bioinformatically informed data matched the unidirectional CMV construct and this is likely due to the TFREs NFkB, GABPA and NFE2l2 composition. Although to prove this, more work is needed.

The other aim of this work was to try and create a bidirectional promoter with a higher final titre than the unidirectional control; unexpectedly, this was achieved twice. The first was in library 1, the B1 construct which put the binding sequence for NFkB as the reverse complement to see if it would affect the titre. From that, it was suspected that design with the sense and anti-sense strands of DNA in mind for bidirectional

promoters was important.

A second library or single sequence was created using the same concepts as B1. The 3' weighted promoter was on the sense and the 300bp promoter was on the anti sense. This appeared to work even better than B1 in terms of fold change with an average value of 1.8 fold versus the unidirectional CMV control. It indicates that for future works that intend to create bidirectional synthetic promoters, the reverse complement should be used for anti-sense transcription.

Overall, the objectives have been achieved, but considerably more work is needed to understand and optimize the system. Bidirectional promoters capped out at 1.8 fold versus U8 which achieved a fold change of 2.7 fold. For now, it appears bidirectional promoters cannot replace unidirectional for raw titre output in a transient expression context. Although, screening in stable systems will be needed to see if the pattern is the same.

6.4 Merck Internal Testing

Unfortunately, due to time restraints and in-house issues at Merck's industrial labs, the stable data is still incomplete. This section is written to describe what is currently available from the internal testing but the final batch runs are not completed. All available data will be shown and discussed here and relevant observations will be mentioned.

The Merck process initially transfects cells and puts them under selection pressure using MSX. From here, they go through cold capture to see what % of the transfection pool for each construct is producing antibody. Figure 6.16 shows the results for the cold capture that Merck provided. Cold capture is usually used to enrich CHO cell populations for higher productivity. More information can be found in Pichler et al. (2009). Please note on all subsequent graphs the code name for bidirectional promoters has changed from, for example, B1 to BD1. BD stands for bidirectional.

The positive control in this experiment is an already producing clonal cell line used in the Merck process. The negative is a non-producing host cell. The other negative control is where no Protein A is present in the cold capture. For this experiment, two controls were provided BD10, the full CMV in a bidirectional context and U13, the unidirectional CMV construct used in previous experiments called FCMV_HC_LC.

The cold capture only gives a vague idea of how well a population of cells is producing and not the actual titre or productivity that will be achieved. For instance, a transfection population with a higher % of GFP-positive cells just indicates it will be more likely to have high-producing clones as the construct produces enough antibody in a large number of clones to be bound by the cold capture process.

From the experiment, some interesting observations can be made. The first is that the FCMV in the bidirectional context is worse than the unidirectional FCMV control in this instance. Fewer cells are producing detectable antibody. Another, in terms of the bidirectional constructs, is that BD1 and BD11 have a large amount of difference in their % of GFP positive cells which may indicate what the transient data suggested. That BD11 is the best bidirectional promoter.

Looking at the unidirectional constructs, U8 again appears to be one of the best overall synthetic constructs that went through cold capture, with approximately 80% of the cells showing expression. Unlike the transient screen, U9 is also one of the best and U12 is doing very poorly in the % of GFP-positive cells. Finally, U14, the best construct from library 2, shows very little activity, nearly as low as the control. This may indicate experimental error as it's next to zero.

If the % of GFP positive cells correlates with the final cell lines tested, it may indicate that the constructs when screened in transient and stable systems act differently in some cases.

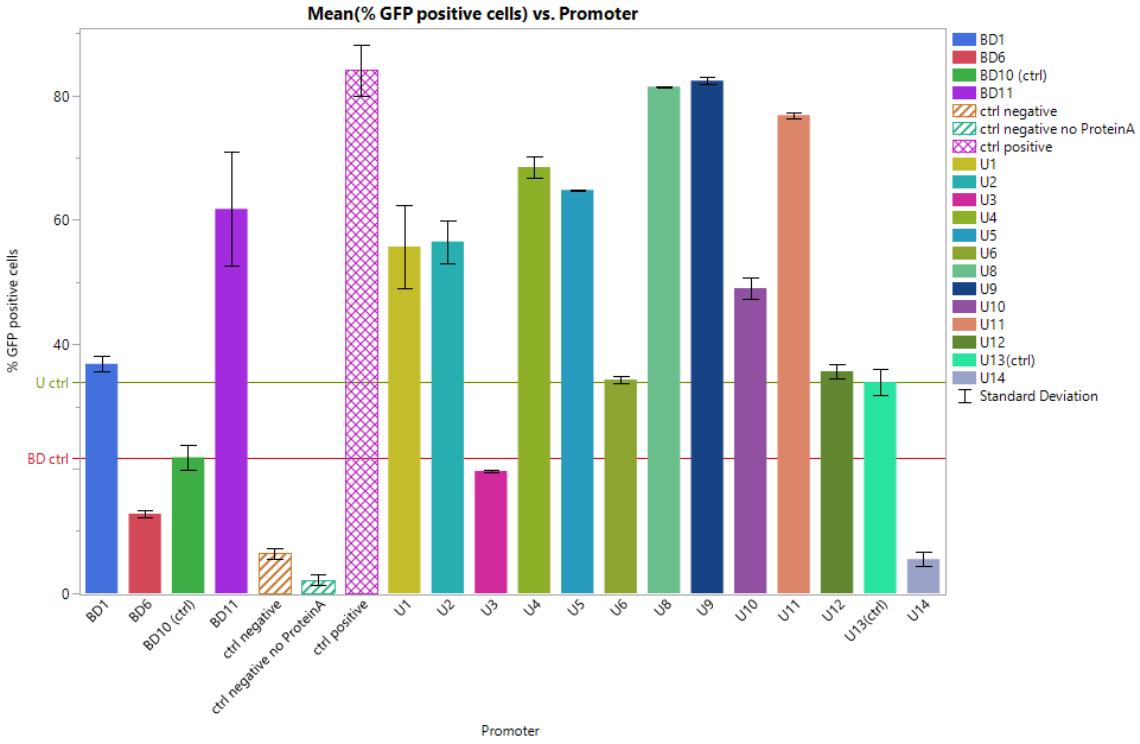


Figure 6.16: Bidirectional and unidirectional GFP enrichment post cold capture. The results for cold capture of the different transfection pools separated by promoter, as provided by Merck are shown. What can be seen is some of the constructs that initially were better than the control are now worse in terms of % of GFP positive cells. Although this doesn't correlate with low productivity later on, it suggests that some constructs' general population, for example, U3 are too low expressing for a high proportion of the population to pass the screening process.

Next, the cells were screened in an instrument called a Beacon (Berkeley Lights, California). This instrument takes pools of cells and allows them to be sorted into single pens using nanofluidic chips (Le et al., 2020). This then measures cell counts, titre and productivity, allowing high throughput screening of potential clones as described in Le et al. (2020).

The data is presented in a boxplot to show the distribution of the data. Figure 6.17 shows the final cell counts of each construct. The control FCMV_HC_LC (U13) is shown at the bottom of the graph with the lowest distribution of final cell counts. This indicates that when seeded into the Beacon (Berkeley Lights, California), the CMV driven control has prolonged growth compared to the synthetic promoters. This is also why the sample size (as shown in black text on the graph) is so small, as most of the FCMV_HC_LC (U13) constructs either didn't grow or died upon Beacon (Berkeley Lights, California) seeding. This likely indicates that overall the synthetic constructs are providing a much lower cellular burden than the CMV control, which may indicate low selection stringency during GS selection.

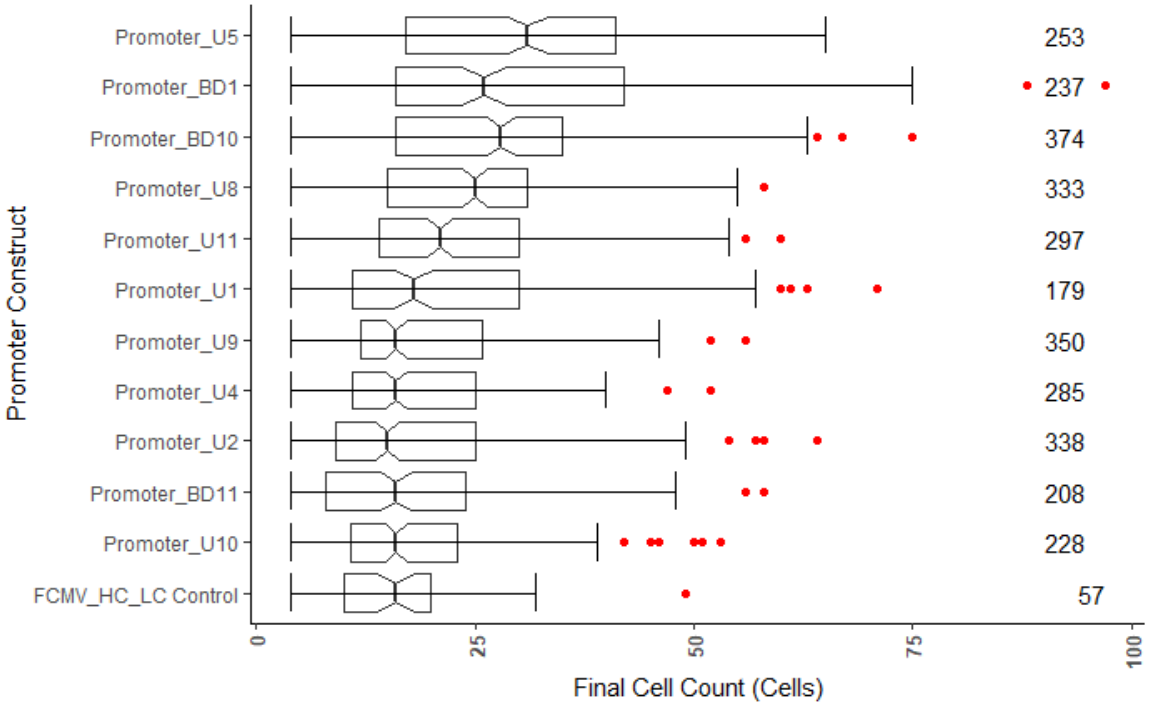


Figure 6.17: The final cell counts reported by the Beacon after 4 days of culture. The data was shown in a boxplot to give an idea of the distribution. The sample size of each construct is shown in black text next to each construct. The red dots indicate outliers. The data shows that the full CMV control is the slowest growing and may suggest that the control has higher selection stringency. Red dots show outliers on the boxplot.

Figure 6.18 depicts the Final AuScore for the constructs during the Beacon (Berkeley Lights, California) experiment. This unit of measurement is Beacon (Berkeley Lights, California) specific and as such doesn't have a unit. It is a measure of the productivity of each pen within the Beacon (Berkeley Lights, California). This does not consider the number of cells present in each pen. What can be seen from this data is that there is a wide distribution of overall productivity. The control is the lowest in this graph, generally being outperformed by every construct. This is likely due to the other cells growing much better than the U13 control and producing more antibodies due to the increased volume of cells.

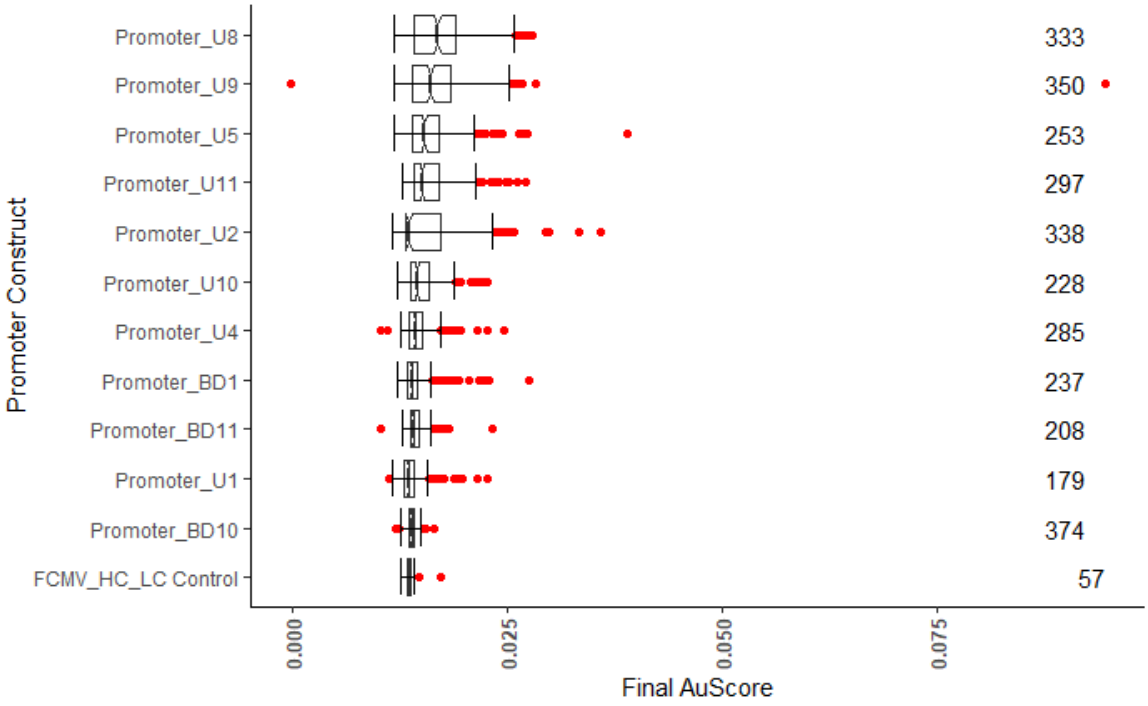


Figure 6.18: The final AuScore reported by the Beacon (Berkeley Lights, California) after 4 days of culture. The AuScore measures productivity. Each construct's name is shown on the y-axis, the Final AuScore on the x-axis and the sample number is shown in black text. In general, it was shown in terms of productivity that most synthetic promoters have higher productivity than the FCMV_HC_LC control (U13). This is likely due to the increased final cell counts of the constructs.

Finally, to account for the differences in final cell count, the beacon uses the final rQp measurement, which is the Final AuScore divided by the final cell count, as can be seen in Figure 6.19. Comparing it to the previous graph, it can be noted that the general trend has completely changed. Promoter_U8 had high productivity but also had very high final cell counts and as such, has been ranked much lower in terms of specific productivity (rQp). When doing clonal selection, the amount of outliers is important. For that reason, the % of outliers for each sample is graphed in red beside the sample numbers.

Overall, looking at the distribution of the data, none are very different from FCMV_HC_LC (U13). Based on the distributions, if one were to pick a construct from this it would likely be BD11, a bidirectional version of U8. The closer the median is to the lower quartile of the boxplot, the more right skewed the data is, which is beneficial for clonal selection. From this plot and looking at individual clones' characteristics, Merck chose the constructs BD1, BD10, BD11, U11, U13, U2 and U9 to carry onto batch testing. From the Beacon (Berkeley Lights, California) data, it appears none are significantly different from the control in terms of generally increasing productivity. What can be seen generally throughout the data is that the synthetic promoters have increased dispersion. This may indicate, that although the medians appear similar, the synthetic constructs' upper boundary is larger, which could indicate they create a higher percentage of high-producing clones, which is very important for clonal selection. Perhaps, if the sample sizes for FCMV_HC_LC (U13) were more equal to the rest of

the constructs, it would show a greater difference.

The big difference between Final AuScore and rQp is likely due to the cells with synthetic constructs growing more and producing more protein overall, but when the cell count is considered, the single cells with synthetic promoters are producing around the equivalent amount per cell. This is important as industry does not want cells that produce too much biomass as it increases resource costs in terms of media consumption etc. The diminished rQp of the synthetic clonal cells may indicate that increased stringency in GS selection is required with synthetic promoters. As many of the synthetic clones are above the maximum rQp of the control batch. Testing will be required to see if the synthetic promoters do outperform the control on a larger scale.

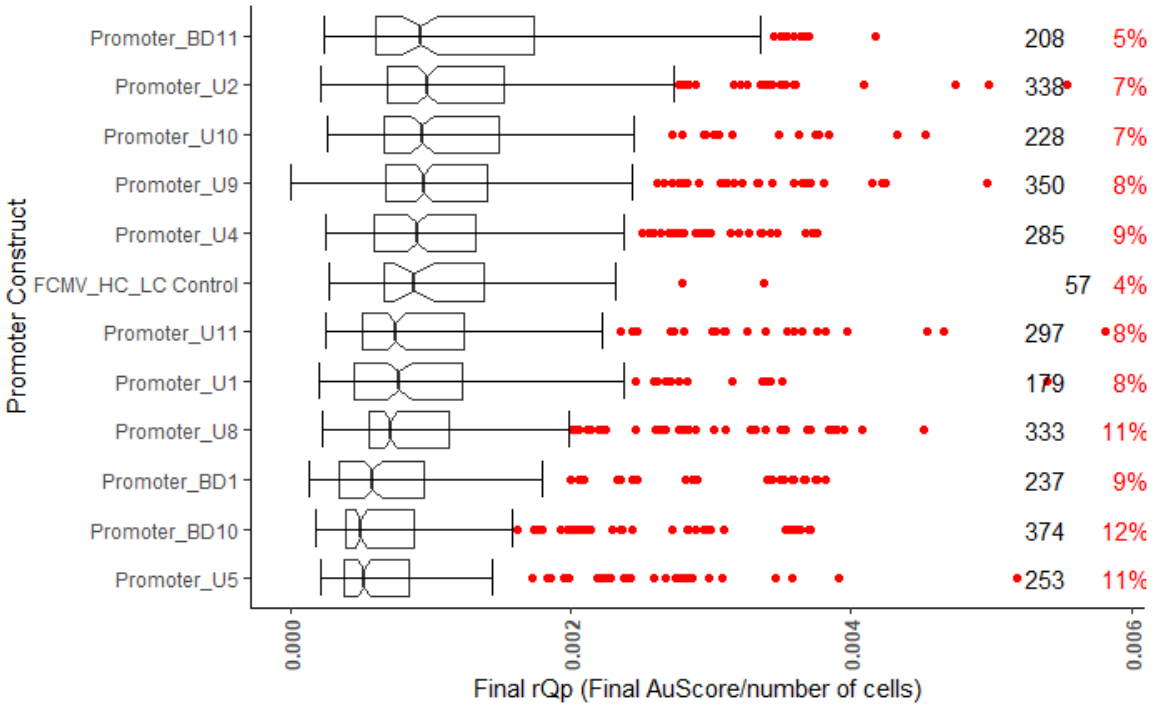


Figure 6.19: The final rQp reported by the Beacon after 4 days of culture. This is a measure of the Final AuScore divided by the Final Cell Count. The black text shows sample numbers for each construct and the red text shows the percentage of outliers. Overall the data is inconclusive. Some synthetic constructs such as U4, U9, U10, U2 and BD11 appear to have skewed the data into generally higher rQps, but batch testing will be required to confirm this. Outlier points on the boxplot are marked in red.

6.5 Final Thoughts

The overall conclusions of the unidirectional and bidirectional works are discussed in their relevant sections but taking the results into account is important. The application of synthetic promoters to produce antibodies is in its infancy and as shown by the unidirectional section is very hard to predict the outcome. The bidirectional work is even less advanced with the application and creation of bidirectional promoters being performed simultaneously. In both cases, more work is needed to understand how to apply synthetic biology technology to its fullest extent.

In terms of the real-world Merck data, it is evident that the transient screens do not directly correlate to how the construct functions in a stable environment. In general, it appears that the trend of what promoter or combination of promoters is best can completely change. The bidirectional construct BD11 is showing the best distribution of data against all of the unidirectional constructs. U2, which was also not a great construct in the unidirectional transient screen, is now the second best performing construct in the stable clonal process. This indicates that in future works, it will be essential to implement a high throughput stable system for better prediction of real-world uses. The batch data that Merck will provide should give an even better idea of just how differently the top clones in these pools perform.

Other molecules must be considered for future works as that may change the pairing kinetics greatly. Other considerations like large scale library screening methodologies should be considered as discussed previously, small libraries will not yield optimal results. The increased growth rate of CHO cells containing synthetic promoters must also be considered and in any future works, GS selection stringency may need to be altered to ensure a higher proportion of high-producers is achieved.

This application chapter has shown that although the understanding of how synthetic promoters function is limited and how they interact is also limited, they can be applied successfully. The most crucial factor is that they are better than the CMV system currently used. For most synthetic promoter constructs this was easily achieved in transient work and as such is easily applicable to industry. In terms of optimisation, it gets more tricky. If optimised synthetic systems are ever to be created, one of two possible outcomes will have to be fulfilled. The first is a high throughput system that can easily find the best construct combination possible and the second is truly understanding the biology to model the biological outcomes before testing. A high throughput system and advanced analytical methods are needed to achieve the second outcome. The Merck stable data has also shown this will have to be done in stable cultures, which is even more resource intensive.

Chapter 7

Conclusions and Future Works

Overview

- This chapter summarises the main findings from the works in this thesis, along with personal views on what the conclusions may mean and how the studies could be improved.
- This is contained within each chapter, but a more defined summary will be shared in this section.
- The future works will discuss where this work could go in the future if continued.

7.1 Conclusions

The overall results presented throughout this work have led to the creation of synthetic promoters, which Merck is testing in their in-house system. This was the ultimate goal of this project. The work has also given some key ideas in synthetic promoter design considerations and how this can be applied to producing an antibody in a transient system. A more detailed conclusion for each chapter will be expanded upon in this section.

7.1.1 Chapter 3 Bioinformatic Analysis of Chinese Hamster Ovary Cells

Although not originally planned, COVID-19 led to a large-scale differential expression screen which was carried out to understand the differences between a producer and a non-producer. A more hypothesis-driven approach was used to test certain theories and see if they hold true in this work. The first step was a large-scale KEGG pathway analysis to see what was changing in these pathways on different days of culture. From this, it was seen that GS selection while producing a recombinant protein appears to create cells with a dysregulated response to protein misfolding and protein degradation.

The analysis then increased in specificity. Instead of using KEGG pathways, all the genes associated with certain pathways were taken and looked at to see if there was overall up or down-regulation on days 2,5 and 10 between a producer and non-producing CHO cell. Overall it appeared the ERAD was downregulated, the UPR upregulated but interestingly, the cellular response to misfolded proteins was down-regulated. Other pathways looked at were oxidative stress, mitochondrial metabolism and fatty acid metabolism, but these results were unclear.

Overall, this chapter presents the differences between a producer and a non-producer and provided a table of possible cellular targets that may be interesting to test in the future. This work could be further improved by employing more cell lines of varying productivities and including a host CHO cell to provide the transcriptomic background they started on before GS selection.

7.1.2 Chapter 4 Finding New Synthetic Promoter Building Blocks

The work in this chapter set out to try and find new building blocks for creating synthetic promoters in CHO cells and expand the design space for heterotypic promoter designs.

Initially, a bioinformatic pipeline that used the over-abundance of TFREs was designed to make this work novel compared to previous works. The benefit was that it was nearly fully automated and allowed no user bias in deciding what TFREs should be screened. This library ultimately had lacklustre results due to decisions such as not allowing the use of TFRE families that had been tested previously and using over-abundance

measurements alone. One new potential TFRE came from this work which was NRF1, but overall the result was disappointing.

To try and understand if sequence variants should be allowed in the screening process, an NFkB variant library was created and it allowed the conclusion that small changes in sequences can have substantial effects on SEAP production, which was important for library 2.

Contrary to what was seen in library 1, the revisions made in library 2 of a new measurement metric (z-score * TPM) and using small promoter lengths appeared to affect the results positively. A new TFRE was found, NFE2l2, which had rivalling activity to NFkB. This was exciting and showed that the new pipeline, which was even more automated than pipeline 1 was a success. It suffered from drawbacks such as information loss, which can be seen by overlooking the NFkB transcription factor.

Ultimately, this work has succeeded in its overall aim of finding new TFRE for designing synthetic building blocks in CHO cells. It was found sequence variants are useful, and unexpectedly, new TFREs were found.

7.1.3 Chapter 5 Unidirectional Synthetic Promoters

Three libraries of heterotypic unidirectional promoters were created, each one advancing and building on the knowledge from the last. Initially, the library was comprised of TFRE sequences taken from literature to ensure that even if there was no advancements Merck would at least have a library to implement in their systems.

Library 2 incorporated the new TFRE NRF1 and then library 3 furthered this by increasing the number of new TFREs tested. This work found that the inclusion of new TFREs, compared to existing ones, showed a slight improvement in SEAP production in transient culture. It was also suggested that 3' weighting of the promoters might be beneficial and the idea of funnelling transcription using 5' repressors does not work, at least in this work. Excitingly, the inclusion of NFE2l2 into sequences that had previously been tested increased the transcriptional activity and as such, it could be a way to improve libraries of previously tested promoters.

7.1.4 Chapter 6 Applying Synthetic Unidirectional and Bidirectional for The Production of Recombinant Antibodies.

This chapter contained the studies for applying synthetic promoters to the production of a recombinant antibody with extra design considerations for the heavy and light chain and expression ratios.

The first application would mimic the currently used system where two unidirectional promoters would be used. To understand what was currently happening in the cells that had undergone RNA-seq, the heavy and light chain of Merck's antibody was aligned and the TPM was measured. Clone 9 showed slightly excess heavy chain on day 5 and

day 10 of culture. Due to the inconclusiveness, an attempt to test both excess heavy chain and excess light chain was performed. For testing, promoter pairings mimicking the CMV system were used, along with combinations of promoters.

The two libraries of testing for unidirectional promoters showed that using a combination of promoters appears to achieve a higher titre than using the same promoter on both heavy and light chain. Although this still performed better than CMV in nearly every case.

The more novel aspect of this work was testing bidirectional synthetic promoters. The first library was very much a shot in the dark with a bioinformatically informed section and a design section. What was found was that using the reverse complement of TFREs appeared to lead to the highest titre. As a proof of concept, the best unidirectional sequence was taken and turned into a bidirectional promoter which performed very well.

Finally, the in-house stable testing by Merck. This was very important as its real-world data. It showed that the synthetic constructs generally don't have much median difference versus the CMV (U13) control, but BD11, U2 and generally most of the synthetic promoters appear to have larger dispersions than the CMV control, which is important for clonal selection. The data comparing the drastic trend change going from Final AuScore to rQP also indicated that selection stringency for cells containing synthetic promoters might need to be considered.

7.2 Future Works

This section will explore the potential future work that could be performed to further the findings of this thesis. This will include areas such as better TFRE identification, improved synthetic promoter understanding and how this can ultimately be implemented in the real world. Much of what is mentioned here is also mentioned throughout this work in the conclusions of sections.

7.2.1 TFRE Identification

Perhaps the largest issue with this work and synthetic biology is how resource intensive and labour intensive the screening processes are. Currently, both pipeline 1 and pipeline 2 are more informed guesses than actual predictions. To change this, more understanding is needed. Instead of performing this analysis with TFRE binding sites, which rely on literature derived databases, actual kinetic data will be required in the future.

With information such as CHIP-seq, TF-seq, GRO-seq and ATAC-seq, much more information about binding kinetic and the functions of these transcription factors could be uncovered. For instance, if the GAGA TFRE actually functioned in the structure of DNA upstream of high expression genes. With this information, a high throughput screening system may not even be required as the actual mechanisms which these TFREs operate under may be understood. This is a very CHO cell problem, if this work was repeated in a human cell line there would be much more information freely available to study these areas and the data sets would not need to be generated in-house. There is also room for machine learning in this area but this will be discussed later in the heterotypic promoter area.

7.2.2 Heterotypic Promoters

The design block model works very well for synthetic promoter construction. Its simplicity, unfortunately, does not allow understanding and thus the prediction of what activity a sequence will produce. For that reason, the future works for heterotypic promoters share some commonality with TFRE identification with the added caveat of TFRE interaction.

In the building block model used in this thesis, interaction is nearly wholly ignored. This is due to the complex nature of knowing what works together. There is literature available that discusses TF - TF interactions but none to say how it should be designed. To truly know how this affects synthetic promoters, a large-scale screening study would be required, which would screen thousands of sequences transfected into CHO along with GFP. They could then be sorted using FACS and the promoter sequences discovered by using an RNA-seq tag. By doing this, a machine learning algorithm such as XGBoost could be used to decipher the patterns within the promoter sequences and build a model that could more accurately predict the activity before testing.

If no bias in the identification of TFRE sites is wanted, deep learning could be used. Although typically used for image identification CNNs can also be used for this genomic type data and Google's deepmind has already paved the way to do this. However, the required expertise and money will likely lead to academia being unable to perform this. This model could even account for the location of integration into the CHO or human genome if it had enough data to learn from.

7.2.3 Future Applications of Synthetic Promoters

For the future works of this thesis for the antibody-producing library, there are some elementary key considerations. This work failed to test different clones with the synthetic promoters and with other molecules. It also failed to check product quality attributes which may vary massively depending on the promoter selection.

Any future works, even in-house at Merck should test how the promoter combinations performed in different clones and with different molecules. One promoter combination will unlikely be the best for all molecules. The results could be completely different with a different molecule and vastly outperform or underperform CMV. Also, as stated previously, selection stringency for constructs containing synthetic promoters may need to be considered to ensure the population of cells is high producing as the cellular burden of synthetic promoters seems to be less than its CMV counterpart.

An experiment that checks for quality product attributes should also be performed to see if the high-producing synthetic promoters are producing a useful product or just aggregates and misfolded junk protein.

Finally, considerations should be taken into implementing this system in industry. How to implement it into the cell selection process. One key way, as aforementioned, would be first to have site-directed integration and create every possible vector combination with barcodes. Once this has been done, a large-scale transfection can occur and the cells selected and sorted. Once the highest producer has been chosen, the barcode could be read and the genetic construct known. This system would add minimal time to the cell screening process and still achieve the GMP standards required. The only drawback would be the cost of the initial synthesis of DNA. Although using the golden gate designed in this work, once the heavy and light chain promoter vectors have been created, it should be relatively easy to create any combination of promoters, even with new molecules, which would reduce costs.

Bibliography

Karen Adelman, Megan A. Kennedy, Sergei Nechaev, Daniel A. Gilchrist, Ginger W. Muse, Yurii Chinenov, and Inez Rogatsky. Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase II stalling. *Proceedings of the National Academy of Sciences*, 106(43): 18207–18212, oct 2009. ISSN 0027-8424. doi: 10.1073/PNAS.0910177106. URL `https://www.pnas.org/content/106/43/18207`.

Jakob Albrethsen, Jaco C. Knol, and Connie R. Jimenez. Unravelling the nuclear matrix proteome, feb 2009. ISSN 18743919. URL `http://www.ncbi.nlm.nih.gov/pubmed/18957335`.

Christina S. Alves and Terrence M. Dobrowsky. Strategies and considerations for improving expression of "difficult to express" proteins in CHO cells. In *Methods in Molecular Biology*, volume 1603, pages 1–23. Humana Press Inc., 2017. doi: 10.1007/978-1-4939-6972-2-1.

Mario Amendola, Mary Anna Venneri, Alessandra Biffi, Elisa Vigna, and Luigi Naldini. Coordinate dual-gene transgenesis by lentiviral vectors carrying synthetic bidirectional promoters. *Nature Biotechnology*, 23(1):108–116, jan 2005. ISSN 10870156. doi: 10.1038/nbt1049. URL `http://www.nature.com/articles/nbt1049`.

Christina Rottbøll Andersen, Lars Søegaard Nielsen, Alexandra Baer, Anne Bondgaard Tolstrup, and Dietmar Weilguny. Efficient expression from one CMV enhancer controlling two core promoters. *Molecular Biotechnology*, 48(2):128–137, jun 2011. ISSN 10736085. doi: 10.1007/s12033-010-9353-7. URL `http://www.ncbi.nlm.nih.gov/pubmed/21113688`.

Robin Andersson, Yun Chen, Leighton Core, John T Lis, Albin Sandelin, and Torben Heick Jensen. Human Gene Promoters Are Intrinsically Bidirectional. *Molecular cell*, 60(3):346–7, nov 2015. ISSN 1097-4164. doi: 10.1016/j.molcel.2015.10.015. URL `http://www.ncbi.nlm.nih.gov/pubmed/26545074http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5485825`.

Yannick Noël Anno, Evelyne Myslinski, Richard Patryk Ngondo-Mbongo, Alain Krol, Olivier Poch, Odile Lecompte, and Philippe Carbon. Genome-wide evidence for an essential role of the human Staf/ZNF143 transcription factor in bidirectional transcription. *Nucleic Acids Research*, 39(8):3116–3127, apr 2011. ISSN 03051048. doi: 10.

1093/nar/gkq1301. URL http://www.ncbi.nlm.nih.gov/pubmed/21177654http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3082894.

Michael Antoniou, Lee Harland, Tracey Mustoe, Steven Williams, Jolyon Holdstock, Ernesto Yague, Tony Mulcahy, Mark Griffiths, Sian Edwards, Panayiotis A. Ioannou, Andrew Mountain, and Robert Crombie. Transgenes encompassing dual-promoter CpG islands from the human TBP and HNRPA2B1 loci are resistant to heterochromatin-mediated silencing. *Genomics*, 82(3):269–279, sep 2003. ISSN 08887543. doi: 10.1016/S0888-7543(03)00107-1. URL https://linkinghub.elsevier.com/retrieve/pii/S0888754303001071.

Cosmas D Arnold, Muhammad A Zabidi, Michaela Pagani, Martina Rath, Katharina Schernhuber, Tomáš Kazmar, and Alexander Stark. Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nature Biotechnology*, 35(2):136–144, feb 2017. ISSN 1087-0156. doi: 10.1038/nbt.3739. URL http://www.nature.com/articles/nbt.3739.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods 2021 18:10*, 18(10):1196–1203, oct 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01252-x. URL https://www.nature.com/articles/s41592-021-01252-x.

Ravichandra Bachu, Iñigo Bergareche, and Lawrence A. Chasin. CRISPR-Cas targeted plasmid integration into mammalian cells via non-homologous end joining. *Biotechnology and Bioengineering*, 112(10):2154–2162, oct 2015. ISSN 00063592. doi: 10.1002/bit.25629. URL http://doi.wiley.com/10.1002/bit.25629.

Dia N Bagchi and Vishwanath R Iyer. The Determinants of Directionality in Transcriptional Initiation, 2016. ISSN 13624555. URL http://www.ncbi.nlm.nih.gov/pubmed/27066865http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4874897.

Laura A Bailey, Diane Hatton, Ray Field, and Alan J Dickson. Determination of Chinese hamster ovary cell line stability and recombinant antibody expression during long-term culture. *Biotechnology and Bioengineering*, 109(8):2093–2103, aug 2012. ISSN 00063592. doi: 10.1002/bit.24485. URL http://doi.wiley.com/10.1002/bit.24485.

Sowmya Balasubramanian, Yashas Rajendra, Lucia Baldi, David L. Hacker, and Florian M. Wurm. Comparison of three transposons for the generation of highly productive recombinant CHO cell pools and cell lines. *Biotechnology and Bioengineering*, 113(6):1234–1243, jun 2016. ISSN 00063592. doi: 10.1002/bit.25888. URL http://doi.wiley.com/10.1002/bit.25888.

Sjoerd Johannes Bastiaan Holwerda and Wouter de Laat. CTCF: The protein, the binding partners, the binding sites and their chromatin loops, jun 2013. ISSN 14712970.

Trish Benton, Tim Chen, Michele McEntee, Brian Fox, David King, Robert Crombie, Thomas C. Thomas, and Christopher Bebbington. The use of UCOE vectors in combination with a preadapted serum free, suspension cell line allows for rapid production of large quantities of protein. In *Cytotechnology*, volume 38, pages 43–46, jan 2002. doi: 10.1023/A:1021141712344. URL http://www.ncbi.nlm.nih.gov/pubmed/19003085http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3449923.

Zeynep Betts and Alan J. Dickson. Ubiquitous Chromatin Opening Elements (UCOEs) effect on transgene position and expression stability in CHO cells following methotrexate (MTX) amplification. *Biotechnology Journal*, 11(4):554–564, apr 2016. ISSN 18607314. doi: 10.1002/biot.201500159. URL http://doi.wiley.com/10.1002/biot.201500159.

Zeynep Betts, Alexandra S Croxford, and Alan J Dickson. Evaluating the interaction between UCOE and DHFR-linked amplification and stability of recombinant protein expression. *Biotechnology Progress*, 31(4):1014–1025, jul 2015. ISSN 87567938. doi: 10.1002/btpr.2083. URL http://doi.wiley.com/10.1002/btpr.2083.

Madhurima Biswas and Jefferson Y. Chan. Role of Nrf1 in antioxidant response element-mediated gene expression and beyond. *Toxicology and applied pharmacology*, 244(1):16, apr 2010. ISSN 0041008X. doi: 10.1016/J.TAAP.2009.07.034. URL /pmc/articles/PMC2837788//pmc/articles/PMC2837788/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2837788/.

William J. Blake, Gábor Balázsi, Michael A. Kohanski, Farren J. Isaacs, Kevin F. Murphy, Yina Kuang, Charles R. Cantor, David R. Walt, and James J. Collins. Phenotypic Consequences of Promoter-Mediated Transcriptional Noise. *Molecular Cell*, 24(6):853–865, dec 2006. ISSN 10972765. doi: 10.1016/j.molcel.2006.11.003. URL https://www.sciencedirect.com/science/article/pii/S1097276506007441?via%3Dihub.

Jens Boch, Heidi Scholze, Sebastian Schornack, Angelika Landgraf, Simone Hahn, Sabine Kay, Thomas Lahaye, Anja Nickstadt, and Ulla Bonas. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, 326(5959):1509–1512, dec 2009. ISSN 00368075. doi: 10.1126/science.1178811.

Jürgen Bode, Thomas Schlake, Michaela Iber, Dirk Schübeler, Jost Seibler, Evgeney Snezhkov, and Lev Nikolaev. The Transgeneticist's Toolbox: Novel Methods for the Targeted Modification of Eukaryotic Genomes. *Biological Chemistry*, 381(9-10):801–813, sep 2000. ISSN 14374315. doi: 10.1515/BC.2000.103.

Adam J. Bogdanove and Daniel F. Voytas. TAL effectors: Customizable proteins for DNA targeting, sep 2011. ISSN 10959203.

Sabrina Boscolo, Francesca Mion, Marta Licciulli, Paolo Macor, Luca De Maso, Martina Brce, Michael N. Antoniou, Roberto Marzari, Claudio Santoro, and Daniele Sblattero. Simple scale-up of recombinant antibody production using an UCOE containing vector. *New Biotechnology*, 29(4):477–484, may 2012. ISSN 18716784. doi: 10.1016/j.nbt.2011.12.005.

Sandra Bosshard, Pierre Olivier Duroy, and Nicolas Mermod. A role for alternative end-joining factors in homologous recombination and genome editing in Chinese hamster ovary cells. *DNA Repair*, 82:102691, oct 2019. ISSN 15687856. doi: 10.1016/j.dnarep.2019.102691.

G. Brightwell, V. Poirier, E. Cole, S. Ivins, and K. W. Brown. Serum-dependent and cell cycle-dependent expression from a cytomegalovirus-based mammalian expression vector. *Gene*, 194(1):115–123, jul 1997. ISSN 03781119. doi: 10.1016/S0378-1119(97)00178-9.

Adam J. Brown and David C. James. Precision control of recombinant gene transcription for CHO cell synthetic biology. *Biotechnology Advances*, 34(5):492–503, 2016. ISSN 07349750. doi: 10.1016/j.biotechadv.2015.12.012. URL `http://dx.doi.org/10.1016/j.biotechadv.2015.12.012`.

Adam J. Brown, Bernie Sweeney, David O. Mainwaring, and David C. James. Synthetic promoters for CHO cell engineering. *Biotechnology and Bioengineering*, 111(8):1638–1647, aug 2014. ISSN 10970290. doi: 10.1002/bit.25227. URL `http://doi.wiley.com/10.1002/bit.25227`.

Adam J. Brown, Bernie Sweeney, David O. Mainwaring, and David C. James. NF-κB, CRE and YY1 elements are key functional regulators of CMV promoter-driven transient gene expression in CHO cells. *Biotechnology Journal*, 10(7):1019–1028, jul 2015. ISSN 18607314. doi: 10.1002/biot.201400744. URL `http://doi.wiley.com/10.1002/biot.201400744`.

Adam J. Brown, Suzanne J. Gibson, Diane Hatton, and David C. James. In silico design of context-responsive mammalian promoters with user-defined functionality. *Nucleic Acids Research*, 45(18):10906–10919, oct 2017. ISSN 13624962. doi: 10.1093/nar/gkx768. URL `http://academic.oup.com/nar/article/45/18/10906/4097617`.

Adam J. Brown, Suzanne J. Gibson, Diane Hatton, Claire L. Arnall, and David C. James. Whole synthetic pathway engineering of recombinant protein production, nov 2019. ISSN 10970290. URL `http://doi.wiley.com/10.1002/bit.26855`.

W. Bruening, B. Giasson, W. Mushynski, and H. D. Durham. Activation of stress-activated MAP protein kinases up-regulates expression of transgenes driven by the cytomegalovirus immediate/early promoter. *Nucleic Acids Research*, 26(2):486–489, jan 1998. ISSN 03051048. doi: 10.1093/NAR/26.2.486. URL `https://www.researchgate.net/publication/13807773_Activation_`

of_stress-activated_MAP_protein_kinases_up-regulates_expression_of_transgenes_driven_by_the_cytomegalovirus_immediateearly_promoter.

Stefania C. Carrara, David Fiebig, Jan P. Bogen, Julius Grzeschik, Björn Hock, and Harald Kolmar. Recombinant antibody production using a dual-promoter single plasmid system. *Antibodies*, 10(2):18, jun 2021. ISSN 20734468. doi: 10.3390/ANTIB10020018/S1. URL https://www.mdpi.com/2073-4468/10/2/18/htmhttps://www.mdpi.com/2073-4468/10/2/18.

Dana Carroll. Genome Engineering with Targetable Nucleases. *Annual Review of Biochemistry*, 83(1):409–439, jun 2014. ISSN 0066-4154. doi: 10.1146/annurev-biochem-060713-035418.

Sittinan Chanarat and Katja Sträßer. Splicing and beyond: The many faces of the Prp19 complex. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1833(10):2126–2134, oct 2013. ISSN 0167-4889. doi: 10.1016/J.BBAMCR.2013.05.023.

Srinivasan Chandrasegaran and Dana Carroll. Origins of Programmable Nucleases for Genome Engineering, feb 2016. ISSN 10898638.

Rachel Chen. Bacterial expression systems for recombinant protein production: E. coli and beyond. *Biotechnology Advances*, 30(5):1102–1107, sep 2012. ISSN 0734-9750. doi: 10.1016/J.BIOTECHADV.2011.09.013.

Si-jia Chen, Wen Wang, Feng-yi Zhang, Yan-long Jia, Xiao-yin Wang, Xiao Guo, Shao-Nan Chen, Jian-hui Gao, and Tian-Yun Wang. A chimeric HS4 insulator-scaffold attachment region enhances transgene expression in transfected Chinese hamster ovary cells. *FEBS Open Bio*, 7(12):2021–2030, dec 2017. ISSN 22115463. doi: 10.1002/2211-5463.12335. URL http://doi.wiley.com/10.1002/2211-5463.12335.

Xiuling Chi, Qi Zheng, Ruhong Jiang, Ruby Yanru Chen-Tsai, and Ling-Jie Kong. A system for site-specific integration of transgenes in mammalian cells. *PLOS ONE*, 14(7):e0219842, jul 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0219842. URL http://dx.plos.org/10.1371/journal.pone.0219842.

Rossitza Christova and Thomas Oelgeschläger. Association of human TFIID–promoter complexes with silenced mitotic chromatin in vivo. *Nature Cell Biology*, 4(1):79–82, jan 2002. ISSN 1465-7392. doi: 10.1038/ncb733. URL http://www.nature.com/articles/ncb733.

Jonathan R. Chubb, Tatjana Trcek, Shailesh M. Shenoy, and Robert H. Singer. Transcriptional Pulsing of a Developmental Gene. *Current Biology*, 16(10):1018–1025, may 2006. ISSN 09609822. doi: 10.1016/j.cub.2006.03.092. URL http://www.ncbi.nlm.nih.gov/pubmed/16713960http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4764056http://linkinghub.elsevier.com/retrieve/pii/S0960982206014266.

L Stirling Churchman and Jonathan S Weissman. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330): 368–373, jan 2011. ISSN 00280836. doi: 10.1038/nature09652. URL http://www.ncbi.nlm.nih.gov/pubmed/21248844http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3880149.

Le Cong, F. Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D. Hsu, Xuebing Wu, Wenyan Jiang, Luciano A. Marraffini, and Feng Zhang. Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121): 819–823, feb 2013. ISSN 10959203. doi: 10.1126/science.1231143.

Brendan Cooney, Susan Dana Jones, and Howard L Levine. Quality by design for monoclonal antibodies, Part 1: Establishing the foundations for process development. *BioProcess International*, 14(6), 2016. ISSN 15426319. URL https://bioprocessintl.com/analytical/upstream-development/quality-by-design-for-monoclonal-antibodies-part-1-establishing-the-foundations-

GM Cooper. Eukaryotic RNA Polymerases and General Transcription Factors. In *The Cell: A Molecular Approach*, pages 4–7. Sinauer Associates, 2015. ISBN -10:0-87893-106-6. URL https://www.ncbi.nlm.nih.gov/books/NBK9935/http://www.ncbi.nlm.nih.gov/books/NBK9935/#A981.

Leighton Core and Karen Adelman. Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes Development*, 33 (15-16):960, aug 2019. ISSN 15495477. doi: 10.1101/GAD.325142.119. URL /pmc/articles/PMC6672056//pmc/articles/PMC6672056/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6672056/.

Leighton J Core, Joshua J Waterfall, and John T Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848, dec 2008. ISSN 00368075. doi: 10.1126/science.1162228. URL http://www.ncbi.nlm.nih.gov/pubmed/19056941http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2833333.

Sandra Cristea, Yevgeniy Freyvert, Yolanda Santiago, Michael C. Holmes, Fyodor D. Urnov, Philip D. Gregory, and Gregory J. Cost. In vivo cleavage of transgene donors promotes nuclease-mediated targeted integration. *Biotechnology and Bioengineering*, 110(3):871–880, mar 2013. ISSN 00063592. doi: 10.1002/bit.24733. URL http://doi.wiley.com/10.1002/bit.24733.

Jennifer A. Doudna and Emmanuelle Charpentier. The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213):1258096, nov 2014. ISSN 0036-8075. doi: 10.1126/science.1258096. URL https://www.sciencemag.org/lookup/doi/10.1126/science.1258096.

Ting Du, Nakita Buenbrazo, Laura Kell, Stephen G Withers, Shawn Defrees, and Warren Wakarchuk. A Bacterial Expression Platform for Production of Therapeutic

Proteins Containing Human-like O-Linked Glycans. *Cell Chemical Biology*, 26: 203–212.e5, 2019. doi: 10.1016/j.chembiol.2018.10.017. URL https://doi.org/10.1016/j.chembiol.2018.10.017.

Jennifer Dumont, Don Euwart, Baisong Mei, Scott Estes, and Rashmi Kshirsagar. Human cell lines for biopharmaceutical manufacturing: history, status, and future perspectives, dec 2016. ISSN 15497801. URL http://www.ncbi.nlm.nih.gov/pubmed/26383226http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5152558.

Sascha H.C. Duttke, Scott A. Lacadie, Mahmoud M. Ibrahim, Christopher K. Glass, David L. Corcoran, Christopher Benner, Sven Heinz, James T. Kadonaga, and Uwe Ohler. Human promoters are intrinsically directional. *Molecular Cell*, 57 (4):674–684, feb 2015. ISSN 10974164. doi: 10.1016/j.molcel.2014.12.029. URL https://www.sciencedirect.com/science/article/pii/S1097276514010077.

Belal El Kaderi, Scott Medler, Sarita Raghunayakula, and Athar Ansari. Gene Looping Is Conferred by Activator-dependent Interaction of Transcription Initiation and Termination Machineries. *Journal of Biological Chemistry*, 284(37):25015–25025, sep 2009. ISSN 0021-9258. doi: 10.1074/jbc.M109.007948. URL http://www.jbc.org/lookup/doi/10.1074/jbc.M109.007948.

David W. Emery, Evangelia Yannaki, Julie Tubb, and George Stamatoyannopoulos. A chromatin insulator protects retrovirus vectors from chromosomal position effects. *Proceedings of the National Academy of Sciences of the United States of America*, 97(16):9150–9155, aug 2000. ISSN 00278424. doi: 10.1073/pnas.160159597. URL http://www.ncbi.nlm.nih.gov/pubmed/10908661http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC16837.

Pär G. Engström, Harukazu Suzuki, Noriko Ninomiya, Altuna Akalin, Luca Sessa, Giovanni Lavorgna, Alessandro Brozzi, Lucilla Luzi, Sin Lam Tan, Liang Yang, Galih Kunarso, Edwin Lian-Chong Ng, Serge Batalov, Claes Wahlestedt, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, Christine Wells, Vladimir B. Bajic, Valerio Orlando, James F. Reid, Boris Lenhard, and Leonard Lipovich. Complex Loci in Human and Mouse Genomes. *PLoS Genetics*, 2(4):e47, 2006. ISSN 1553-7390. doi: 10.1371/journal.pgen.0020047. URL https://dx.plos.org/10.1371/journal.pgen.0020047.

Laure Escoubet-Lozach, Christopher Benner, Minna U. Kaikkonen, Jean Lozach, Sven Heinz, Nathan J. Spann, Andrea Crotti, Josh Stender, Serena Ghisletti, Donna Reichart, Christine S. Cheng, Rosa Luna, Colleen Ludka, Roman Sasik, Ivan Garcia-Bassets, Alexander Hoffmann, Shankar Subramaniam, Gary Hardiman, Michael G. Rosenfeld, and Christopher K. Glass. Mechanisms Establishing TLR4-Responsive Activation States of Inflammatory Response Genes. *PLoS Genetics*, 7(12):e1002401, dec 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002401. URL https://dx.plos.org/10.1371/journal.pgen.1002401.

Susan K Eszterhas, Eric E Bouhassira, David I K Martin, and Steven Fiering. Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. *Molecular and cellular biology*, 22(2):469–79, jan 2002. ISSN 0270-7306. doi: 10.1128/mcb.22.2.469-479.2002. URL http://www.ncbi.nlm.nih.gov/pubmed/11756543http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC139736.

Dan Y. Even, Adi Kedmi, Shani Basch-Barzilay, Diana Ideses, Ravid Tikotzki, Hila Shir-Shapira, Orit Shefi, and Tamar Juven-Gershon. Engineered Promoters for Potent Transient Overexpression. *PLOS ONE*, 11(2):e0148918, feb 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0148918. URL https://dx.plos.org/10.1371/journal.pone.0148918.

Domènec Farré, Nicolás Bellora, Loris Mularoni, Xavier Messeguer, and M. Mar Albà. Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biology*, 8(7):R140, jul 2007. ISSN 14747596. doi: 10.1186/gb-2007-8-7-r140. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-7-r140.

Cédric Feschotte and Ellen J. Pritham. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review of Genetics*, 41(1):331–368, dec 2007. ISSN 0066-4197. doi: 10.1146/annurev.genet.40.110405.090448.

J Fivaz, M C Bassi, S Pinaud, and J Mirkovitch. RNA polymerase II promoter-proximal pausing upregulates c-fos gene expression. *Gene*, 255(2):185–94, sep 2000. ISSN 0378-1119. URL http://www.ncbi.nlm.nih.gov/pubmed/11024278.

Nicholas J. Fuda, M. Behfar Ardehali, and John T. Lis. Defining mechanisms that regulate RNA polymerase II transcription in vivo, sep 2009. ISSN 00280836. URL http://www.nature.com/doifinder/10.1038/nature08449.

Daniel L. Galvan, Yozo Nakazawa, Aparna Kaja, Claudia Kettlun, Laurence J.N. N. Cooper, Cliona M. Rooney, and Matthew H. Wilson. No Title. *Journal of Immunotherapy*, 32(8):837–844, oct 2009. doi: 10.1097/CJI.0b013e3181b2914c. URL http://journals.lww.com/00002371-200910000-00007.

Brigitte Gasser and Diethard Mattanovich. Antibody production with yeasts and filamentous fungi: on the road to large scale? *Biotechnology Letters 2006 29:2*, 29 (2):201–212, nov 2006. ISSN 1573-6776. doi: 10.1007/S10529-006-9237-X. URL https://link.springer.com/article/10.1007/s10529-006-9237-x.

Aron M Geurts, Christopher S Hackett, Jason B Bell, Tracy L Bergemann, Lara S Collier, Corey M Carlson, David A Largaespada, and Perry B Hackett. Structure-based prediction of insertion-site preferences of transposons into chromosomes. *Nucleic Acids Research*, 34(9):2803–2811, jan 2006. ISSN 0305-1048. doi: 10.1093/nar/gkl301. URL https://doi.org/10.1093/nar/gkl301.

Darius Ghaderi, Mai Zhang, Nancy Hurtado-Ziola, and Ajit Varki. Production platforms for biotherapeutic glycoproteins. Occurrence, impact, and challenges of non-human sialylation. *Biotechnology and Genetic Engineering Reviews*, 28:147–175, 2012. ISSN 02648725. doi: 10.5661/bger-28-147. URL http://www.ncbi.nlm.nih.gov/pubmed/22616486.

Rodolfo Ghirlando, Keith Giles, Humaira Gowher, Tiaojiang Xiao, Zhixiong Xu, Hongjie Yao, and Gary Felsenfeld. Chromatin domains, insulators, and the regulation of gene expression, jul 2012. ISSN 18749399.

Pierre Alain Girod, Monique Zahn-Zabal, and Nicolas Mermod. Use of the chicken lysozyme 5 prime matrix attachment region to generate high producer CHO cell lines. *Biotechnology and Bioengineering*, 91(1):1–11, jul 2005. ISSN 00063592. doi: 10.1002/bit.20563. URL http://www.ncbi.nlm.nih.gov/pubmed/15889435.

Pierre Alain Girod, Duc Quang Nguyen, David Calabrese, Stefania Puttini, Mélanie Grandjean, Danielle Martinet, Alexandre Regamey, Damien Saugy, Jacques S. Beckmann, Philipp Bucher, and Nicolas Mermod. Genome-wide prediction of matrix attachment regions that increase gene expression in mammalian cells. *Nature Methods*, 4(9):747–753, sep 2007. ISSN 15487091. doi: 10.1038/nmeth1076.

Ivana Grabundzija, Markus Irgang, Lajos Mátés, Eyayu Belay, Janka Matrai, Andreas Gogol-Döring, Koichi Kawakami, Wei Chen, Patricia Ruiz, Marinee K.L. Chuah, Thierry Vandendriessche, Zsuzsanna Izsvák, and Zoltán Ivics. Comparative analysis of transposable element vector systems in human cells. *Molecular Therapy*, 18(6): 1200–1209, jun 2010. ISSN 15250016. doi: 10.1038/mt.2010.47.

Jay Gralla. Transcription Reinitiation Rate: a Special Role for the TATA Box. Technical Report 7, 1997. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC232232/pdf/173809.pdf.

Mélanie Grandjean, Pierre-Alain Girod, David Calabrese, Kaja Kostyrko, Marianne Wicht, Florence Yerly, Christian Mazza, Jacques S Beckmann, Danielle Martinet, and Nicolas Mermod. High-level transgene expression by homologous recombination-mediated gene transfer. *Nucleic acids research*, 39(15):e104, aug 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr436. URL http://www.ncbi.nlm.nih.gov/pubmed/21652640http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3159483.

Anthony JF Griffiths, Jeffrey H Miller, David T Suzuki, Richard C Lewontin, and William M Gelbart. Transcription and RNA polymerase. 2000. URL https://www.ncbi.nlm.nih.gov/books/NBK22085/.

Matthew G. Guenther, Stuart S. Levine, Laurie A. Boyer, Rudolf Jaenisch, and Richard A. Young. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell*, 130(1):77–88, jul 2007. ISSN 00928674. doi: 10.1016/

j.cell.2007.05.042. URL https://www.sciencedirect.com/science/article/pii/S0092867407006812?via%3Dihub.

Patrick Guye, Yinqing Li, Liliana Wroblewska, Xavier Duportet, and Ron Weiss. Rapid, modular and reliable construction of complex mammalian gene circuits. *Nucleic Acids Research*, 41(16):e156–e156, jul 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt605. URL https://doi.org/10.1093/nar/gkt605.

Vanja Haberle and Alexander Stark. Eukaryotic core promoters and the functional basis of transcription initiation, oct 2018. ISSN 14710080. URL http://www.nature.com/articles/s41580-018-0028-8.

Steven Hahn. Structure and mechanism of the RNA polymerase II transcription machinery. *Nature structural molecular biology*, 11(5): 394–403, may 2004. ISSN 1545-9993. doi: 10.1038/nsmb763. URL http://www.ncbi.nlm.nih.gov/pubmed/15114340http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1189732.

Antti Häkkinen, Shannon Healy, Howard T. Jacobs, and Andre S. Ribeiro. Genome wide study of NF-Y type CCAAT boxes in unidirectional and bidirectional promoters in human and mouse. *Journal of Theoretical Biology*, 281(1):74–83, jul 2011. ISSN 00225193. doi: 10.1016/j.jtbi.2011.04.027. URL https://www.sciencedirect.com/science/article/pii/S002251931100227X?via%3Dihub#bib13.

Hideki Hanawa, Motoko Yamamoto, Huifen Zhao, Takashi Shimada, and Derek A. Persons. Optimized lentiviral vector design improves titer and transgene expression of vectors containing the chicken β-globin locus HS4 insulator element. *Molecular Therapy*, 17(4):667–674, apr 2009. ISSN 15250016. doi: 10.1038/mt.2009.1. URL http://www.ncbi.nlm.nih.gov/pubmed/19223867http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2835111.

Niamh Harraghy, Armelle Gaussin, and Nicolas Mermod. Sustained Transgene Expression Using MAR Elements. *Current Gene Therapy*, 8(5):353–366, oct 2008. ISSN 15665232. doi: 10.2174/156652308786071032. URL http://www.ncbi.nlm.nih.gov/pubmed/18855632.

Niamh Harraghy, Alexandre Regamey, Pierre Alain Girod, and Nicolas Mermod. Identification of a potent MAR element from the mouse genome and assessment of its activity in stable and transient transfections. *Journal of Biotechnology*, 154(1): 11–20, jun 2011. ISSN 01681656. doi: 10.1016/j.jbiotec.2011.04.004.

Xiangjun He, Chunlai Tan, Feng Wang, Yaofeng Wang, Rui Zhou, Dexuan Cui, Wenxing You, Hui Zhao, Jianwei Ren, and Bo Feng. Knock-in of large reporter genes in human cells via CRISPR/Cas9-induced homology-dependent and independent DNA repair. *Nucleic acids research*, 44(9):e85, 2016. ISSN 1362-4962. doi: 10.1093/nar/gkw064. URL http://www.ncbi.nlm.nih.gov/pubmed/26850641http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4872082.

Ximiao He, Khund Sayeed Syed, Desiree Tillo, Ishminder Mann, Matthew T Weirauch, and Charles Vinson. GABPα Binding to Overlapping ETS and CRE DNA Motifs Is Enhanced by CREB1: Custom DNA Microarrays. *G358; Genes|Genomes|Genetics*, 5(9):1909–1918, jul 2015. ISSN 2160-1836. doi: 10.1534/g3.115.020248. URL http://www.ncbi.nlm.nih.gov/pubmed/26185160http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4555227.

Martin Herold, Marek Bartkuhn, and Rainer Renkawitz. CTCF: Insights into insulator function during development. *Development*, 139(6):1045–1057, mar 2012. ISSN 09501991. doi: 10.1242/dev.065268. URL http://www.ncbi.nlm.nih.gov/pubmed/22354838.

Mitchell Ho and Ira Pastan. Display and selection of scFv antibodies on HEK-293T cells. *Methods in molecular biology (Clifton, N.J.)*, 562:99, 2009. ISSN 10643745. doi: 10.1007/978-1-60327-302-2_8. URL /pmc/articles/PMC2790380//pmc/articles/PMC2790380/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2790380/.

Femke Hoeksema, Rik Van Blokland, Michel Siep, Karien Hamer, Tjalling Siersma, Jan Den Blaauwen, John Verhees, and Arie P. Otte. The use of a stringent selection system allows the identification of DNA elements that augment gene expression. *Molecular Biotechnology*, 48(1):19–29, may 2011. ISSN 10736085. doi: 10.1007/s12033-010-9344-8.

Sun Woo Hong, Seong Min Hong, Jae Wook Yoo, Young Chul Lee, Soyoun Kim, John T Lis, and Dong-Ki Lee. Phosphorylation of the RNA polymerase II C-terminal domain by TFIIH kinase is not essential for transcription of Saccharomyces cerevisiae genome. *Proceedings of the National Academy of Sciences*, 106(34):14276–14280, aug 2009. ISSN 0027-8424. doi: 10.1073/pnas.0903642106. URL http://www.ncbi.nlm.nih.gov/pubmed/19666497http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2732804http://www.pnas.org/content/106/34/14276.abstract%5Cnhttp://www.pnas.org/content/106/34/14276.full.pdf.

Patrick D. Hsu, Eric S. Lander, and Feng Zhang. Development and applications of CRISPR-Cas9 for genome engineering, jun 2014. ISSN 10974172.

Jianwen Hu, Jizhong Han, Haoran Li, Xian Zhang, Lan Lan Liu, Fei Chen, and Bin Zeng. Human Embryonic Kidney 293 Cells: A Vehicle for Biopharmaceutical Manufacturing, Structural Biology, and Electrophysiology. *Cells Tissues Organs*, 205(1):1–8, apr 2018. ISSN 1422-6405. doi: 10.1159/000485501. URL https://www.karger.com/Article/FullText/485501https://www.karger.com/Article/Abstract/485501.

Ying Huang, Yan Li, Yu Gang Wang, Xin Gu, Yan Wang, and Bei Fen Shen. An efficient and targeted gene integration system for high-level antibody expression. *Journal of Immunological Methods*, 322(1-2):28–39, apr 2007. ISSN 00221759. doi: 10.1016/j.jim.2007.01.022.

Ildikó Huliák, Ádám Sike, Sevil Zencir, and Imre M. Boros. The Objectivity of Reporters: Interference Between Physically Unlinked Promoters Affects Reporter Gene Expression in Transient Transfection Experiments. *DNA and Cell Biology*, 31 (11):1580–1584, 2012. ISSN 1044-5498. doi: 10.1089/dna.2012.1711.

Molly Hunter, Ping Yuan, Divya Vavilala, and Mark Fox. Optimization of Protein Expression in Mammalian Cells. *Current Protocols in Protein Science*, 95(1), feb 2019. ISSN 19343663. doi: 10.1002/CPPS.77.

Mara C. Inniss, Kalpanie Bandara, Barbara Jusiak, Timothy K. Lu, Ron Weiss, Liliana Wroblewska, and Lin Zhang. A novel Bxb1 integrase RMCE system for high fidelity site-specific integration of mAb expression cassette in CHO Cells. *Biotechnology and Bioengineering*, 114(8):1837–1846, aug 2017. ISSN 00063592. doi: 10.1002/bit.26268. URL `http://doi.wiley.com/10.1002/bit.26268`.

Masako Izumi and David M. Gilbert. Homogeneous tetracycline-regulatable gene expression in mammalian fibroblasts. *Journal of Cellular Biochemistry*, 76(2):280–289, feb 2000. ISSN 1097-4644. doi: 10.1002/(SICI)1097-4644(20000201)76:2<280::AID-JCB11>3.0.CO;2-0.

Linda B. Jacobsen, Susan A. Calvin, and Edward K. Lobenhofer. Transcriptional effects of transfection: the potential for misinterpretation of gene expression data generated from transiently transfected cells. *BioTechniques*, 47(1):617–624, jul 2009. ISSN 0736-6205. doi: 10.2144/000113132. URL `https://www.future-science.com/doi/10.2144/000113132`.

Karthik P. Jayapal, Katie F. Wlaschin, Wei Shou Hu, and Miranda G.S. Yap. Recombinant protein therapeutics from CHO Cells - 20 years and counting. *Chemical Engineering Progress*, 103(10):40–47, 2007. ISSN 15206106. doi: 10.1021/jp076244o.

Zhilian Jia, Jingwei Li, Xiao Ge, Yonghu Wu, Ya Guo, and Qiang Wu. Tandem CTCF sites function as insulators to balance spatial chromatin contacts and topological enhancer-promoter selection. *Genome Biology*, 21(1):75, mar 2020. ISSN 1474760X. doi: 10.1186/s13059-020-01984-7. URL `https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-01984-7`.

Ying Jiang and Jay D Gralla. Uncoupling of initiation and reinitiation rates during HeLa RNA polymerase II transcription in vitro. *Mol Cell Biol*, 13(8):4572–4577, 1993. ISSN 0270-7306. doi: 10.1128/MCB.13.8.4572. URL `http://mcb.asm.org/`.

Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821, aug 2012. ISSN 10959203. doi: 10.1126/science.1225829.

Yusuf B. Johari, Adam J. Brown, Christina S. Alves, Yizhou Zhou, Chapman M. Wright, Scott D. Estes, Rashmi Kshirsagar, and David C. James. CHO genome

mining for synthetic promoter design. *Journal of Biotechnology*, 294:1–13, mar 2019. ISSN 0168-1656. doi: 10.1016/J.JBIOTEC.2019.01.015. URL https://www.sciencedirect.com/science/article/pii/S0168165619300227.

Abayomi Oluwanbe Johnson, Miriam Gonzalez-Villanueva, Kang Lan Tee, and Tuck Seng Wong. No Title. *ACS Synthetic Biology*, 7(8):1918–1928, aug 2018. doi: 10.1021/acssynbio.8b00136. URL http://pubs.acs.org/doi/10.1021/acssynbio.8b00136.

Tamar Juven-Gershon, Susan Cheng, and James T Kadonaga. Rational design of a super core promoter that enhances gene expression. *Nature Methods*, 3(11):917–922, nov 2006. ISSN 15487091. doi: 10.1038/nmeth937. URL http://www.nature.com/articles/nmeth937.

Yujiro Kameyama, Yoshinori Kawabe, Akira Ito, and Masamichi Kamihira. An accumulative site-specific gene integration system using cre recombinase-mediated cassette exchange. *Biotechnology and Bioengineering*, 105(6):n/a–n/a, apr 2010. ISSN 00063592. doi: 10.1002/bit.22619. URL http://doi.wiley.com/10.1002/bit.22619.

Shin Young Kang, Yeon Gu Kim, Seunghee Kang, Hong Weon Lee, and Eun Gyo Lee. A novel regulatory element (E77) isolated from CHO-K1 genomic DNA enhances stable gene expression in Chinese hamster ovary cells. *Biotechnology Journal*, 11(5):633–641, may 2016. ISSN 18607314. doi: 10.1002/biot.201500464. URL http://www.ncbi.nlm.nih.gov/pubmed/26762773http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5067685.

Yoshinori Kawabe, Takanori Inao, Shodai Komatsu, Akira Ito, and Masamichi Kamihira. Cre-mediated cellular modification for establishing producer CHO cells of recombinant scFv-Fc. *BMC Proceedings*, 9(S9):1–3, dec 2015. ISSN 1753-6561. doi: 10.1186/1753-6561-9-s9-p5.

Jee Yon Kim, Yeon-Gu Kim, and Gyun Min Lee. CHO cells in biotechnology for production of recombinant proteins: current state and further potential. *Applied Microbiology and Biotechnology*, 93(3):917–930, feb 2012. ISSN 0175-7598. doi: 10.1007/s00253-011-3758-5. URL http://link.springer.com/10.1007/s00253-011-3758-5.

Jong Mook Kim, Jung Seob Kim, Doo Hong Park, Ho Sung Kang, Jaeseung Yoon, Kwanghee Baek, and Yeup Yoon. Improved recombinant gene expression in CHO cells using matrix attachment regions. *Journal of Biotechnology*, 107(2):95–105, jan 2004. ISSN 01681656. doi: 10.1016/j.jbiotec.2003.09.015.

Minsoo Kim, Peter M. O'Callaghan, Kurt A. Droms, and David C. James. A mechanistic understanding of production instability in CHO cell lines expressing recombinant monoclonal antibodies. *Biotechnology and Bioengineering*, 108(10):2434–2446, oct

2011. ISSN 00063592. doi: 10.1002/bit.23189. URL http://doi.wiley.com/10.1002/bit.23189.

Tae Hoon Kim, Ziedulla K. Abdullaev, Andrew D. Smith, Keith A. Ching, Dmitri I. Loukinov, Roland D D. Green, Michael Q. Zhang, Victor V. Lobanenkov, and Bing Ren. Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell*, 128(6):1231–1245, mar 2007. ISSN 00928674. doi: 10.1016/j.cell.2006.12.048.

Tae Kyung Kim and James H. Eberwine. Mammalian cell transfection: The present and the future. *Analytical and Bioanalytical Chemistry*, 397(8):3173–3178, aug 2010. ISSN 16182642. doi: 10.1007/s00216-010-3821-6.

Masahiro Kito, S. Itami, Y. Fukano, K. Yamana, and T. Shibui. Construction of engineered cho strains for high-level production of recombinant proteins. *Applied Microbiology and Biotechnology*, 60(4):442–448, dec 2002. ISSN 01757598. doi: 10.1007/s00253-002-1134-1.

Robert J. Klose and Adrian P. Bird. Genomic DNA methylation: The mark and its mediators, feb 2006. ISSN 09680004.

Igor Kondrychyn, Marta Garcia-Lecea, Alexander Emelyanov, Sergey Parinov, and Vladimir Korzh. Genome-wide analysis of Tol2 transposon reintegration in zebrafish. *BMC Genomics*, 10(1):418, sep 2009. ISSN 14712164. doi: 10.1186/1471-2164-10-418. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-418.

Kaja Kostyrko and Nicolas Mermod. Assays for DNA double-strand break repair by microhomology-based end-joining repair mechanisms. *Nucleic acids research*, 44(6):e56, apr 2016. ISSN 1362-4962. doi: 10.1093/nar/gkv1349. URL http://www.ncbi.nlm.nih.gov/pubmed/26657630http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4824085.

Kaja Kostyrko, Samuel Neuenschwander, Thomas Junier, Alexandre Regamey, Christian Iseli, Emanuel Schmid-Siegert, Sandra Bosshard, Stefano Majocchi, Valérie Le Fourn, Pierre-Alain Girod, Ioannis Xenarios, and Nicolas Mermod. MAR-Mediated transgene integration into permissive chromatin and increased expression by recombination pathway engineering. *Biotechnology and Bioengineering*, 114(2):384–396, feb 2017. ISSN 00063592. doi: 10.1002/bit.26086. URL http://doi.wiley.com/10.1002/bit.26086.

Simone Krinner. *The impact of intragenic CpG content on epigenetic control of transgene expression in mammalian cells*. PhD thesis, 2012. URL https://epub.uni-regensburg.de/27410/1/Dissertation-KrinnerS.pdf.

Shankarling Krishnamurthy and Michael Hampsey. Eukaryotic transcription initiation. *Current biology : CB*, 19(4):R153–6, feb 2009. ISSN 1879-0445. doi: 10.1016/j.cub.2008.11.052. URL http://www.ncbi.nlm.nih.gov/pubmed/19243687.

A Krumm, L B Hickey, and M Groudine. Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes development*, 9(5):559–72, mar 1995. ISSN 0890-9369. doi: 10.1101/GAD.9.5.559. URL http://www.ncbi.nlm.nih.gov/pubmed/7698646.

Stephanie H. Kung, Adam C. Retchless, Jessica Y. Kwan, and Rodrigo P.P. Almeida. Effects of DNA size on transformation and recombination efficiencies in xylella fastidiosa. *Applied and Environmental Microbiology*, 79(5):1712–1717, mar 2013. ISSN 00992240. doi: 10.1128/AEM.03525-12.

Jessica Kunkiel, Natascha Gödecke, Mania Ackermann, Dirk Hoffmann, Axel Schambach, Nico Lachmann, Dagmar Wirth, and Thomas Moritz. The CpG-sites of the CBX3 ubiquitous chromatin opening element are critical structural determinants for the anti-silencing function. *Scientific Reports*, 7(1):7919, dec 2017. ISSN 20452322. doi: 10.1038/s41598-017-04212-8. URL http://www.nature.com/articles/s41598-017-04212-8.

Ted H.J. Kwaks, Phil Barnett, Wieger Hemrika, Tjalling Siersma, Richard G.A.B. Sewalt, David P.E. Satijn, Janynke F. Brons, Rik Van Blokland, Paul Kwakman, Arle L. Kruckeberg, Angèle Kelder, and Arie P. Otte. Identification of anti-repressor elements that confer high and stable protein production in mammalian cells. *Nature Biotechnology*, 21(5):553–558, may 2003. ISSN 10870156. doi: 10.1038/nbt814.

Kim Le, Christopher Tan, Huong Le, Jasmine Tat, Ewelina Zasadzinska, Jonathan Diep, Ryan Zastrow, Chun Chen, and Jennitte Stevens. Assuring Clonality on the Beacon Digital Cell Line Development Platform. *Biotechnology journal*, 15(1), jan 2020. ISSN 1860-7314. doi: 10.1002/BIOT.201900247. URL https://pubmed.ncbi.nlm.nih.gov/31743597/.

Jae Seong Lee, Lise Marie Grav, Nathan E. Lewis, and Helene Faustrup Kildegaard. CRISPR/Cas9-mediated genome engineering of CHO cell factories: Application and perspectives, jul 2015a. ISSN 18607314.

Jae Seong Lee, Thomas Beuchert Kallehauge, Lasse Ebdrup Pedersen, and Helene Faustrup Kildegaard. Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. *Scientific Reports*, 5(1):1–11, feb 2015b. ISSN 20452322. doi: 10.1038/srep08572.

Jae Seong Lee, Lise Marie Grav, Lasse Ebdrup Pedersen, Gyun Min Lee, and Helene Faustrup Kildegaard. Accelerated homology-directed targeted integration of transgenes in Chinese hamster ovary cells via CRISPR/Cas9 and fluorescent enrichment. *Biotechnology and Bioengineering*, 113(11):2518–2523, nov 2016. ISSN 00063592. doi: 10.1002/bit.26002. URL http://doi.wiley.com/10.1002/bit.26002.

Aimin Li, Zengshan Liu, Qianxue Li, Lu Yu, Dacheng Wang, and Xuming Deng. Construction and characterization of bidirectional expression vectors in Saccha-

romyces cerevisiae. *FEMS Yeast Research*, 8(1):6–9, 2008. ISSN 15671356. doi: 10.1111/j.1567-1364.2007.00335.x. URL https://academic.oup.com/femsyr/article-abstract/8/1/6/563020.

Wang Li, Heng-Xin Li, Sheng-Yue Ji, Shuang Li, Yue-Sheng Gong, Ming-Ming Yang, and Yu-Lin Chen. Characterization of two temperature-inducible promoters newly isolated from B. subtilis. *Biochemical and Biophysical Research Communications*, 358(4):1148–1153, jul 2007. ISSN 0006-291X. doi: 10.1016/J.BBRC.2007.05.064. URL https://www.sciencedirect.com/science/article/pii/S0006291X07010376?via%3Dihub.

Qi Liang, Jun Kong, James Stalker, and Allan Bradley. Chromosomal mobilization and reintegration of Sleeping Beauty and PiggyBac transposons. *genesis*, 47(6):404–408, jun 2009. ISSN 1526954X. doi: 10.1002/dvg.20508. URL http://doi.wiley.com/10.1002/dvg.20508.

Jane M. Lin, Patrick J. Collins, Nathan D. Trinklein, Yutao Fu, Hualin Xi, Richard M. Myers, and Zhiping Weng. No Title. 17(6), jun 2007. doi: 10.1101/gr.5623407. URL http://www.ncbi.nlm.nih.gov/pubmed/17568000http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1891341.

Beihui Liu, Julian F Paton, and Sergey Kasparov. Viral vectors based on bidirectional cell-specific mammalian promoters and transcriptional amplification strategy for use in vitro and in vivo. *BMC Biotechnology*, 8(1):49, may 2008. ISSN 1472-6750. doi: 10.1186/1472-6750-8-49. URL http://bmcbiotechnol.biomedcentral.com/articles/10.1186/1472-6750-8-49.

Bo Liu, Zhanjiang Yuan, Kazuyuki Aihara, and Luonan Chen. Reinitiation enhances reliable transcriptional responses in eukaryotes. *Journal of the Royal Society Interface*, 11(97):20140326, aug 2014. ISSN 17425662. doi: 10.1098/rsif.2014.0326. URL http://www.ncbi.nlm.nih.gov/pubmed/24850905http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4208363.

Xiuli Liu, W. Lee Kraus, and Xiaoying Bai. Ready, pause, go: Regulation of RNA polymerase II pausing and release by cellular signaling pathways, sep 2015. ISSN 13624326. URL https://linkinghub.elsevier.com/retrieve/pii/S096800041500122X.

Yubing Liu, Soumyadeep Nandi, André Martel, Alen Antoun, Ilya Ioshikhes, and Alexandre Blais. Discovery, optimization and validation of an optimal DNA-binding sequence for the Six1 homeodomain transcription factor. *Nucleic Acids Research*, 40(17):8227–8239, sep 2012. ISSN 03051048. doi: 10.1093/nar/gks587. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks587.

Ruei Min Lu, Yu Chyi Hwang, I. Ju Liu, Chi Chiu Lee, Han Zen Tsai, Hsin Jung Li, and Han Chung Wu. Development of therapeutic antibodies for the treatment of diseases, jan 2020. ISSN 14230127. URL https://jbiomedsci.biomedcentral.com/articles/10.1186/s12929-019-0592-z.

Stefano Majocchi, Elena Aritonovska, and Nicolas Mermod. Epigenetic regulatory elements associate with specific histone modifications to prevent silencing of telomeric genes. *Nucleic Acids Research*, 42(1):193–204, 2014. ISSN 03051048. doi: 10.1093/nar/gkt880. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3874193/.

O. Maksimenko, N. B. Gasanov, and P. Georgiev. Regulatory elements in vectors for efficient generation of cell lines producing target proteins, 2015. ISSN 20758251.

V. H. Mann, M. E. Morales, K. J. Kines, and P. J. Brindley. Transgenesis of schistosomes: Approaches employing mobile genetic elements. *Parasitology*, 135(2):141–153, feb 2008. ISSN 00311820. doi: 10.1017/S0031182007003824.

Stefan J. Marciniak, Chi Y. Yun, Seiichi Oyadomari, Isabel Novoa, Yuhong Zhang, Rivka Jungreis, Kazuhiro Nagata, Heather P. Harding, and David Ron. CHOP induces death by promoting protein synthesis and oxidation in the stressed endoplasmic reticulum. *Genes Development*, 18(24):3066–3077, dec 2004. ISSN 0890-9369. doi: 10.1101/GAD.1250704. URL http://genesdev.cshlp.org/content/18/24/3066.fullhttp://genesdev.cshlp.org/content/18/24/3066http://genesdev.cshlp.org/content/18/24/3066.abstract.

Mariati, Esther YC Koh, Jessna HM Yeo, Steven CL Ho, and Yuansheng Yang. Toward stable gene expression in CHO cells: Preventing promoter silencing with core CpG island elements. *Bioengineered Bugs*, 5(5):340–345, sep 2014. ISSN 19491026. doi: 10.4161/bioe.32111. URL http://www.tandfonline.com/doi/abs/10.4161/bioe.32111.

Mattia Matasci, Lucia Baldi, David L. Hacker, and Florian M. Wurm. The PiggyBac transposon enhances the frequency of CHO stable cell line generation and yields recombinant lines with superior productivity and stability. *Biotechnology and Bioengineering*, 108(9):2141–2150, sep 2011. ISSN 00063592. doi: 10.1002/bit.23167. URL http://doi.wiley.com/10.1002/bit.23167.

Emma J. Mead, Lesley M. Chiverton, C. Mark Smales, and Tobias Der Von Haar. Identification of the limitations on recombinant gene expression in CHO cell lines with varying luciferase production rates. *Biotechnology and Bioengineering*, 102 (6):1593–1602, apr 2009. ISSN 00063592. doi: 10.1002/bit.22201. URL http://www.ncbi.nlm.nih.gov/pubmed/19090535.

Yaa Jyuhn J. Meir, Matthew T. Weirauch, Herng Shing Yang, Pei Cheng Chung, Robert K. Yu, and Sareina C.Y. Wu. Genome-wide target profiling of piggyBac and Tol2 in HEK 293: Pros and cons for gene discovery and gene therapy. *BMC Biotechnology*, 11(1):28, mar 2011. ISSN 14726750. doi: 10.1186/1472-6750-11-28. URL https://bmcbiotechnol.biomedcentral.com/articles/10.1186/1472-6750-11-28.

Djalila Mekahli, Geert Bultynck, Jan B. Parys, Humbert de Smedt, and Ludwig Missiaen. Endoplasmic-Reticulum Calcium Depletion and Disease. *Cold Spring Harbor Perspectives in Biology*, 3(6):1–30, jun 2011.

ISSN 19430264. doi: 10.1101/CSHPERSPECT.A004317. URL /pmc/articles/PMC3098671//pmc/articles/PMC3098671/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3098671/.

Olga Mikhaylichenko, Vladyslav Bondarenko, Dermot Harnett, Ignacio E Schor, Matilda Males, Rebecca R Viales, and Eileen E.M. Furlong. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes and Development*, 32(1):42–57, jan 2018. ISSN 15495477. doi: 10.1101/gad.308619.117. URL http://www.ncbi.nlm.nih.gov/pubmed/29378788http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5828394.

Kathryn Miller-Jensen, Ron Skupsky, Priya S. Shah, Adam P. Arkin, and David V. Schaffer. Genetic Selection for Context-Dependent Stochastic Phenotypes: Sp1 and TATA Mutations Increase Phenotypic Noise in HIV-1 Gene Expression. *PLoS Computational Biology*, 9(7):e1003135, jul 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003135. URL http://www.ncbi.nlm.nih.gov/pubmed/23874178http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3708878https://dx.plos.org/10.1371/journal.pcbi.1003135.

Jovan Mirkovitch, Marc Edouard Mirault, and Ulrich K. Laemmli. Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold. *Cell*, 39(1):223–232, 1984. ISSN 00928674. doi: 10.1016/0092-8674(84)90208-3.

Vladislav V. Mokhonov, Veena P. Theendakara, Yekaterina E. Gribanova, Novruz B. Ahmedli, and Debora B. Farber. Sequence-Specific Binding of Recombinant Zbed4 to DNA: Insights into Zbed4 Participation in Gene Transcription and Its Association with Other Proteins. *PLoS ONE*, 7(5):35317, may 2012. ISSN 19326203. doi: 10.1371/JOURNAL.PONE.0035317. URL /pmc/articles/PMC3365051//pmc/articles/PMC3365051/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3365051/.

Daniel Montiel, Hahk-Soo Kang, Fang-Yuan Chang, Zachary Charlop-Powers, and Sean F Brady. Yeast homologous recombination-based promoter engineering for the activation of silent natural product biosynthetic gene clusters. *Proceedings of the National Academy of Sciences*, 112(29):8953–8958, jul 2015. ISSN 0027-8424. doi: 10.1073/pnas.1507606112. URL http://www.ncbi.nlm.nih.gov/pubmed/26150486http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4517240.

Evelien Moorkens, Nicolas Meuwissen, Isabelle Huys, Paul Declerck, Arnold G Vulto, and Steven Simoens. The market of biopharmaceutical medicines: A snapshot of a diverse industrial landscape. *Frontiers in Pharmacology*, 8(JUN):314, 2017. ISSN 16639812. doi: 10.3389/fphar.2017.00314. URL http://www.ncbi.nlm.nih.gov/pubmed/28642701.

Matthew J. Moscou and Adam J. Bogdanove. A simple cipher governs DNA recognition by TAL effectors. *Science*, 326(5959):1501, dec 2009. ISSN 00368075. doi: 10.1126/science.1178817.

Théo Mozzanino. Engineering of the Secretory Pathway of CHO Cells for Recombinant Protein Production : Manipulation of SNARE Proteins. *Kent Academic Repository*, 20 (2):252–257, 2018. URL `http://kar.kent.ac.uk/contact.htmlhttp://kar.kent.ac.uk/contact.html%0Ahttps://doi.org/10.1080/14616742.2018.1457881`.

Fatemeh Naderi, Mehrdad Hashemi, Hadi Bayat, Omid Mohammadian, Es'hagh Pourmaleki, Mohammad Hossein Etemadzadeh, and Azam Rahimpour. The Augmenting Effects of the tDNA Insulator on Stable Expression of Monoclonal Antibody in Chinese Hamster Ovary Cells. *Monoclonal Antibodies in Immunodiagnosis and Immunotherapy*, 37(5):200–206, nov 2018. ISSN 21679436. doi: 10.1089/mab.2018.0015. URL `https://www.liebertpub.com/doi/10.1089/mab.2018.0015`.

Shota Nakade, Takuya Tsubota, Yuto Sakane, Satoshi Kume, Naoaki Sakamoto, Masanobu Obara, Takaaki Daimon, Hideki Sezutsu, Takashi Yamamoto, Tetsushi Sakuma, and Ken Ichi T. Suzuki. Microhomology-mediated end-joining-dependent integration of donor DNA in cells and animals using TALENs and CRISPR/Cas9. *Nature Communications*, 5(1):1–8, nov 2014. ISSN 20411723. doi: 10.1038/ncomms6560.

Toshiyuki Nakagawa, Hong Zhu, Nobuhiro Morishima, En Li, Jin Xu, Bruce A. Yankner, and Junying Yuan. Caspase-12 mediates endoplasmic-reticulum-specific apoptosis and cytotoxicity by amyloid-β. *Nature 2000 403:6765*, 403(6765):98–103, jan 2000. ISSN 1476-4687. doi: 10.1038/47513. URL `https://www.nature.com/articles/47513`.

Takahiro Nakayama, Tsukasa Shimojima, and Susumu Hirose. The PBAP remodeling complex is required for histone H3.3 replacement at chromatin boundaries and for boundary functions. *Development (Cambridge)*, 139(24):4582–4590, dec 2012. ISSN 09501991. doi: 10.1242/dev.083246. URL `http://www.ncbi.nlm.nih.gov/pubmed/23136390`.

Fatemeh Nematpour, Fereidoun Mahboudi, Behrouz Vaziri, Vahid Khalaj, Samira Ahmadi, Maryam Ahmadi, Saedeh Ebadat, and Fatemeh Davami. Evaluating the expression profile and stability of different UCOE containing vector combinations in mAb-producing CHO cells. *BMC Biotechnology*, 17(1):18, feb 2017. ISSN 14726750. doi: 10.1186/s12896-017-0330-0. URL `http://bmcbiotechnol.biomedcentral.com/articles/10.1186/s12896-017-0330-0`.

A. A. Nemudryi, K. R. Valetdinova, S. P. Medvedev, and S. M. Zakian. TALEN and CRISPR/Cas genome editing systems: Tools of discovery, 2014. ISSN 20758251.

Jonathan J. Neville, Joe Orlando, Kimberly Mann, Bethany McCloskey, and Michael N. Antoniou. Ubiquitous Chromatin-opening Elements (UCOEs): Applications in biomanufacturing and gene therapy, sep 2017. ISSN 07349750.

Hiroshi Nishimasu, F. Ann Ran, Patrick D. Hsu, Silvana Konermann, Soraya I. Shehata, Naoshi Dohmae, Ryuichiro Ishitani, Feng Zhang, and Osamu Nureki. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, 156(5):935–949, feb 2014. ISSN 00928674. doi: 10.1016/j.cell.2014.02.001.

Soo Min Noh, Seunghyeon Shin, and Gyun Min Lee. Comprehensive characterization of glutamine synthetase-mediated selection for the establishment of recombinant CHO cells producing monoclonal antibodies. *Scientific Reports*, 8(1):1–11, mar 2018. ISSN 20452322. doi: 10.1038/s41598-018-23720-9. URL `https://www.nature.com/articles/s41598-018-23720-9`.

Ajit Nott, Hervé Le Hir, and Melissa J. Moore. Splicing enhances translation in mammalian cells: an additional function of the exon junction complex. *Genes Development*, 18(2):210, jan 2004. ISSN 08909369. doi: 10.1101/GAD.1163204. URL `/pmc/articles/PMC324426//pmc/articles/PMC324426/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC324426/`.

Hirokazu Obayashi, Yoshinori Kawabe, Hirokatsu Makitsubo, Ryoko Watanabe, Yujiro Kameyama, Shuohao Huang, Yuta Takenouchi, Akira Ito, and Masamichi Kamihira. Accumulative gene integration into a pre-determined site using Cre/loxP. *Journal of Bioscience and Bioengineering*, 113(3):381–388, mar 2012. ISSN 13891723. doi: 10.1016/j.jbiosc.2011.10.027.

A. S. Orekhova and P. M. Rubtsov. No Title. *Biochemistry (Moscow)*, 78 (4), apr 2013. ISSN 0006-2979. doi: 10.1134/S0006297913040020. URL `https://link.springer.com/content/pdf/10.1134%2FS0006297913040020.pdfhttp://link.springer.com/10.1134/S0006297913040020`.

Salvatore J Orlando, Yolanda Santiago, Russell C DeKelver, Yevgeniy Freyvert, Elizabeth A Boydston, Erica A Moehle, Vivian M Choi, Sunita M Gopalan, Jacqueline F Lou, James Li, Jeffrey C Miller, Michael C Holmes, Philip D Gregory, Fyodor D Urnov, and Gregory J Cost. Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology. *Nucleic acids research*, 38(15):e152, aug 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq512. URL `http://www.ncbi.nlm.nih.gov/pubmed/20530528http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2926620`.

Andrea Osterlehner, Silke Simmeth, and Ulrich Göpfert. Promoter methylation and transgene copy numbers predict unstable protein production in recombinant chinese hamster ovary cell lines. *Biotechnology and Bioengineering*, 108(11):2670–2681, nov 2011. ISSN 00063592. doi: 10.1002/bit.23216. URL `http://doi.wiley.com/10.1002/bit.23216`.

A.P. Otte, T.H.J. Kwaks, R.J.M. VanBlokland, R.G.A.B. Sewalt, J. Verhees, V.N.A. Klaren, T.K. Siersma, H.W.M. Korse, N.C. Teunissen, S. Botschuijver, C. VanMer,

and S.Y. Man. Various Expression-Augmenting DNA Elements Benefit from STAR-Select, a Novel High Stringency Selection System for Protein Expression. *Biotechnology Progress*, 23(4):801–807, aug 2007. ISSN 8756-7938. doi: 10.1021/bp070107r. URL http://doi.wiley.com/10.1021/bp070107r.

Serguei Parinov, Igor Kondrichin, Vladimir Korzh, and Alexander Emelyanov. Tol2 transposon-mediated enhancer trap to identify developmentally regulated zebrafish genes in vivo. *Developmental Dynamics*, 231(2):449–459, oct 2004. ISSN 10588388. doi: 10.1002/dvdy.20157.

Siavash Partow, Verena Siewers, Sara Bjørn, Jens Nielsen, and Jérôme Maury. Characterization of different promoters for designing a new expression vector in Saccharomyces cerevisiae. *Yeast*, 27(11):955–964, nov 2010. ISSN 0749503X. doi: 10.1002/yea.1806. URL http://doi.wiley.com/10.1002/yea.1806.

John Patton, Scott Block, Chris Coombs, and Mark E. Martin. Identification of functional elements in the murine Gabpα/ATP synthase coupling factor 6 bi-directional promoter. *Gene*, 369(1-2):35–44, mar 2006. ISSN 03781119. doi: 10.1016/j.gene.2005.10.009. URL https://www.sciencedirect.com/science/article/pii/S037811190500613X.

B. Matija Peterlin and David H. Price. Controlling the Elongation Phase of Transcription with P-TEFb. *Molecular Cell*, 23(3):297–305, aug 2006. ISSN 1097-2765. doi: 10.1016/J.MOLCEL.2006.06.014. URL https://www.sciencedirect.com/science/article/pii/S1097276506004291?via%3Dihub.

Johannes Pichler, Friedemann Hesse, Matthias Wieser, Renate Kunert, Sybille S. Galosy, John E. Mott, and Nicole Borth. A study on the temperature dependency and time course of the cold capture antibody secretion assay. *Journal of Biotechnology*, 141 (1-2):80–83, apr 2009. ISSN 01681656. doi: 10.1016/J.JBIOTEC.2009.03.001. URL https://www.researchgate.net/publication/24414401_A_study_on_the_temperature_dependency_and_time_course_of_the_cold_capture_antibody_secretion_assay.

Jens Pontiller, Stefan Gross, Haruthai Thaisuchat, Friedemann Hesse, and Wolfgang Ernst. Identification of CHO endogenous promoter elements based on a genomic library approach. In *Molecular Biotechnology*, volume 39, pages 135–139. Springer, jun 2008. doi: 10.1007/s12033-008-9044-9.

Jens Pontiller, Andreas Maccani, Martina Baumann, Ingo Klancnik, and Wolfgang Ernst. Identification of CHO endogenous gene regulatory elements. *Molecular Biotechnology*, 45(3):235–240, jul 2010. ISSN 10736085. doi: 10.1007/s12033-010-9278-1.

Rui M C Portela, Thomas Vogl, Claudia Kniely, Jasmin E Fischer, Rui Oliveira, and Anton Glieder. Synthetic Core Promoters as Universal Parts for Fine-Tuning Expression in Different Yeast Species. *ACS synthetic biology*, 6(3):471–484, 2017. ISSN 2161-5063. doi: 10.1021/acssynbio.

6b00178. URL http://www.ncbi.nlm.nih.gov/pubmed/27973777http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5359585.

Nick J. Proudfoot. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science*, 352(6291), jun 2016. ISSN 10959203. doi: 10.1126/SCIENCE.AAD9926/ASSET/4BAC16B8-0F78-40E3-AE87-7EA560D74D36/ASSETS/GRAPHIC/352_AAD9926_FA.JPEG. URL https://www.science.org/doi/10.1126/science.aad9926.

Stefania Puttini, Ruthger W. van Zwieten, Damien Saugy, Małgorzata Lekka, Florence Hogger, Deborah Ley, Andrzej J. Kulik, and Nicolas Mermod. MAR-mediated integration of plasmid vectors for in vivo gene transfer and regulation. *BMC Molecular Biology*, 14:26, dec 2013. ISSN 14712199. doi: 10.1186/1471-2199-14-26. URL http://www.ncbi.nlm.nih.gov/pubmed/24295286http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4219123.

Jane Yuxia Qin, Li Zhang, Kayla L. Clift, Imge Hulur, Andy Peng Xiang, Bing Zhong Ren, and Bruce T. Lahn. Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS ONE*, 5(5):e10611, may 2010. ISSN 19326203. doi: 10.1371/journal.pone.0010611. URL http://dx.plos.org/10.1371/journal.pone.0010611.

Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biology*, 4(10):e309, sep 2006. ISSN 1545-7885. doi: 10.1371/journal.pbio.0040309. URL http://www.ncbi.nlm.nih.gov/pubmed/17048983http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1563489https://dx.plos.org/10.1371/journal.pbio.0040309.

F. Ann Ran, Patrick D. Hsu, Chie Yu Lin, Jonathan S. Gootenberg, Silvana Konermann, Alexandro E. Trevino, David A. Scott, Azusa Inoue, Shogo Matoba, Yi Zhang, and Feng Zhang. Double nicking by RNA-guided CRISPR cas9 for enhanced genome editing specificity. *Cell*, 154(6):1380–1389, sep 2013. ISSN 00928674. doi: 10.1016/j.cell.2013.08.021.

Deepak Reyon, Shengdar Q. Tsai, Cyd Khgayter, Jennifer A. Foden, Jeffry D. Sander, and J. Keith Joung. FLASH assembly of TALENs for high-throughput genome editing. *Nature Biotechnology*, 30(5):460–465, may 2012. ISSN 10870156. doi: 10.1038/nbt.2170.

Maria del Refugio Rocha-Pizaña, Guadalupe Ascencio-Favela, Brenda Maribell Soto-García, Margarita de la Luz Martinez-Fierro, and Mario Moisés Alvarez. Evaluation of changes in promoters, use of UCOES and chain order to improve the antibody production in CHO cells. *Protein Expression and Purification*, 132:108–115, apr 2017. ISSN 10465928. doi: 10.1016/j.pep.2017.01.014.

Nadiya Romanova and Thomas Noll. Engineered and Natural Promoters and Chromatin-Modifying Elements for Recombinant Protein Expression in CHO Cells, mar 2018. ISSN 18607314. URL http://doi.wiley.com/10.1002/biot.201700232.

Lorena Romero-Santacreu, Helena Orozco, Elena Garre, and Paula Alepuz. The bidirectional cytomegalovirus immediate/early promoter is regulated by Hog1 and the stress transcription factors Sko1 and Hot1 in yeast. *Molecular Genetics and Genomics*, 283(5):511–518, may 2010. ISSN 16174615. doi: 10.1007/s00438-010-0537-4. URL https://pubmed.ncbi.nlm.nih.gov/20364387/.

Carlotta Ronda, Lasse Ebdrup Pedersen, Henning Gram Hansen, Thomas Beuchert Kallehauge, Michael J. Betenbaugh, Alex Toftgaard Nielsen, and Helene Faustrup Kildegaard. Accelerating genome editing in CHO cells using CRISPR Cas9 and CRISPy, a web-based target finding tool. *Biotechnology and Bioengineering*, 111 (8):1604–1616, aug 2014. ISSN 00063592. doi: 10.1002/bit.25233. URL http://doi.wiley.com/10.1002/bit.25233.

Germán L. Rosano and Eduardo A. Ceccarelli. Recombinant protein expression in Escherichia coli: Advances and challenges. *Frontiers in Microbiology*, 5(APR):172, 2014. ISSN 1664302X. doi: 10.3389/FMICB.2014.00172/BIBTEX.

Emanuel Rosonina, Syuzo Kaneko, and James L Manley. Terminating the transcript: breaking up is hard to do. *Genes development*, 20(9):1050–6, may 2006. ISSN 0890-9369. doi: 10.1101/gad.1431606. URL http://www.ncbi.nlm.nih.gov/pubmed/16651651.

Tetsushi Sakuma, Mitsumasa Takenaga, Yoshinori Kawabe, Takahiro Nakamura, Masamichi Kamihira, and Takashi Yamamoto. Homologous Recombination-Independent Large Gene Cassette Knock-in in CHO Cells Using TALEN and MMEJ-Directed Donor Plasmids. *International Journal of Molecular Sciences*, 16(10):23849–23866, oct 2015. ISSN 1422-0067. doi: 10.3390/ijms161023849. URL http://www.mdpi.com/1422-0067/16/10/23849.

Tetsushi Sakuma, Shota Nakade, Yuto Sakane, Ken Ichi T. Suzuki, and Takashi Yamamoto. MMEJ-Assisted gene knock-in using TALENs and CRISPR-Cas9 with the PITCh systems. *Nature Protocols*, 11(1):118–133, jan 2016. ISSN 17502799. doi: 10.1038/nprot.2015.140.

Fay Saunders, Berni Sweeney, Michael N. Antoniou, Paul Stephens, and Katharine Cain. Chromatin Function Modifying Elements in an Industrial Antibody Production Platform - Comparison of UCOE, MAR, STAR and cHS4 Elements. *PLOS ONE*, 10 (4):e0120096, apr 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0120096. URL https://dx.plos.org/10.1371/journal.pone.0120096.

Michael R Schlabach, Jimmy K Hu, Mamie Li, and Stephen J Elledge. Synthetic design of strong promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 107(6):2538–43, feb 2010. ISSN 1091-6490. doi: 10.1073/pnas.

0914803107. URL http://www.ncbi.nlm.nih.gov/pubmed/20133776http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2823900.

Stefan Schlatter, Scott H. Stansfield, Diane M. Dinnis, Andrew J. Racher, John R. Birch, and David C. James. On the optimal ratio of heavy to light chain genes for efficient recombinant antibody production by CHO cells. *Biotechnology Progress*, 21(1): 122–133, jan 2005. ISSN 87567938. doi: 10.1021/bp049780w. URL https://aiche.onlinelibrary.wiley.com/doi/full/10.1021/bp049780whttps://aiche.onlinelibrary.wiley.com/doi/abs/10.1021/bp049780whttps://aiche.onlinelibrary.wiley.com/doi/10.1021/bp049780w.

Randy J. Seeley, Keith A. Yagaloff, Stewart L. Fisher, Paul Burn, Todd E. Thiele, Gertjan van Dijk, Denis G. Baskin, and Michael W. Schwartz. Melanocortin receptors in leptin effects. *Nature*, 390(6658):349–349, nov 1997. ISSN 0028-0836. doi: 10.1038/37016. URL http://www.nature.com/articles/37016.

Jayasha Shandilya and Stefan G E Roberts. The transcription cycle in eukaryotes: From productive initiation to RNA polymerase II recycling, may 2012. ISSN 18749399. URL https://www.sciencedirect.com/science/article/pii/S1874939912000284?via%3Dihub#f0015.

Xianzong Shi and Donald L. Jarvis. Protein N-Glycosylation in the Baculovirus-Insect Cell System. *Current drug targets*, 8(10):1116, oct 2007. ISSN 13894501. doi: 10.2174/138945007782151360. URL /pmc/articles/PMC3647355//pmc/articles/PMC3647355/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3647355/.

L S Shopland, K Hirayoshi, M Fernandes, and J T Lis. HSF access to heat shock elements in vivo depends critically on promoter architecture defined by GAGA factor, TFIID, and RNA polymerase II binding sites. *Genes development*, 9(22): 2756–69, nov 1995. ISSN 0890-9369. doi: 10.1101/GAD.9.22.2756. URL http://www.ncbi.nlm.nih.gov/pubmed/7590251.

Badri Nath Singh and Michael Hampsey. A Transcription-Independent Role for TFIIB in Gene Looping. *Molecular Cell*, 27(5):806–816, sep 2007. ISSN 1097-2765. doi: 10.1016/J.MOLCEL.2007.07.013. URL https://www.sciencedirect.com/science/article/pii/S1097276507004868.

Hanna L. Sladitschek and Pierre A. Neveu. Bidirectional Promoter Engineering for Single Cell MicroRNA Sensors in Embryonic Stem Cells. *PLOS ONE*, 11(5): e0155177, may 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0155177. URL https://dx.plos.org/10.1371/journal.pone.0155177.

Surabhi Srivastava, Deepika Puri, Hita Sony Garapati, Jyotsna Dhawan, and Rakesh K. Mishra. Vertebrate GAGA factor associated insulator elements demarcate homeotic genes in the HOX clusters. *Epigenetics chromatin*, 6(1), 2013. ISSN 1756-8935. doi: 10.1186/1756-8935-6-8. URL https://pubmed.ncbi.nlm.nih.gov/23607454/.

Kevin Struhl. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes, jul 1999. ISSN 00928674. URL http://www.ncbi.nlm.nih.gov/pubmed/10412974http://linkinghub.elsevier.com/retrieve/pii/S0092867400805991.

Kamilla Swiech, Virgínia Picanço-Castro, and Dimas Tadeu Covas. Human cells: new platform for recombinant therapeutic protein production. *Protein expression and purification*, 84(1):147–153, jul 2012. ISSN 1096-0279. doi: 10.1016/J.PEP.2012.04.023. URL https://pubmed.ncbi.nlm.nih.gov/22580292/.

Yasuhiro Takagi. Strategies To Improve Recombinant Protein Production. page 115, 2017.

Reza Taleei and Hooshang Nikjoo. Biochemical DSB-repair model for mammalian cells in G1 and early S phases of the cell cycle. *Mutation Research - Genetic Toxicology and Environmental Mutagenesis*, 756(1-2):206–212, may 2013. ISSN 13835718. doi: 10.1016/j.mrgentox.2013.06.004.

Sue Mei Tan-Wong, Hashanthi D. Wijayatilake, and Nick J. Proudfoot. Gene loops function to maintain transcriptional memory through interaction with the nuclear pore complex. *Genes and Development*, 23(22):2610–2624, nov 2009. ISSN 08909369. doi: 10.1101/gad.1823209. URL http://www.ncbi.nlm.nih.gov/pubmed/19933151http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2779764http://genesdev.cshlp.org/lookup/doi/10.1101/gad.1823209.

Katjana Tantale, Florian Mueller, Alja Kozulic-Pirher, Annick Lesne, Jean-Marc Victor, Marie-Cécile Robert, Serena Capozi, Racha Chouaib, Volker Bäcker, Julio Mateos-Langerak, Xavier Darzacq, Christophe Zimmer, Eugenia Basyuk, and Edouard Bertrand. A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. *Nature Communications*, 7(1):12248, dec 2016. ISSN 2041-1723. doi: 10.1038/ncomms12248. URL http://www.nature.com/articles/ncomms12248.

Naoko Tokuda, Masaki Sasai, and George Chikenji. Roles of DNA looping in enhancer-blocking activity. *Biophysical Journal*, 100(1):126–134, jan 2011. ISSN 15420086. doi: 10.1016/j.bpj.2010.11.016.

Nathan D Trinklein, Shelley Force Aldred, Sara J Hartman, Diane I Schroeder, Robert P Otillar, and Richard M Myers. An abundance of bidirectional promoters in the human genome. *Genome research*, 14(1):62–6, jan 2004. ISSN 1088-9051. doi: 10.1101/gr.1982804. URL http://www.ncbi.nlm.nih.gov/pubmed/14707170http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC314279.

Nagesh K. Tripathi and Ambuj Shrivastava. Recent Developments in Bioprocessing of Recombinant Proteins: Expression Hosts and Process Development, dec 2019. ISSN 22964185.

Jacqueline Unsinger, Andrea Kröger, Hansjörg Hauser, and Dagmar Wirth. Retroviral vectors for the transduction of autoregulated, bidirectional expression cassettes. *Molecular Therapy*, 4(5):484–489, nov 2001. ISSN 15250016. doi: 10.1006/mthe.2001.0480. URL https://www.sciencedirect.com/science/article/pii/S1525001601904800?via%3Dihub#bib17.

H. J.M. van Blokland, T. H.J. Kwaks, R. G.A.B. Sewalt, J. A. Verhees, V. N.A. Klaren, T. K. Siersma, J. W.M. Korse, N. C. Teunissen, S. Botschuijver, C. van Mer, S. Y. Man, and A. P. Otte. A novel, high stringency selection system allows screening of few clones for high protein expression. *Journal of Biotechnology*, 128(2):237–245, feb 2007. ISSN 01681656. doi: 10.1016/j.jbiotec.2006.09.023.

Thomas Vogl, Thomas Kickenweiz, Julia Pitzer, Lukas Sturmberger, Astrid Weninger, Bradley W. Biggs, Eva-Maria Maria Köhler, Armin Baumschlager, Jasmin Elgin Fischer, Patrick Hyden, Marlies Wagner, Martina Baumann, Nicole Borth, Martina Geier, Parayil Kumaran Ajikumar, Anton Glieder, Parayil Kumaran Ajikumar, and Anton Glieder. Engineered bidirectional promoters enable rapid multi-gene co-expression optimization. *Nature Communications*, 9 (1):3589, dec 2018. ISSN 20411723. doi: 10.1038/s41467-018-05915-w. URL www.nature.com/naturecommunicationshttp://www.nature.com/articles/s41467-018-05915-whttp://www.ncbi.nlm.nih.gov/pubmed/30181586http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6123417.

Nadezhda E. Vorobyeva, Marina U. Mazina, Anton K. Golovnin, Daria V. Kopytova, Dmitriy Y. Gurskiy, Elena N. Nabirochkina, Sofia G. Georgieva, Pavel G. Georgiev, and Aleksey N. Krasnov. Insulator protein Su(Hw) recruits SAGA and Brahma complexes and constitutes part of Origin Recognition Complex-binding sites in the Drosophila genome. *Nucleic Acids Research*, 41(11):5717–5730, 2013. ISSN 13624962. doi: 10.1093/nar/gkt297. URL https://www.ncbi.nlm.nih.gov/pubmed/23609538.

Gary Walsh. Biopharmaceutical benchmarks 2018. *Nature Biotechnology 2018 36:12*, 36(12):1136–1145, dec 2018. ISSN 1546-1696. doi: 10.1038/nbt.4305. URL https://www.nature.com/articles/nbt.4305.

Danyang Wang, Wei Dai, Jian Wu, and Jinke Wang. Improving transcriptional activity of human cytomegalovirus major immediate-early promoter by mutating NF-κB binding sites. *Protein Expression and Purification*, 142:16–24, feb 2018. ISSN 10465928. doi: 10.1016/j.pep.2017.09.008. URL https://www.sciencedirect.com/science/article/pii/S1046592817304345?via%3Dihub.

Tian-Yun Wang, Jun-He Zhang, Chang-Qin Jing, Xian-Jun Yang, and Jun-Tang Lin. Positional effects of the matrix attachment region on transgene expression in stably transfected CHO cells. *Cell biology international*, 34(2):141–5, feb 2010. ISSN 1095-8355. doi: 10.1042/CBI20090017. URL http://www.ncbi.nlm.nih.gov/pubmed/19947951.

Xiao-Yin Wang, Xi Zhang, Tian-Yun Wang, Yan-Long Jia, Dan-Hua Xu, and Dan-Dan Yi. Shortened nuclear matrix attachment regions are sufficient for replication and maintenance of episomes in mammalian cells. *Molecular Biology of the Cell*, 30 (22):2761–2770, oct 2019. ISSN 1059-1524. doi: 10.1091/mbc.E19-02-0108. URL https://www.molbiolcell.org/doi/10.1091/mbc.E19-02-0108.

Wu Wei, Vicent Pelechano, Aino I Jä, Lars M Steinmetz, Aino I Järvelin, and Lars M Steinmetz. Functional consequences of bidirectional promoters. 27(7), jul 2011. ISSN 0168-9525. doi: 10.1016/j.tig.2011.04.002. URL http://www.ncbi.nlm.nih.gov/pubmed/21601935http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3123404https://www.cell.com/trends/genetics/pdf/S0168-9525(11)00058-8.pdf.

Adam G. West, Miklos Gaszner, and Gary Felsenfeld. Insulators: Many functions, many mechanisms, feb 2002. ISSN 08909369.

Nathan R West. Development of a Tunable Mammalian Protein Expression System and an Investigation of Promoter Interference in Three Promoters Often Utilized in the Production of Biopharmaceuticals, aug 2014. URL http://etheses.whiterose.ac.uk/6701/1/NathanWest-FinalThesis140819.pdfhttp://etheses.whiterose.ac.uk/6701/.

Steven Williams, Tracey Mustoe, Tony Mulcahy, Mark Griffiths, David Simpson, Michael Antoniou, Alistair Irvine, Andrew Mountain, and Robert Crombie. CpG-island fragments from the HNRPA2B1/CBX3 genomic locus reduce silencing and enhance transgene expression from the hCMV promoter/enhancer in mammalian cells. *BMC Biotechnology*, 5(1):17, jun 2005. ISSN 14726750. doi: 10.1186/1472-6750-5-17. URL http://bmcbiotechnol.biomedcentral.com/articles/10.1186/1472-6750-5-17.

Matthew H. Wilson, Craig J. Coates, and Alfred L. George. PiggyBac transposon-mediated gene transfer in human cells. *Molecular Therapy*, 15(1):139–145, jan 2007. ISSN 15250016. doi: 10.1038/sj.mt.6300028.

Daniel Wong, Ana Teixeira, Spyros Oikonomopoulos, Peter Humburg, Imtiaz Nisar Lone, David Saliba, Trevor Siggers, Martha Bulyk, Dimitar Angelov, Stefan Dimitrov, Irina A Udalova, and Jiannis Ragoussis. Extensive characterization of NF-κB binding uncovers non-canonical motifs and advances the interpretation of genetic functional traits. *Genome Biology*, 12(7):R70, jul 2011a. ISSN 1474760X. doi: 10.1186/gb-2011-12-7-r70. URL http://www.ncbi.nlm.nih.gov/pubmed/21801342http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3218832.

S. P. Wong, O. Argyros, C. Coutelle, and R. P. Harbottle. Non-viral S/MAR vectors replicate episomally in vivo when provided with a selective advantage. *Gene Therapy*, 18(1):82–87, jan 2011b. ISSN 09697128. doi: 10.1038/gt.2010.116.

Olivier J Wouters, Martin McKee, and Jeroen Luyten. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, 323(9):844–853, mar 2020. ISSN 0098-7484. doi: 10.1001/jama.2020.1166. URL https://doi.org/10.1001/jama.2020.1166.

Sareina Chiung Yuan Wu, Yaa Jyuhn James Meir, Craig J. Coates, Alfred M. Handler, Pawel Pelczar, Stefan Moisyadi, and Joseph M. Kaminski. piggyBac is a flexible and highly active transposon as compared to Sleeping Beauty, Tol2, and Mos1 in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, 103(41):15008–15013, oct 2006. ISSN 00278424. doi: 10.1073/pnas.0606979103.

Chao Xu, Jiajia Chen, and Bairong Shen. The preservation of bidirectional promoter architecture in eukaryotes: What is the driving force? *BMC Systems Biology*, 6 (SUPPL.1):S21, jul 2012. ISSN 17520509. doi: 10.1186/1752-0509-6-S1-S21. URL http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-6-S1-S21http://www.ncbi.nlm.nih.gov/pubmed/23046569http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3403606.

Mamiko Yajima, William G. Fairbrother, and Gary M. Wessel. ISWI contributes to Arsi insulator function in development of the sea urchin. *Development (Cambridge)*, 139(19):3613–3622, oct 2012. ISSN 09501991. doi: 10.1242/dev.081828. URL http://www.ncbi.nlm.nih.gov/pubmed/22949616http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3436113.

Mary Qu Yang and Laura L Elnitski. Diversity of core promoter elements comprising human bidirectional promoters. In *BMC Genomics*, volume 9, page S3, sep 2008. doi: 10.1186/1471-2164-9-S2-S3. URL http://www.ncbi.nlm.nih.gov/pubmed/18831794http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2559893http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-9-S2-S3.

Song Yang, Sean C Sleight, and Herbert M Sauro. Rationally designed bidirectional promoter improves the evolutionary stability of synthetic genetic circuits. *Nucleic Acids Research*, 41(1):e33, jan 2013. ISSN 03051048. doi: 10.1093/nar/gks972. URL http://www.ncbi.nlm.nih.gov/pubmed/23093602http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3592475.

Yuansheng Yang, Mariati, Janet Chusainow, and Miranda G.S. Yap. DNA methylation contributes to loss in productivity of monoclonal antibody-producing CHO cell lines. *Journal of Biotechnology*, 147(3-4):180–185, jun 2010. ISSN 01681656. doi: 10.1016/j.jbiotec.2010.04.004.

Natalya Yudkovsky, Jeffrey A. Ranish, and Steven Hahn. A transcription reinitiation intermediate that is stabilized by activator. *Nature*, 408(6809):225–229, nov 2000. ISSN 0028-0836. doi: 10.1038/35041603. URL http://www.nature.com/articles/35041603.

Timur M. Yusufzai and Gary Felsenfeld. The 5 prime-HS4 chicken β-globin insulator is a CTCF-dependent nuclear matrix-associated element. *Proceedings of the National Academy of Sciences of the United States of America*, 101(23):8620–8624, jun 2004. ISSN 00278424. doi: 10.1073/pnas.0402938101. URL http://www.ncbi.nlm.nih.gov/pubmed/15169959http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC423244.

Monique Zahn-Zabal, Michel Kobr, Pierre Alain Girod, Markus Imhof, Philippe Chatellard, Maria De Jesus, Florian Wurm, and Nicolas Mermod. Development of stable cell lines for production or regulated expression using matrix attachment regions. *Journal of Biotechnology*, 87(1):29–42, apr 2001. ISSN 01681656. doi: 10.1016/S0168-1656(00)00423-5.

Ernesto Zanotto, Antti Häkkinen, Gabriel Teku, Bairong Shen, Andre S. Ribeiro, and Howard T. Jacobs. NF-Y influences directionality of transcription from the bidirectional Mrps12/Sarsm promoter in both mouse and human cells. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1789(5):432–442, may 2009. ISSN 18749399. doi: 10.1016/j.bbagrm.2009.05.001. URL https://www.sciencedirect.com/science/article/pii/S1874939909000546?via%3Dihub.

Fang Zhang, Amy R. Frost, Mike P. Blundell, Olivia Bales, Michael N. Antoniou, and Adrian J. Thrasher. A Ubiquitous Chromatin Opening Element (UCOE) confers resistance to DNA methylation-mediated silencing of lentiviral vectors. *Molecular Therapy*, 18(9):1640–1649, sep 2010. ISSN 15250024. doi: 10.1038/mt.2010.132. URL http://www.ncbi.nlm.nih.gov/pubmed/20588258http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2956914.

Fang Zhang, Giorgia Santilli, and Adrian J. Thrasher. Characterization of a core region in the A2UCOE that confers effective anti-silencing activity. *Scientific Reports*, 7(1):1–9, dec 2017. ISSN 20452322. doi: 10.1038/s41598-017-10222-3.

Xiao Hui Zhang, Louis Y. Tee, Xiao Gang Wang, Qun Shan Huang, and Shi Hua Yang. Off-target effects in CRISPR/Cas9-mediated genome engineering, nov 2015. ISSN 21622531.

Meiqi Menglin Zhao, Jiaxian Wang, Manyu Luo, Han Luo, Meiqi Menglin Zhao, Lei Han, Mengxiao Zhang, Hui Yang, Yueqing Xie, Hua Jiang, Lei Feng, Huili Lu, and Jianwei Zhu. Rapid development of stable transgene CHO cell lines by CRISPR/Cas9-mediated site-specific integration into C12orf35. *Applied Microbiology and Biotechnology*, 102(14):6105–6117, jul 2018. ISSN 14320614. doi: 10.1007/s00253-018-9021-6.

Hong Zhou, Zhi gang Liu, Zhi wei Sun, Ying Huang, and Wei yuan Yu. Generation of stable cell lines by site-specific integration of transgenes into engineered Chinese hamster ovary strains using an FLP-FRT system. *Journal of Biotechnology*, 147(2):122–129, may 2010. ISSN 01681656. doi: 10.1016/j.jbiotec.2010.03.020. URL http://www.ncbi.nlm.nih.gov/pubmed/20371256.

Roberto A. Zúñiga, Matías Gutiérrez-González, Norberto Collazo, Pablo Hérnan Sotelo, Carolina H. Ribeiro, Claudia Altamirano, Carmen Lorenzo, Juan Carlos Aguillón, and María Carmen Molina. Development of a new promoter to avoid the silencing of genes in the production of recombinant antibodies in chinese hamster ovary cells. *Journal of Biological Engineering*, 13(1):1–12, jun 2019. ISSN 17541611. doi: 10.1186/s13036-019-0187-y.

Appendix A

Vector Maps and Library Sequences

A.1 pSEAP2-CMVCore Vector map

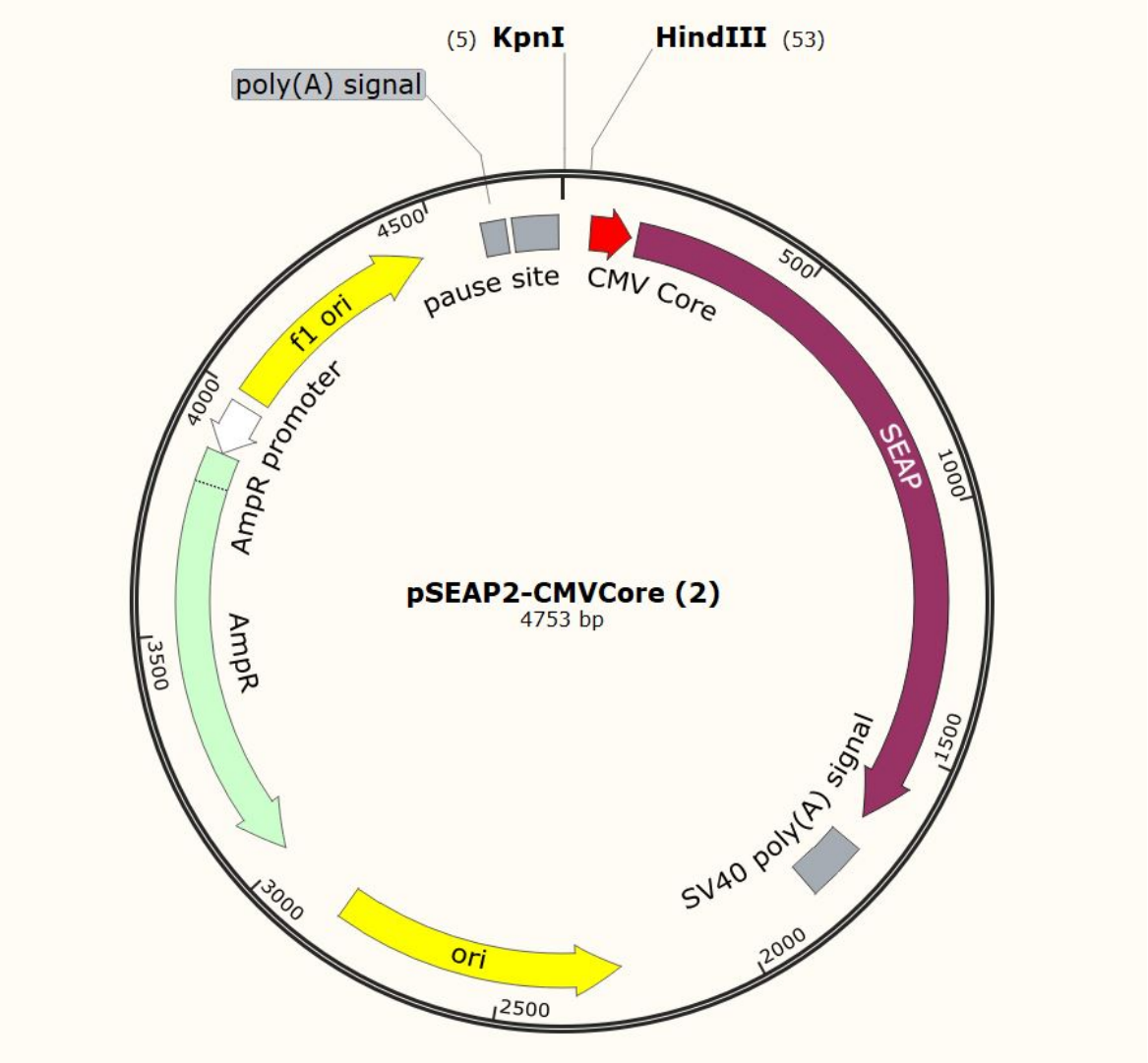


Figure A.1: pSEAP2-CMVCore vector map used in homotypic and heterotypic testing for SEAP

A.2 Full Sequences tested in TFRE Discovery Pipeline 1

>V$AHRARNT.03
GAAGCTCGAGTCGATGGTACCAGTGCGTGGGAATAAGTGC
GTGGGAATAAGTGCGTGGGAATAAGTGCGTGGGAATAAGT
GCGTGGGAATAAGTGCGTGGGAAAAGCTTGGTCAACGTCG
>V$BACH2.01
GAAGCTCGAGTCGATGGTACCCGTGAGTCATCTACGTGAGT
CATCTACGTGAGTCATCTACGTGAGTCATCTACGTGAGTCAT
CTACGTGAGTCATCAAGCTTGGTCAACGTCGA
>V$BARBIE.01
GAAGCTCGAGTCGATGGTACCAGCTAAAGCAGGAGGTAAG
CTAAAGCAGGAGGTAAGCTAAAGCAGGAGGTAAGCTAAAG
CAGGAGGTAAGCTAAAGCAGGAGGTAAGCTAAAGCAGGA
>V$CSRNP1.01
GAAGCTCGAGTCGATGGTACCAGAGTTAAGAGTTAAGAGTT
AAGAGTTAAGAGTTAAGAGTAAGCTTGGTCAACGTCGA
>V$CTCF.01
GAAGCTCGAGTCGATGGTACCCTCCCCGGCCGCTAGGGGG
CGGGCTACTCCCCGGCCGCTAGGGGGCGGGCTACTCCCCG
GCCGCTAGGGGGCGGGCTACTCCCCGGCCGCTAGGGGGCG
GGCTACTCCCCGGCCGCTAGGGGGCGGGCTACTCCCCGGCC
>V$E2F2.01
GAAGCTCGAGTCGATGGTACCAAGGCGCGCGATAAAGGCG
CGCGATAAAGGCGCGCGATAAAGGCGCGCGATAAAGGCGC
GCGATAAAGGCGCGCGAAAGCTTGGTCAACGTCGA
>V$EGR2.01
GAAGCTCGAGTCGATGGTACCTTGCGTGGGCGTTATTGCGT
GGGCGTTATTGCGTGGGCGTTATTGCGTGGGCGTTATTGCG
TGGGCGTTATTGCGTGGGCGTAAGCTTGGTCAACGTCGA
>V$GAGA.01
GAAGCTCGAGTCGATGGTACCGGGAGAGAGAGAGAGAGA
GAGAGATAGGGAGAGAGAGAGAGAGAGAGAGATAGGGA
GAGAGAGAGAGAGAGAGAGATAGGGAGAGAGAGAGAGA
GAGAGAGATAGGGAGAGAGAGAGAGAGAGAGAGATAGG
>V$GATA4.01
GAAGCTCGAGTCGATGGTACCAGAGATAAGATTAAGAGAT
AAGATTAAGAGATAAGATTAAGAGATAAGATTAAGAGATA
AGATTAAGAGATAAGATAAGCTTGGTCAACGTCGA
>V$HIC1.01
GAAGCTCGAGTCGATGGTACCTTATGCCAACCTATATTATGC
CAACCTATATTATGCCAACCTATATTATGCCAACCTATATTAT
GCCAACCTATATTATGCCAACCTAAAGCTTGGTCAACGTCGA
>V$KLF2.01
GAAGCTCGAGTCGATGGTACCAGGGGTGGGGTAAGGGGT
GGGGTAAGGGGTGGGGTAAGGGGTGGGGTAAGGGGTGG
GGTAAGGGGTGGGGAAGCTTGGTCAACGTCGA
>V$MAZR.01
GAAGCTCGAGTCGATGGTACCTGGGGGGGGGCCATATGGG
GGGGGGCCATATGGGGGGGGGCCATATGGGGGGGGGCCA
TATGGGGGGGGGCCATATGGGGGGGGGCCAAAGCTTGGT
>V$MEIS1A_HOXA9.01

GAAGCTCGAGTCGATGGTACCTGACAGTTTTACGATATGAC
AGTTTTACGATATGACAGTTTTACGATATGACAGTTTTACGA
TATGACAGTTTTACGATATGACAGTTTTACGAAAGCTTGGTC
>V$MIT.01
GAAGCTCGAGTCGATGGTACCGAGATCATGTGATGATAGA
GATCATGTGATGATAGAGATCATGTGATGATAGAGATCATG
TGATGATAGAGATCATGTGATGATAGAGATCATGTGATGAA
>V$NANOG.01
GAAGCTCGAGTCGATGGTACCTACTCATTCATTTATACTCAT
TCATTTATACTCATTCATTTATACTCATTCATTTATACTCATTC
ATTTATACTCATTCATTAAGCTTGGTCAACGTCGA
>V$NFIB.01
GAAGCTCGAGTCGATGGTACCCCTGGCTCCGTGCCAGCTTA
CCTGGCTCCGTGCCAGCTTACCTGGCTCCGTGCCAGCTTACC
TGGCTCCGTGCCAGCTTACCTGGCTCCGTGCCAGCTTACCTG
>V$NRF1.01
GAAGCTCGAGTCGATGGTACCGCCGCGCATGCGCATCTAGC
CGCGCATGCGCATCTAGCCGCGCATGCGCATCTAGCCGCGC
ATGCGCATCTAGCCGCGCATGCGCATCTAGCCGCGCATGCG
>V$RREB1.01
GAAGCTCGAGTCGATGGTACCCCCCAAACCACCCATACCCC
AAACCACCCATACCCCAAACCACCCATACCCCAAACCACCCA
TACCCCAAACCACCCATACCCCAAACCACCCAAAGCTTGGTC
>V$SMAD.01
GAAGCTCGAGTCGATGGTACCTGTCTGGCTTATGTCTGGCT
TATGTCTGGCTTATGTCTGGCTTATGTCTGGCTTATGTCTGG
>V$SOX6.01
GAAGCTCGAGTCGATGGTACCTCCTTTGTCTTATCCTTTGTC
TTATCCTTTGTCTTATCCTTTGTCTTATCCTTTGTCTTATCCTT
>V$TEAD4.01
GAAGCTCGAGTCGATGGTACCCTGCATTCCTCATACTGCATT
CCTCATACTGCATTCCTCATACTGCATTCCTCATACTGCATTC
CTCATACTGCATTCCTCAAAGCTTGGTCAACGTCGA
>V$YB1.01
GAAGCTCGAGTCGATGGTACCCTGATTGGCCAATACTGATT
GGCCAATACTGATTGGCCAATACTGATTGGCCAATACTGATT
GGCCAATACTGATTGGCCAAAAGCTTGGTCAACGTCGA

A.3 Full Sequences tested in TFRE Discovery Pipeline 2

>NF1.02
TATGGCACCATGCCAAGATATATGGCACCATGCCAAGATATAT
GGCACCATGCCAAGATATATGGCACCATGCCAAGATATATGGC
ACCATGCCAAGATATATGGCACCATGCCAAGATA

>nrf1.01
TAgCGCAGgcgcTATAgCGCAGgcgcTATAgCGCAGgcgcTATAgC
GCAGgcgcTATAgCGCAGgcgcTATAgCGCAGgcgcTA

>USF1.02
TAggtcacgtgTATAggtcacgtgTATAggtcacgtgTATAggtcacgtgTA
TAggtcacgtgTATAggtcacgtgTA
>v$atf2.01
TATGACGTAATATATGACGTAATATATGACGTAATATATGACGT
AATATATGACGTAATATATGACGTAATA
>V$ATF6.02
TATGACGTGTATATGACGTGTATATGACGTGTATATGACGTGT
ATATGACGTGTATATGACGTGTA
>V$BBX.01
TATGAACGACGTTCATATATGAACGACGTTCATATATGAACGA
CGTTCATATATGAACGACGTTCATATATGAACGACGTTCATATA
TGAACGACGTTCATA
>V$CLOCK_BMAL1.01
TAgggtCACGtgTATAgggtCACGtgTATAgggtCACGtgTATAgggtC
ACGtgTATAgggtCACGtgTATAgggtCACGtgTA
>V$E2F6.01
TAGGCGGGATATAGGCGGGATATAGGCGGGATATAGGCGGG
ATATAGGCGGGATATAGGCGGGATA
>V$EGR1.04
TAGGGCGGGGGCGGGGTATAGGGCGGGGGCGGGGTATAGG
GCGGGGGCGGGGTATAGGGCGGGGGCGGGGTATAGGGCGG
GGGCGGGGTATAGGGCGGGGGCGGGGTA
>V$ESRRA.02
TAAGGGGTCATATAAGGGGTCATATAAGGGGTCATATAAGGG
GTCATATAAGGGGTCATATAAGGGGTCATA
>V$ETS1.01
TACAGGAAGTGTATACAGGAAGTGTATACAGGAAGTGTATAC
AGGAAGTGTATACAGGAAGTGTATACAGGAAGTGTA

>V$ETV4.01
TACCGGAAGTTATACCGGAAGTTATACCGGAAGTTATACCGGA
AGTTATACCGGAAGTTATACCGGAAGTTA
>V$GABPA.02
TAACCGGAAGTTATAACCGGAAGTTATAACCGGAAGTTATAAC
CGGAAGTTATAACCGGAAGTTATAACCGGAAGTTA

>V$GABPA.02v.1
TACCGGAAGTGGTATACCGGAAGTGGTATACCGGAAGTGGTA
TACCGGAAGTGGTATACCGGAAGTGGTATACCGGAAGTGGTA

>v$gabpb1.01
TACCCGGAAGTGACTATACCCGGAAGTGACTATACCCGGAAGT
GACTATACCCGGAAGTGACTATACCCGGAAGTGACTATACCCG
GAAGTGACTA
>V$HIVEP1.01
TAGGGACTTTCCTATAGGGACTTTCCTATAGGGACTTTCCTATA
GGGACTTTCCTATAGGGACTTTCCTATAGGGACTTTCCTA

>V$HPF1.01
TAAGGACAAAGGCCAGCCTATAAGGACAAAGGCCAGCCTATA
AGGACAAAGGCCAGCCTATAAGGACAAAGGCCAGCCTATAAG
GACAAAGGCCAGCCTATAAGGACAAAGGCCAGCCTA

>V$HSF1.02
TAGAAGATTCGAGAACATTCTATAGAAGATTCGAGAACATTCT
ATAGAAGATTCGAGAACATTCTATAGAAGATTCGAGAACATTC
TATAGAAGATTCGAGAACATTCTATAGAAGATTCGAGAACATT
CTA
>V$HSF1.04
TATTCTGGAAGCTTCTTATATTCTGGAAGCTTCTTATATTCTGGA
AGCTTCTTATATTCTGGAAGCTTCTTATATTCTGGAAGCTTCTTA
TATTCTGGAAGCTTCTTA
>V$HSF1.05
TATTCCAGAATATATTCCAGAATATATTCCAGAATATATTCCAG
AATATATTCCAGAATATATTCCAGAATA
>V$HSF2.01
TAGAACATTTATAGAACATTTATAGAACATTTATAGAACATTTA
TAGAACATTTATAGAACATTTA
>V$KLF6.01
TAGGGGGCGGTATAGGGGGCGGTATAGGGGGCGGTATAGGG
GGCGGTATAGGGGGCGGTATAGGGGGCGGTA
>V$MAFK.01
TAAGTCAGCATTTTTATAAGTCAGCATTTTTATAAGTCAGCATTT
TTA
>V$MAZ.04
TAGGGAGGGGGTATAGGGAGGGGGTATAGGGAGGGGGTATA
GGGAGGGGGTATAGGGAGGGGGTATAGGGAGGGGGTA

>V$NFE2L2.01
TAtgctGAGTcat□TATAtgctGAGTcat□TATAtgctGAGTcat□TAT
AtgctGAGTcat□TATAtgctGAGTcat□TATAtgctGAGTcat□TA

>V$NM23.01
TAGGGTGGGGGGGGGGTATAGGGTGGGGGGGGGGTATAGGGT
GGGGGGGGGGTATAGGGTGGGGGGGGGGTATAGGGTGGGGGG
GGGTATAGGGTGGGGGGGGGGTA
>V$NR2F6.01
TAGGTCAAAGGTCTATAGGTCAAAGGTCTATAGGTCAAAGGTC
TATAGGTCAAAGGTCTATAGGTCAAAGGTCTATAGGTCAAAGG
TCTA

>V$PREB.01
TACATCATCAGACACCTATACATCATCAGACACCTATACATCAT
CAGACACCTATACATCATCAGACACCTATACATCATCAGACACC
TATACATCATCAGACACCTA
>V$RAR_RXR.03
TAGGGTCACAGAGAGTTCATATAGGGTCACAGAGAGTTCATAT
AGGGTCACAGAGAGTTCATATAGGGTCACAGAGAGTTCATATA
GGGTCACAGAGAGTTCATATAGGGTCACAGAGAGTTCATA

>V$RARG.01
TATGACCTTTTGTATATGACCTTTTGTATATGACCTTTTGTATAT
GACCTTTTGTATATGACCTTTTGTATATGACCTTTTGTA

>V$USF.04
TAGTCACGTGGTATAGTCACGTGGTATAGTCACGTGGTATAGT
CACGTGGTATAGTCACGTGGTATAGTCACGTGGTA
>V$XBP1.01
TAGATGACGTGTATAGATGACGTGTATAGATGACGTGTATAGA
TGACGTGTATAGATGACGTGTATAGATGACGTGTA

>V$YY1.04
TAGCCGCCATCTTGTATAGCCGCCATCTTGTATAGCCGCCATCT
TGTATAGCCGCCATCTTGTATAGCCGCCATCTTGTATAGCCGCC
ATCTTGTA

A.4 Heterotypic Promoters Tested in Library 1

>3'Weighted Promoter
GAAGCTCGAGTCGATGGTACCGTTGCGTGCGAAaaATGACA
CAGCAATaaGTTGCGTGCGAAaaGTTGCGTGCGAAaaATTGC
ATCAaaGTACGTGCaaATTGCATCAaaATTGCATCAaaAGGGG
CGGGGTaaGTACGTGCaaAGGGGCGGGGTaaATGACACAGC
AATaaGTACGTGCaaAGGGGCGGGGTaaAGGGGCGGGGTaa
ATGACACAGCAATaaGTACGTGCaaATGACACAGCAATaaAT
GACACAGCAATaaGTACGTGCaaGTTGCGTGCGAAaaAGGG
GCGGGGTaaATTGCATCAaaGTTGCGTGCGAAaaATTGCATC
AaaGGGACTTTCCaaGACCCGGATGTAGaaCCCCGGAAGTGA
CaaGACCCGGATGTAGaaGACCCGGATGTAGaaCCCCGGAA
GTGACaaGGGACTTTCCaaGGGACTTTCCaaGGGACTTTCCaa
GACCCGGATGTAGaaGGGACTTTCCaaCCCCGGAAGTGACaa
CCCCGGAAGTGACaaCCCCGGAAGTGACaaGGGACTTTCCaa
CCCCGGAAGTGACaaGACCCGGATGTAGaaGGGACTTTCCaa
>5' Weighted Promoter
GAAGCTCGAGTCGATGGTACCGGGACTTTCCaaGACCCGGA
TGTAGaaCCCCGGAAGTGACaaGACCCGGATGTAGaaGACCC
GGATGTAGaaCCCCGGAAGTGACaaGGGACTTTCCaaGGGAC
TTTCCaaGGGACTTTCCaaGACCCGGATGTAGaaGGGACTTTC
CaaCCCCGGAAGTGACaaCCCCGGAAGTGACaaCCCCGGAAG
TGACaaGGGACTTTCCaaCCCCGGAAGTGACaaGACCCGGAT
GTAGaaGGGACTTTCCaaCCCCGGAAGTGACaaGTTGCGTGC
GAAaaATGACACAGCAATaaGTTGCGTGCGAAaaGTTGCGTG
CGAAaaATTGCATCAaaGTACGTGCaaATTGCATCAaaATTGC
ATCAaaAGGGGCGGGGTaaGTACGTGCaaAGGGGCGGGGTa
aATGACACAGCAATaaGTACGTGCaaAGGGGCGGGGTaaAG
GGGCGGGGTaaATGACACAGCAATaaGTACGTGCaaATGAC
ACAGCAATaaATGACACAGCAATaaGTACGTGCaaGTTGCGT
GCGAAaaAGGGGCGGGGTaaATTGCATCAaaGTTGCGTGCG
>Heterotypic TFRE Compliment Prom 1
GAAGCTCGAGTCGATGGTACCCTGGGCCTACATCttGGGGCC
TTCACTGttCAACGCACGCTTttCCCTGAAAGGttTCCCCGCCC
CAttCCCCGCCCCCCTCttCCCTGAAAGGttTACTGTGTCGTTAt
tGGGGCCTTCACTGttCATGCACGttCCCCGCCCCCCTCttCTGG
GCCTACATCttTACTGTGTCGTTAttCAACGCACGCTTttCCCTG
AAAGGttGGGGCCTTCACTGttCAACGCACGCTTttCATGCACG
ttTAACGTAGTttCTGGGCCTACATCttTAACGTAGTttCTGGGC
CTACATCttCAACGCACGCTTttGGGGCCTTCACTGttCCCTGA
AAGGttGGGGCCTTCACTGttCATGCACGttTAACGTAGTttTC
CCCGCCCCAttCATGCACGAAGCTTGGTCAACGTCGA
>Heterotypic TFRE Reverse Compitent Prom 1

GAAGCTCGAGTCGATGGTACCGCACGTACttACCCCGCCCCTt
tTGATGCAATttGCACGTACttGTCACTTCCGGGGttGGAAAGT
CCCttGTCACTTCCGGGGttTTCGCACGCAACttCTACATCCGG
GTCttTGATGCAATttCTACATCCGGGTCttTGATGCAATttGCA
CGTACttTTCGCACGCAACttGTCACTTCCGGGGttGGAAAGTC
CCttTTCGCACGCAACttATTGCTGTGTCATttCTACATCCGGGT
CttCTCCCCCCGCCCCttGCACGTACttGTCACTTCCGGGGttAT
TGCTGTGTCATttGGAAAGTCCCttCTCCCCCCGCCCCttACCCC
GCCCCTttGGAAAGTCCCttTTCGCACGCAACttGTCACTTCCG
GGGttCTACATCCGGGTCAAGCTTGGTCAACGTCGA
>Hetrotypic High Transcription Promoter
GAAGCTCGAGTCGATGGTACCGACCCGGATGTAGaaCCCCG
GAAGTGACaaGTTGCGTGCGAAaaGGGACTTTCCaaAGGGG
CGGGGTaaGGGGCGGGGGGAGaaGGGACTTTCCaaATGACA
CAGCAATaaCCCCGGAAGTGACaaGTACGTGCaaGGGGCGG
GGGGAGaaGACCCGGATGTAGaaATGACACAGCAATaaGTT
GCGTGCGAAaaGGGACTTTCCaaCCCCGGAAGTGACaaGTTG
CGTGCGAAaaGTACGTGCaaATTGCATCAaaGACCCGGATGT
AGaaATTGCATCAaaGACCCGGATGTAGaaGTTGCGTGCGAA
aaCCCCGGAAGTGACaaGGGACTTTCCaaCCCCGGAAGTGAC
aaGTACGTGCaaATTGCATCAaaAGGGGCGGGGTaaGTACGT
>Hetrotypic High Transcription Promoter without
GAAGCTCGAGTCGATGGTACCGACCCGGATGTAGaaCCCCG
GAAGTGACaaGTTGCGTGCGAAaaGGGACTTTCCaaGGGACT
TTCCaaATGACACAGCAATaaCCCCGGAAGTGACaaGTACGT
GCaaGACCCGGATGTAGaaATGACACAGCAATaaGTTGCGTG
CGAAaaGGGACTTTCCaaCCCCGGAAGTGACaaGTTGCGTGC
GAAaaGTACGTGCaaATTGCATCAaaGACCCGGATGTAGaaA
TTGCATCAaaGACCCGGATGTAGaaGTTGCGTGCGAAaaCCC
CGGAAGTGACaaGGGACTTTCCaaCCCCGGAAGTGACaaGTA
CGTGCaaATTGCATCAaaGTACGTGCAAGCTTGGTCAACGTC
>Promoter with 3' repressor
GAAGCTCGAGTCGATGGTACCGACCCGGATGTAGaaCCCCG
GAAGTGACaaGTTGCGTGCGAAaaGGGACTTTCCaaGGGACT
TTCCaaATGACACAGCAATaaCCCCGGAAGTGACaaGTACGT
GCaaGACCCGGATGTAGaaATGACACAGCAATaaGTTGCGTG
CGAAaaGGGACTTTCCaaCCCCGGAAGTGACaaGTTGCGTGC
GAAaaGTACGTGCaaATTGCATCAaaGACCCGGATGTAGaaA
TTGCATCAaaGACCCGGATGTAGaaGTTGCGTGCGAAaaCCC
CGGAAGTGACaaGGGACTTTCCaaCCCCGGAAGTGACaaGTA
CGTGCaaATTGCATCAaaGTACGTGCaaCGCCATTTTaaCGCC
ATTTTaaCGCCATTTTAAGCTTGGTCAACGTCGA
>Promoter With 5' repressor

GAAGCTCGAGTCGATGGTACCCGCCATTTTaaCGCCATTTTaa
CGCCATTTTaaGACCCGGATGTAGaaCCCCGGAAGTGACaaG
TTGCGTGCGAAaaGGGACTTTCCaaGGGACTTTCCaaATGAC
ACAGCAATaaCCCCGGAAGTGACaaGTACGTGCaaGACCCGG
ATGTAGaaATGACACAGCAATaaGTTGCGTGCGAAaaGGGAC
TTTCCaaCCCCGGAAGTGACaaGTTGCGTGCGAAaaGTACGT
GCaaATTGCATCAaaGACCCGGATGTAGaaATTGCATCAaaGA
CCCGGATGTAGaaGTTGCGTGCGAAaaCCCCGGAAGTGACaa
GGGACTTTCCaaCCCCGGAAGTGACaaGTACGTGCaaATTGC
ATCAaaGTACGTGCAAGCTTGGTCAACGTCGA
>Really High TFRE
GAAGCTCGAGTCGATGGTACCGACCCGGATGTAGaaATGAC
ACAGCAATaaATTGCATCAaaGTTGCGTGCGAAaaAGGGGCG
GGGTaaCCCCGGAAGTGACaaCCCCGGAAGTGACaaCCCCGG
AAGTGACaaGGGACTTTCCaaGACCCGGATGTAGaaATTGCA
TCAaaGTTGCGTGCGAAaaGTACGTGCaaATTGCATCAaaAG
GGGCGGGGTaaATTGCATCAaaATGACACAGCAATaaGGGA
CTTTCCaaGTTGCGTGCGAAaaCCCCGGAAGTGACaaGTTGC
GTGCGAAaaGTACGTGCaaGGGACTTTCCaaGACCCGGATGT
AGaaGACCCGGATGTAGaaGTACGTGCaaCCCCGGAAGTGAC
aaATGACACAGCAATaaCCCCGGAAGTGACaaCCCCGGAAGT
GACaaGGGACTTTCCaaGGGACTTTCCaaATTGCATCAaaAGG
GGCGGGGTaaGGGACTTTCCaaGTTGCGTGCGAAaaGTACGT
GCaaAGGGGCGGGGTaaGGGACTTTCCaaATGACACAGCAA
TaaAGGGGCGGGGTaaGTACGTGCaaGACCCGGATGTAGaa
>Remixed High TFRE
GAAGCTCGAGTCGATGGTACCGTACGTGCaaATGACACAGC
AATaaCCCCGGAAGTGACaaATTGCATCAaaGTACGTGCaaAT
TGCATCAaaGGGACTTTCCaaGACCCGGATGTAGaaGGGACT
TTCCaaGACCCGGATGTAGaaATGACACAGCAATaaCCCCGG
AAGTGACaaAGGGGCGGGGTaaGGGACTTTCCaaCCCCGGA
AGTGACaaGACCCGGATGTAGaaAGGGGCGGGGTaaGTACG
TGCaaCCCCGGAAGTGACaaGTTGCGTGCGAAaaGTTGCGTG
CGAAaaGGGACTTTCCaaAGGGGCGGGGTaaGGGACTTTCCa
aCCCCGGAAGTGACaaATTGCATCAaaGGGACTTTCCaaGTTG
CGTGCGAAaaGACCCGGATGTAGaaCCCCGGAAGTGACaaA
GGGGCGGGGTaaGTACGTGCaaAGGGGCGGGGTaaATGACA
CAGCAATaaATGACACAGCAATaaGACCCGGATGTAGaaATT
GCATCAaaGTTGCGTGCGAAaaATGACACAGCAATaaATTGC
ATCAaaGGGACTTTCCaaCCCCGGAAGTGACaaGTACGTGCaa

A.5 Heterotypic Promoters Tested in Library 2

>100bp Promoter single copies
GAAGCTCGAGTCGATGGTACCGGGACTTTCCaaGTACG
TGCaaATGACACAGCAATaaGACCCGGATGTAGaaCCCC
GGAAGTGACaaGTTGCGTGCGAAaaGCCGCGCATGCG
CATCaaATTGCATCAAAGCTTGGTCAACGTCGA

>200bp Promoter two copies
GAAGCTCGAGTCGATGGTACCGGGACTTTCCaaGTACG
TGCaaATGACACAGCAATaaGACCCGGATGTAGaaCCCC
GGAAGTGACaaGTTGCGTGCGAAaaGCCGCGCATGCG
CATCaaATTGCATCAaaGGGACTTTCCaaGTACGTGCaa
ATGACACAGCAATaaGACCCGGATGTAGaaCCCCGGAA
GTGACaaGTTGCGTGCGAAaaGCCGCGCATGCGCATCa
aATTGCATCAAAGCTTGGTCAACGTCGA

>3' Weighted Promoter with 5' Repressor
GAAGCTCGAGTCGATGGTACCCGCCATTTTaaCGCCATT
TTaaCGCCATTTTaaGTTGCGTGCGAAaaGTACGTGCaa
GTACGTGCaaGTACGTGCaaATTGCATCAaaGTTGCGTG
CGAAaaGTTGCGTGCGAAaaATTGCATCAaaGTACGTGC
aaGTTGCGTGCGAAaaGACCCGGATGTAGaaCCCCGGA
AGTGACaaATGACACAGCAATaaGACCCGGATGTAGaa
GGGACTTTCCaaGGGACTTTCCaaGGGACTTTCCaaGAC
CCGGATGTAGaaCCCCGGAAGTGACaaGACCCGGATGT
AGaaGGGACTTTCCaaGGGACTTTCCaaCCCCGGAAGT
GACaaATGACACAGCAATaaCCCCGGAAGTGACaaGCC
GCGCATGCGCATCaaGCCGCGCATGCGCATCaaGCCGC
GCATGCGCATCaaGCCGCGCATGCGCATCaaCCCCGGA
AGTGACaaGGGACTTTCCaaGCCGCGCATGCGCATCaaA
TGACACAGCAATaaCCCCGGAAGTGACaaATGACACAG
CAATAAGCTTGGTCAACGTCGA

>300bp Promoter
GAAGCTCGAGTCGATGGTACCGGGACTTTCCaaGTACG
TGCaaATGACACAGCAATaaGACCCGGATGTAGaaCCCC
GGAAGTGACaaGTTGCGTGCGAAaaGCCGCGCATGCG
CATCaaATTGCATCAaaGGGACTTTCCaaGTACGTGCaa
ATGACACAGCAATaaGACCCGGATGTAGaaCCCCGGAA
GTGACaaGTTGCGTGCGAAaaGCCGCGCATGCGCATCa
aATTGCATCAaaGGGACTTTCCaaGTACGTGCaaATGAC
ACAGCAATaaGACCCGGATGTAGaaCCCCGGAAGTGAC
aaGTTGCGTGCGAAaaGCCGCGCATGCGCATCaaATTG
CATCAAAGCTTGGTCAACGTCGA

>Balanced Promoter

GAAGCTCGAGTCGATGGTACCGGGACTTTCCaaGTTGC
GTGCGAAaaCCCCGGAAGTGACaaATGACACAGCAATa
aGGGACTTTCCaaGGGACTTTCCaaATGACACAGCAATa
aGCCGCGCATGCGCATCaaGGGACTTTCCaaATGACACA
GCAATaaCCCCGGAAGTGACaaCCCCGGAAGTGACaaG
ACCCGGATGTAGaaATTGCATCAaaATGACACAGCAATa
aGGGACTTTCCaaGACCCGGATGTAGaaGTACGTGCaa
GACCCGGATGTAGaaGACCCGGATGTAGaaCCCCGGAA
GTGACaaGGGACTTTCCaaGTACGTGCaaGTTGCGTGC
GAAaaCCCCGGAAGTGACaaATTGCATCAaaGTACGTG
CaaGCCGCGCATGCGCATCaaGCCGCGCATGCGCATCaa
CCCCGGAAGTGACaaGCCGCGCATGCGCATCaaGCCGC
GCATGCGCATCaaGTTGCGTGCGAAaaGTTGCGTGCGA
AaaGTACGTGCAAGCTTGGTCAACGTCGA

>Balanced Promoter without Spacers
GAAGCTCGAGTCGATGGTACCGGGACTTTCCGTTGCGT
GCGAACCCCGGAAGTGACATGACACAGCAATGGGACT
TTCCGGGACTTTCCATGACACAGCAATGCCGCGCATGC
GCATCGGGACTTTCCATGACACAGCAATCCCCGGAAGT
GACCCCCGGAAGTGACGACCCGGATGTAGATTGCATC
AATGACACAGCAATGGGACTTTCCGACCCGGATGTAG
GTACGTGCGACCCGGATGTAGGACCCGGATGTAGCCC
CGGAAGTGACGGGACTTTCCGTACGTGCGTTGCGTGC
GAACCCCGGAAGTGACATTGCATCAGTACGTGCGCCG
CGCATGCGCATCGCCGCGCATGCGCATCCCCCGGAAGT
GACGCCGCGCATGCGCATCGCCGCGCATGCGCATCGTT
GCGTGCGAAGTTGCGTGCGAAGTACGTGCAAGCTTGG
TCAACGTCGA

>Reduced Promoter
GAAGCTCGAGTCGATGGTACCCCCCGGAAGTGACaaG
CCGCGCATGCGCATCaaGACCCGGATGTAGaaGTTGCG
TGCGAAaaGTACGTGCaaGCCGCGCATGCGCATCaaGA
CCCGGATGTAGaaGACCCGGATGTAGaaATGACACAGC
AATaaATGACACAGCAATaaCCCCGGAAGTGACaaCCCC
GGAAGTGACaaGTTGCGTGCGAAaaGGGACTTTCCaaG
TACGTGCaaGCCGCGCATGCGCATCaaGGGACTTTCCaa
ATTGCATCAaaGGGACTTTCCAAGCTTGGTCAACGTCG
A
>Reduced Promoter_without_Spacer
GAAGCTCGAGTCGATGGTACCCCCCGGAAGTGACGCC
GCGCATGCGCATCGACCCGGATGTAGGTTGCGTGCGA
AGTACGTGCGCCGCGCATGCGCATCGACCCGGATGTA
GGACCCGGATGTAGATGACACAGCAATATGACACAGC
AATCCCCGGAAGTGACCCCCGGAAGTGACGTTGCGTG
CGAAGGGACTTTCCGTACGTGCGCCGCGCATGCGCATC
GGGACTTTCCATTGCATCAGGGACTTTCCAAGCTTGGT
CAACGTCGA
>Super_Promoter

GAAGCTCGAGTCGATGGTACCCCCCGGAAGTGACaaG
CCGCGCATGCGCATCaaGACCCGGATGTAGaaGTACGT
GCaaGAAATGGGTGGTTCCGGaaGCCGCGCATGCGCAT
CaaATGACACAGCAATaaGCCGCGCATGCGCATCaaGTT
GCGTGCGAAaaATGACACAGCAATaaCCCCGGAAGTGA
CaaATTGCATCAaaGAAATGGGTGGTTCCGGaaGTACG
TGCaaGACCCGGATGTAGaaGTACGTGCaaGGGACTTT
CCaaATTGCATCAaaGGGACTTTCCaaGGGACTTTCCaaA
TGACACAGCAATaaCCCCGGAAGTGACaaGTTGCGTGC
GAAaaGAAATGGGTGGTTCCGGaaCCCCGGAAGTGACa
aGGGACTTTCCaaGAAATGGGTGGTTCCGGaaGGGACT
TTCCaaATGACACAGCAATaaGCCGCGCATGCGCATCaa
GGGACTTTCCaaGACCCGGATGTAGaaGACCCGGATGT
AGaaGTTGCGTGCGAAaaGTACGTGCaaCCCCGGAAGT
GACaaGTTGCGTGCGAAaaGCCGCGCATGCGCATCaaC
CCCGGAAGTGACAAGCTTGGTCAACGTCGA

>Super_Promoter_without_Spacers
GAAGCTCGAGTCGATGGTACCCCCCGGAAGTGACGCC
GCGCATGCGCATCGACCCGGATGTAGGTACGTGCGAA
ATGGGTGGTTCCGGGCCGCGCATGCGCATCATGACAC
AGCAATGCCGCGCATGCGCATCGTTGCGTGCGAAATG
ACACAGCAATCCCCGGAAGTGACATTGCATCAGAAATG
GGTGGTTCCGGGTACGTGCGACCCGGATGTAGGTACG
TGCGGGACTTTCCATTGCATCAGGGACTTTCCGGGACT
TTCCATGACACAGCAATCCCCGGAAGTGACGTTGCGTG
CGAAGAAATGGGTGGTTCCGGCCCCGGAAGTGACGGG
ACTTTCCGAAATGGGTGGTTCCGGGGGACTTTCCATGA
CACAGCAATGCCGCGCATGCGCATCGGGACTTTCCGAC
CCGGATGTAGGACCCGGATGTAGGTTGCGTGCGAAGT
ACGTGCCCCCGGAAGTGACGTTGCGTGCGAAGCCGCG
CATGCGCATCCCCCGGAAGTGACAAGCTTGGTCAACGT
CGA

A.6 Heterotypic Promoters Tested in Library 3

>5' homeotypic NFE2l2 on 100bp Promoter
TAtgctGAGTcat□ATAtgctGAGTcat□ATAtgctGAGTcat□ATAtgctGAGTcat□ATA
tgctGAGTcat□ATAtgctGAGTcat□AGGGACTTTCCaaGTACGTGCaaATGACACAGC
AATaaGACCCGGATGTAGaaCCCCGGAAGTGACaaGTTGCGTGCGAAaaGCCGCGCA
TGCGCATCaaATTGCATCA
>Balanced_Promoter(Weak Sequences Replaced with new ones
GGGACTTTCCaaTGCTGAGTCATaaCCCCGGAAGTGACaaATGACACAGCAATaaGG
GACTTTCCaaGGGACTTTCCaaATGACACAGCAATaaGGGTCACGTGaaGGGACTTTC
CaaATGACACAGCAATaaCCCCGGAAGTGACaaCCCCGGAAGTGACaaGACCCGGAT
GTAGaaGGGCGGGGGCGGGGaaATGACACAGCAATaaGGGACTTTCCaaGACCCGG
ATGTAGaaGTACGTGCaaGACCCGGATGTAGaaGACCCGGATGTAGaaCCCCGGAAG
TGACaaGGGACTTTCCaaGTACGTGCaaTGCTGAGTCATaaCCCCGGAAGTGACaaG
GGCGGGGGCGGGGaaGTACGTGCaaGGGTCACGTGaaGGGTCACGTGaaCCCCGGA
AGTGACaaGGGTCACGTGaaGGGTCACGTGaaTGCTGAGTCATaaTGCTGAGTCAT□
aaGTACGTGC
>Even_Promoter
GAAATGGGTGGTTCCGGTAATTGCATCATAGTACGTGCTACGCCATTTTTAGGGACT
TTCCTAtgctGAGTcat□ACCCCGGAAGTGACTAGAAATGGGTGGTTCCGGTAGACCC
GGATGTAGTAtgctGAGTcat□ACGCCATTTTTAATGACACAGCAATTAGAAATGGGT
GGTTCCGGTAtgctGAGTcat□AGTACGTGCTACGCCATTTTTAGACCCGGATGTAGT
AGGGACTTTCCTAGTTGCGTGCGAATAATTGCATCATAGACCCGGATGTAGTAGTTG
CGTGCGAATAGGGACTTTCCTACCCCGGAAGTGACTAGCCGCGCATGCGCATCTAAT
GACACAGCAATTAGTACGTGCTAATGACACAGCAATTACCCCGGAAGTGACTAATTG
CATCATAGCCGCGCATGCGCATCTAGCCGCGCATGCGCATCTAGTTGCGTGCGAA

>Even_Promoter wo spacer
GAAATGGGTGGTTCCGGATTGCATCAGTACGTGCCGCCATTTTGGGACTTTCCtgctG
AGTcat□CCCGGAAGTGACGAAATGGGTGGTTCCGGGACCCGGATGTAGtgctGAGT
cat□GCCATTTTATGACACAGCAATGAAATGGGTGGTTCCGGtgctGAGTcat□GTACG
TGCCGCCATTTTGACCCGGATGTAGGGGACTTTCCGTTGCGTGCGAAATTGCATCAG
ACCCGGATGTAGGTTGCGTGCGAAGGGACTTTCCCCCCGGAAGTGACGCCGCGCAT
GCGCATCATGACACAGCAATGTACGTGCATGACACAGCAATCCCCGGAAGTGACATT
GCATCAGCCGCGCATGCGCATCGCCGCGCATGCGCATCGTTGCGTGCGAA

>New_Sequence_Promoter_200bp
CCCGGAAGTGACTATATGCTGAGTCATTATAGGGCGGGGGCGGGGTATACCCGGAA
GTGACTATAGGGTCACGTGTATATGCTGAGTCATTATAGGGCGGGGGCGGGGTATA
CCCGGAAGTGACTATACCGGAAGTTATATGCTGAGTCATTATACCGGAAGTTATATG
CTGAGTCATTATACCCGGAAGTGACTATAGGGTCACGTG
>New_Sequence_Promoter_300bp
CCGGAAGTTATAGGGTCACGTGTATAGGGCGGGGGCGGGGTATACCCGGAAGTGA
CTATACCGGAAGTTATAGGGTCACGTGTATAGGGCGGGGGCGGGGTATACCCGGAA
GTGACTATACCCGGAAGTGACTATAGGGTCACGTGTATACCGGAAGTTATAGGGTCA
CGTGTATAGGGCGGGGGCGGGGTATATGCTGAGTCATTATACCCGGAAGTGACTAT
AGGGCGGGGGCGGGGTATATGCTGAGTCATTATATGCTGAGTCATTATACCCGGAA
GTGACTATATGCTGAGTCATTATACCCGGAAGTGACTATATGCTGAGTCATTATACCG
GAAGTTATATGCTGAGTCAT
>New_Sequence_Promoter_Max

TGCTGAGTCATTATACCGGAAGTTATACCCGGAAGTGACTATAGGGCGGGGGCGGG
GTATAGGGTCACGTGTATATGCTGAGTCATTATAGGGTCACGTGTATACCCGGAAGT
GACTATACCGGAAGTTATACCGGAAGTTATATGCTGAGTCATTATATGCTGAGTCATT
ATACCCGGAAGTGACTATACCCGGAAGTGACTATAGGGTCACGTGTATACCGGAAGT
TATAGGGTCACGTGTATAGGGCGGGGGCGGGGTATATGCTGAGTCATTATACCGGA
AGTTATAGGGCGGGGGCGGGGTATAGGGTCACGTGTATAGGGCGGGGGCGGGGT
ATATGCTGAGTCATTATACCCGGAAGTGACTATACCCGGAAGTGACTATAGGGCGGG
GGCGGGGTATATGCTGAGTCATTATACCCGGAAGTGACTATAGGGCGGGGGCGGG
GTATACCGGAAGTTATAGGGTCACGTG
>NFE2l2 Absent Promoter
GTTGCGTGCGAATACCCCGGAAGTGACTAGCCGCGCATGCGCATCTAGCCGCGCAT
GCGCATCTAGGGACTTTCCTACCCCGGAAGTGACTACCCCGGAAGTGACTAGGGACT
TTCCTAGGGACTTTCCTAGAAATGGGTGGTTCCGGTAATTGCATCATAGACCCGGAT
GTAGTAGTTGCGTGCGAATACGCCATTTTTAATGACACAGCAATTAGAAATGGGTGG
TTCCGGATCCCCGGAAGTGACTAGACCCGGATGTAGTAGTACGTGCTAATGACACAG
CAATTAGGGACTTTCCTAATTGCATCATAGTACGTGC

>NFEL2_3' Weighted_Promoter
tgctGAGTcatTAGCCGCGCATGCGCATCTAtgctGAGTcatTAtgctGAGTcatTAATT
GCATCATAATTGCATCATAGCCGCGCATGCGCATCTAtgctGAGTcatTAGGGACTTTC
CTAGAAATGGGTGGTTCCGGTAGACCCGGATGTAGTAATGACACAGCAATTAATGAC
ACAGCAATTAGGGACTTTCCTAGAAATGGGTGGTTCCGGTAGACCCGGATGTAGTAC
CCCGGAAGTGACTAGTTGCGTGCGAATAGGGACTTTCCTAGTACGTGCTACCCCGGA
AGTGACTAGGGACTTTCCTAGTACGTGCTACCCCGGAAGTGACTACGCCATTTTTAGT
TGCGTGCGAATACCCCGGAAGTGAC
>NFEL2_3' Weighted_Promoter without spacer
tgctGAGTcatGCCGCGCATGCGCATCtgctGAGTcatTgctGAGTcatATTGCATCAAT
TGCATCAGCCGCGCATGCGCATCtgctGAGTcatGGGACTTTCCGAAATGGGTGGTTC
CGGGACCCGGATGTAGATGACACAGCAATATGACACAGCAATGGGACTTTCCGAAA
TGGGTGGTTCCGGGACCCGGATGTAGCCCCGGAAGTGACGTTGCGTGCGAAGGGAC
TTTCCGTACGTGCCCCCGGAAGTGACGGGACTTTCCGTACGTGCCCCCGGAAGTGAC
CGCCATTTTGTTGCGTGCGAACCCCGGAAGTGAC
>NFkB Absent Promoter
GCCGCGCATGCGCATCtaGTACGTGCtaGACCCGGATGTAGtaGACCCGGATGTAGtat
gctGAGTcatTatgctGAGTcatTaATTGCATCAtatgctGAGTcatTaCCCCGGAAGTGAC
taCCCCGGAAGTGACtaGCCGCGCATGCGCATCtaCCCCGGAAGTGACtaATGACACA
GCAATtaGTACGTGCtaGTTGCGTGCGAAtaCGCCATTTTtaATGACACAGCAATtaGAA
ATGGGTGGTTCCGGtaGTTGCGTGCGAAtaATTGCATCAtaGAAATGGGTGGTTCCGG
tatgctGAGTcatTaCCCCGGAAGTGAC
>Reduced_Promoter_Without_Spacer (Weak Sequences Replced with new
ones)
CCCCGGAAGTGACGGGTCACGTGGACCCGGATGTAGTGCTGAGTCATGGTACGTGC
GGGTCACGTGGACCCGGATGTAGGACCCGGATGTAGATGACACAGCAATATGACAC
AGCAATCCCCGGAAGTGACCCCCGGAAGTGACTGCTGAGTCATGGGGACTTTCCGTA
CGTGCGGGTCACGTGGGGACTTTCCGGGCGGGGGCGGGGGGGACTTTCC

A.7 Bidirectional Promoter Sequences

>BD_1 Bidirectional_Using_NFkB_as_Spacer_reverse_Complement

ATGACACAGCAATGGAAAGTCCCGAAATGGGTGGTTCCGGGGAAAGTCCCtgct
GAGTcat□GGAAAGTCCCATTGCATCAGGAAAGTCCCGACCCGGATGTAGGGAA
AGTCCCGTACGTGCGGAAAGTCCCGCCGCGCATGCGCATCGGAAAGTCCCGTTG
CGTGCGAAGGAAAGTCCCCCCCGGAAGTGACGGAAAGTCCCtgctGAGTcat□GG
AAAGTCCCCGCCATTTTGGAAAGTCCCCCCCGGAAGTGACGGAAAGTCCCCCCC
GGAAGTGACGGAAAGTCCCtgctGAGTcat□

>BD_11 U8 Bidirectional Promoter
TCTAGAATGACACAGCAATGATGCGCATGCGCGGCGTTGCGTGCGAATTCGCAC
GCAACGTTGCGTGCGAAGTCACTTCCGGGGATTGCATCACTACATCCGGGTCGT
ACGTGCATTGCTGTGTCATATTGCATCAGCACGTACttGGAAAGTCCCATTGCATC
ATGATGCAATAGGGGCGGGGTGATGCGCATGCGCGGCGTACGTGCTTCGCACG
CAACAGGGGCGGGGTGTCACTTCCGGGGATGACACAGCAATCTACATCCGGGT
CGTACGTGCATTGCTGTGTCATAGGGGCGGGGTGCACGTACAGGGGCGGGGTG
GAAAGTCCCATGACACAGCAATGGAAAGTCCCGTACGTGCTGATGCAATATGAC
ACAGCAATGATGCGCATGCGCGGCATGACACAGCAATTTCGCACGCAACGTACG
TGCCTACATCCGGGTCGTTGCGTGCGAAATTGCTGTGTCATAGGGGCGGGGTGC
ACGTACATTGCATCAGGAAAGTCCCGGTACCGTTGCGTGCGAAaaATTGCATCAa
aGGGACTTTCCaaGACCCGGATGTAGaaCCCCGGAAGTGACaaGACCCGGATGTA
GaaGACCCGGATGTAGaaCCCCGGAAGTGACaaGGGACTTTCCaaGGGACTTTCC
aaGGGACTTTCCaaGACCCGGATGTAGaaGGGACTTTCCaaCCCCGGAAGTGACa
aCCCCGGAAGTGACaaCCCCGGAAGTGACaaGGGACTTTCCaaCCCCGGAAGTGA
CaaGACCCGGATGTAGaaGGGACTTTCCaaCCCCGGAAGTGACAAGCTT

>BD_2 Bidirectional_Using NFkB as spacer complement

ATGACACAGCAATCCCTGAAAGGGAAATGGGTGGTTCCGGCCCTGAAAGGtgct
GAGTcat□CCTGAAAGGATTGCATCACCCTGAAAGGGACCCGGATGTAGCCCTG
AAAGGGTACGTGCCCCTGAAAGGGCCGCGCATGCGCATCCCCTGAAAGGGTTG
CGTGCGAACCCTGAAAGGCCCCGGAAGTGACCCCTGAAAGGtgctGAGTcat□C
CTGAAAGGCGCCATTTTCCCTGAAAGGCCCCGGAAGTGACCCCTGAAAGGCCCC
GGAAGTGACCCCTGAAAGGtgctGAGTcat□

>BD_3 Bidirectional_Balanced_Promoter

GGGACTTTCCaaGTTGCGTGCGAAaaCCCCGGAAGTGACaaATGACACAGCAATa
aGGGACTTTCCaaGGGACTTTCCaaATGACACAGCAATaaGCCGCGCATGCGCAT
CaaGGGACTTTCCaaATGACACAGCAATaaCCCCGGAAGTGACaaCCCCGGAAGT
GACaaGACCCGGATGTAGaaATTGCATCAaaATGACACAGCAATaaGGGACTTTC
CaaGACCCGGATGTAGaaGTACGTGCaaGACCCGGATGTAGaaGACCCGGATGTA
GaaCCCCGGAAGTGACaaGGGACTTTCCaaGTACGTGCaaGTTGCGTGCGAAaaC
CCCGGAAGTGACaaATTGCATCAaaGTACGTGCaaGCCGCGCATGCGCATCaaGC
CGCGCATGCGCATCaaCCCCGGAAGTGACaaGCCGCGCATGCGCATCaaGCCGCG
CATGCGCATCaaGTTGCGTGCGAAaaGTTGCGTGCGAAaaGTACGTGC

>BD_4 Bidirectional_Super_Prom_W/O_spacer
CCCCGGAAGTGACGCCGCGCATGCGCATCGACCCGGATGTAGGTACGTGCGAA
ATGGGTGGTTCCGGGCCGCGCATGCGCATCATGACACAGCAATGCCGCGCATGC
GCATCGTTGCGTGCGAAATGACACAGCAATCCCCGGAAGTGACATTGCATCAGA
AATGGGTGGTTCCGGGTACGTGCGACCCGGATGTAGGTACGTGCGGGACTTTCC
ATTGCATCAGGGACTTTCCGGGACTTTCCATGACACAGCAATCCCCGGAAGTGA
CGTTGCGTGCGAAGAAATGGGTGGTTCCGGCCCCGGAAGTGACGGGACTTTCC
GAAATGGGTGGTTCCGGGGGACTTTCCATGACACAGCAATGCCGCGCATGCGCA
TCGGGACTTTCCGACCCGGATGTAGGACCCGGATGTAGGTTGCGTGCGAAGTAC
GTGCCCCCGGAAGTGACGTTGCGTGCGAAGCCGCGCATGCGCATCCCCCGGAA
GTGAC

>BD_5 One_Block_OF_Each_TFRE_Bi
GCCGCGCATGCGCATCGACCCGGATGTAGCGCCATTTTGGGACTTTCCATGACA
CAGCAATGTACGTGCGAAATGGGTGGTTCCGGGGGGAGGGGGCGGGGtgctGA
GTcat□gcgtgGGCGCCCCGGAAGTGACGTTGCGTGCGAAATTGCATCAccaccagg
gggcgc

>BD_6 NFKB_NFE2l2_GABPA_Skewed_Promoter
GGGACTTTCCtgctGAGTcat□ATGACACAGCAATtgcgtgGGCGtgctGAGTcat□ca
ccaggggcgcCCCCGGAAGTGACATTGCATCACCCCGGAAGTGACGTACGTGCCC
CCGGAAGTGACCCCCGGAAGTGACGGGACTTTCCGTTGCGTGCGAAGCCGCGC
ATGCGCATCtgctGAGTcat□GGGGAGGGGGCGGGGGGGACTTTCCGACCCGGAT
GTAGGAAATGGGTGGTTCCGGCGCCATTTTGGGACTTTCCtgctGAGTcat□

>BD_7 Increased_SP1_Promoter

GACCCGGATGTAGGGGACTTTCCGGGGAGGGGGCGGGGccaccaggggcgcGGG
GAGGGGGCGGGGATGACACAGCAATGGGACTTTCCtgctGAGTcatGAAATGGG
TGGTTCCGGtgcgtgGGCGGGGACTTTCCtgctGAGTcatGGCCATTTTATTGCATC
AtgctGAGTcatCCCCGGAAGTGACGTACGTGCtgctGAGTcatCCCCGGAAGTG
ACGGGACTTTCCGTTGCGTGCGAAGCCGCGCATGCGCATCCCCCGGAAGTGACG
GGGAGGGGGCGGGGCCCCGGAAGTGAC

>BD_8 Increased_SP1_and_CTCF_Promoter_Bi
tgctGAGTcatGGGGAGGGGGCGGGGGCCGCGCATGCGCATCGTTGCGTGCGA
AGACCCGGATGTAGccaccaggggcgcGGGACTTTCCCCCCGGAAGTGACGGGAC
TTTCCGGGGAGGGGGCGGGGATGACACAGCAATccaccaggggcgctgcgtgGGCG
ccaccaggggcgcCCCCGGAAGTGACGGGACTTTCCATTGCATCAtgctGAGTcatG
GGGAGGGGGCGGGGCCCCGGAAGTGACGGGACTTTCCtgctGAGTcatGGCCAT
TTTtgctGAGTcatGAAATGGGTGGTTCCGGCCCCGGAAGTGACGTACGTGC

>BD_9 Complicated_Promoter_With_elevated_everything

CGCCATTTTGCCGCGCATGCGCATCCCCCGGAAGTGACGGGACTTTCCtgcgtgGG
CGCGCCATTTTATTGCATCACCCCGGAAGTGACGTACGTGCGACCCGGATGTAG
ATGACACAGCAATccaccaggggcgcCGCCATTTTATGACACAGCAATGTTGCGTG
CGAAGTACGTGCtgcgtgGGCGGGGGAGGGGGCGGGGGAAATGGGTGGTTCCG
GGCCGCGCATGCGCATCATTGCATCAtgctGAGTcatGTTGCGTGCGAAGAAATG
GGTGGTTCCGGtgctGAGTcatGGGGAGGGGGCGGGGGGGACTTTCCGGGGAG
GGGGCGGGGGAAATGGGTGGTTCCGGGCCGCGCATGCGCATCGACCCGGATG
TAGtgctGAGTcatGTTGCGTGCGAACCCCGGAAGTGACccaccaggggcgctgctG
AGTcatGGGACTTTCCtgcgtgGGCGATGACACAGCAATccaccaggggcgcGACCC
GGATGTAGATTGCATCAGGGACTTTCCCCCCGGAAGTGACGTACGTGC

A.8 Diagrams of Promoters

Heterotypic Library 1 Design Diagrams

_ → A SINGLE UNDERSCORE INDICATES NO SPACER

____ → Indicates spacers

> Heterotypic High Transcription Promoter

D*esign: Random assortment of high and weak blocks*

DMP1___GABP beta___AhR/ARNT___NFkB___Sp1___ZBED4___NFkB___ARE___GABP beta___HRE___ZBED4___DMP1___ARE___AhR/ARNT___NFkB___GABP beta___AhR/ARNT___HRE___AARE___DMP1___AARE___DMP1___AhR/ARNT___GABP beta___NFkB___GABP beta___HRE___AARE___Sp1___HRE

> Heterotypic High Transcription Promoters without stability

D*esign: Random assortment of high and weak blocks with stability element removed*

DMP1___GABP beta___AhR/ARNT___NFkB___NFkB___ARE___GABP beta___HRE___DMP1___ARE___AhR/ARNT___NFkB___GABP beta___AhR/ARNT___HRE___AARE___DMP1___AARE___DMP1___AhR/ARNT___GABP beta___NFkB___GABP beta___HRE___AARE___HRE

> Promoter With 5' repressor

D*esign: Using stacks of repressors on the 5' end*

YY1___YY1___YY1___DMP1___GABP beta___AhR/ARNT___NFkB___NFkB___ARE___GABP beta___HRE___DMP1___ARE___AhR/ARNT___NFkB___GABP beta___AhR/ARNT___HRE___AARE___DMP1___AARE___DMP1___AhR/ARNT___GABP beta___NFkB___GABP beta___HRE___AARE___HRE

> Promoter With 3' repressor

D*esign: Using stacks of repressors on the 3' end*

DMP1___GABP beta___AhR/ARNT___NFkB___NFkB___ARE___GABP beta___HRE___DMP1___ARE___AhR/ARNT___NFkB___GABP beta___AhR/ARNT___HRE___AARE___DMP1___AARE___DMP1___AhR/ARNT___GABP beta___NFkB___GABP beta___HRE___AARE___HRE___YY1___YY1___YY1

> Really High TFRE

D*esign: Making a promoter with increased amount of building blocks.*

DMP1___ ARE___ AARE___ AhR/ARNT___ Sp1___ GABP beta___ GABP beta___ GABP beta___ NFkB___ DMP1___ AARE___ AhR/ARNT___ HRE___ AARE___ Sp1___ AARE___ ARE___ NFkB___ AhR/ARNT___ GABP beta___ AhR/ARNT___ HRE___ NFkB___ DMP1___ DMP1___ HRE___ GABP beta___ ARE___ GABP beta___ GABP beta___ NFkB___ NFkB___ AARE___ Sp1___ NFkB___ AhR/ARNT___ HRE___ Sp1___ NFkB___ ARE___ Sp1___ HRE___ DMP1___ ARE

> Remixed High TFRE

D*esign: Remixed version of really high TFRE.*

HRE___ ARE___ GABP beta___ AARE___ HRE___ AARE___ NFkB___ DMP1___ NFkB___ DMP1___ ARE___ GABP beta___ Sp1___ NFkB___ GABP beta___ DMP1___ Sp1___ HRE___ GABP beta___ AhR/ARNT___ AhR/ARNT___ NFkB___ Sp1___ NFkB___ GABP beta___ AARE___ NFkB___ AhR/ARNT___ DMP1___ GABP beta___ Sp1___ HRE___ Sp1___ ARE___ ARE___ DMP1___ AARE___ AhR/ARNT___ ARE___ AARE___ NFkB___ GABP beta___ HRE___ AhR/ARNT

> 5’ weighted promoter

D*esign: Weighing transcriptionally strong TFREs to the 5’ end of the DNA.*

NFkB___ DMP1___ GABP beta___ DMP1___ DMP1___ GABP beta___ NFkB___ NFkB___ NFkB___ DMP1___ NFkB___ GABP beta___ GABP beta___ GABP beta___ NFkB___ GABP beta___ DMP1___ NFkB___ GABP beta___ AhR/ARNT___ ARE___ AhR/ARNT___ AhR/ARNT___ AARE___ HRE___ AARE___ AARE___ Sp1___ HRE___ Sp1___ ARE___ HRE___ Sp1___ Sp1___ ARE___ HRE___ ARE___ ARE___ HRE___ AhR/ARNT___ Sp1___ AARE___ AhR/ARNT___ AARE

> 3’ weighted promoter

D*esign: Weighing transcriptionally strong TFREs to the 3’ end of the DNA.*

AhR/ARNT___ ARE___ AhR/ARNT___ AhR/ARNT___ AARE___ HRE___ AARE___ AARE___ Sp1___ HRE___ Sp1___ ARE___ HRE___ Sp1___ Sp1___ ARE___ HRE___ ARE___ ARE___ HRE___ AhR/ARNT___ Sp1___ AARE___ AhR/ARNT___ AARE___ NFkB___ DMP1___ GABP beta___ DMP1___ DMP1___ GABP beta___ NFkB___ NFkB___ NFkB___ DMP1___ NFkB___ GABP beta___ GABP beta___ GABP beta___ NFkB___ GABP beta___ DMP1___ NFkB___ GABP beta

> Heterotypic High Transcription Promoters Complement

D*esign: Copy of Heterotypic High Transcription Promoter using the Complement of TFRE sequence*

DMP1___GABP beta___AhR/ARNT___NFkB___Sp1___ZBED4___NFkB___ARE___GABP beta___HRE___ZBED4___DMP1___ARE___AhR/ARNT___NFkB___GABP beta___AhR/ARNT___HRE___AARE___DMP1___AARE___DMP1___AhR/ARNT___GABP beta___NFkB___GABP beta___HRE___AARE___Sp1___HRE

> Heterotypic High Transcription Promoters Reverse Complement

D*esign: Copy of Heterotypic High Transcription Promoter using the Reverse Complement of TFRE sequence*

HRE___Sp1___AARE___HRE___beta GABP___NFkB___beta GABP___ARNT/AhR___DMP1___AARE___DMP1___AARE___HRE___ARNT/AhR___beta GABP___NFkB___ARNT/AhR___ARE___DMP1___ZBED4___HRE___beta GABP___ARE___NFkB___ZBED4___Sp1___NFkB___ARNT/AhR___beta GABP___DMP1

Heterotypic Library 2 Design Diagrams

_ → A SINGLE UNDERSCORE INDICATES NO SPACER

____ → Indicates spacers

> Balanced Promoter

D*esign: Promoter without excessive amounts of any TFRE and mixed randomly*

NFkB____ AhR/ARNT____ GABP beta ____ ARE____ NFkB____ NFkB____ ARE____ NRF1____ NFkB____ ARE____ GABP beta ____ GABP beta ____ DMP1____ AARE____ ARE____ NFkB____ DMP1____ HRE____ DMP1____ DMP1____ GABP beta ____ NFkB____ HRE____ AhR/ARNT____ GABP beta ____ AARE____ HRE____ NRF1____ NRF1____ GABP beta ____ NRF1____ NRF1____ AhR/ARNT____ AhR/ARNT____ HRE_

> Balanced Promoter without spacer

D*esign: Promoter without excessive amounts of any TFRE and mixed randomly without spacer*

NFkB_AhR/ARNT_GABP beta _ARE_NFkB_NFkB_ARE_NRF1_NFkB_ARE_GABP beta _GABP beta _DMP1_AARE_ARE_NFkB_DMP1_HRE_DMP1_DMP1_GABP beta _NFkB_HRE_AhR/ARNT_GABP beta _AARE_HRE_NRF1_NRF1_GABP beta _NRF1_NRF1_AhR/ARNT_AhR/ARNT_HRE_

> 3' Weighted Promoter with 5' Repressor

D*esign: Trying 3' weighting with 5' repression*

YY1____ YY1____ YY1____ AhR/ARNT____ HRE____ HRE____ HRE____ AARE____ AhR/ARNT____ AhR/ARNT____ AARE____ HRE____ AhR/ARNT____ DMP1____ GABP beta ____ ARE____ DMP1____ NFkB____ NFkB____ NFkB____ DMP1____ GABP beta ____ DMP1____ NFkB____ NFkB____ GABP beta ____ ARE____ GABP beta ____ NRF1____ NRF1____ NRF1____ NRF1____ GABP beta ____ NFkB____ NRF1____ ARE____ GABP beta ____ ARE_

> Reduced Promoter

D*esign: Balanced Promoter but with less TFREs*

GABP beta ____ NRF1____ DMP1____ AhR/ARNT____ HRE____ NRF1____ DMP1____ DMP1____ ARE____ ARE____ GABP beta ____ GABP beta ____ AhR/ARNT____ NFkB____ HRE____ NRF1____ NFkB____ AARE____ NFkB_

> Reduced Promoter_without_Spacer

D*esign: Balanced Promoter but with less TFREs without spacer*

GABP beta _NRF1_DMP1_AhR/ARNT_HRE_NRF1_DMP1_DMP1_ARE_ARE_GABP beta _GABP beta _AhR/ARNT_NFkB_HRE_NRF1_NFkB_AARE_NFkB_

> 100bp Promoter single copies

D*esign: Testing promoter lengths*

NFkB____ HRE____ ARE____ DMP1____ GABP beta ____ AhR/ARNT____ NRF1____ AARE_

> 200bp Promoter two copies

D*esign: Testing promoter lengths*

NFkB____ HRE____ ARE____ DMP1____ GABP beta ____ AhR/ARNT____ NRF1____ AARE____
NFkB____ HRE____ ARE____ DMP1____ GABP beta ____ AhR/ARNT____ NRF1____ AARE_

> 300bp Promoter

D*esign: Testing promoter lengths*

NFkB____ HRE____ ARE____ DMP1____ GABP beta ____ AhR/ARNT____ NRF1____ AARE____
NFkB____ HRE____ ARE____ DMP1____ GABP beta ____ AhR/ARNT____ NRF1____ AARE____
NFkB____ HRE____ ARE____ DMP1____ GABP beta ____ AhR/ARNT____ NRF1____ AARE_

> Super_Promoter

D*esign: Enhanced heterotypic promoter which contains GC box also.*

GABP beta ____ NRF1____ DMP1____ HRE____ GC Box____ NRF1____ ARE____ NRF1____
AhR/ARNT____ ARE____ GABP beta ____ AARE____ GC Box____ HRE____ DMP1____ HRE____
NFkB____ AARE____ NFkB____ NFkB____ ARE____ GABP beta ____ AhR/ARNT____ GC Box____
GABP beta ____ NFkB____ GC Box____ NFkB____ ARE____ NRF1____ NFkB____ DMP1____
DMP1____ AhR/ARNT____ HRE____ GABP beta ____ AhR/ARNT____ NRF1____ GABP beta _

> Super_Promoter_without_Spacers

D*esign: Enhanced heterotypic promoter which contains GC box also.*

GABP beta _NRF1_DMP1_HRE_GC Box_NRF1_ARE_NRF1_AhR/ARNT_ARE_GABP beta _AARE_GC Box_HRE_DMP1_HRE_NFkB_AARE_NFkB_NFkB_ARE_GABP beta _AhR/ARNT_GC Box_GABP beta _NFkB_GC Box_NFkB_ARE_NRF1_NFkB_DMP1_DMP1_AhR/ARNT_HRE_GABP beta _AhR/ARNT_NRF1_GABP beta _

Heterotypic Library 3 Design Diagrams

_ → A SINGLE UNDERSCORE INDICATES NO SPACER

____ → Indicates spacers

> NFEL2_3' Weighted_Promoter

D*esign: Creating a 3' weighted promoter that uses NFE2l2.*

NFe2l2____ NRF1____ NFe2l2____ NFe2l2____ AARE____ AARE____ NRF1____ NFe2l2____ NFkB____ GC Box____ DMP1____ ARE____ ARE____ NFkB____ GC Box____ DMP1____ GABP beta____ V$AHRARNT.03____ NFkB____ HRE____ GABP beta____ NFkB____ HRE____ GABP beta____ YY1____ V$AHRARNT.03____ GABP beta_

> Even Promoter

D*esign: Everything in equal amounts for promoter composition*

GC Box____ AARE____ HRE____ YY1____ NFkB____ NFe2l2____ GABP beta____ GC Box____ DMP1____ NFe2l2____ YY1____ ARE____ GC Box____ NFe2l2____ HRE____ YY1____ DMP1____ NFkB____ V$AHRARNT.03____ AARE____ DMP1____ V$AHRARNT.03____ NFkB____ GABP beta____ NRF1____ ARE____ HRE____ ARE____ GABP beta____ AARE____ NRF1____ NRF1____ V$AHRARNT.03_

> NFkB absent Promoter

D*esign: A promoter with NFE2l2 instead of NFkB*

NRF1____ HRE____ DMP1____ DMP1____ NFe2l2____ NFe2l2____ AARE____ NFe2l2____ GABP beta____ GABP beta____ NRF1____ GABP beta____ ARE____ HRE____ V$AHRARNT.03____ YY1____ ARE____ GC Box____ V$AHRARNT.03____ AARE____ GC Box____ NFe2l2____ GABP beta_

> NFE2l2 absent Promoter

D*esign: A promoter with NFkB instead of NFE2l2*

V$AHRARNT.03____ GABP beta____ NRF1____ NRF1____ NFkB____ GABP beta____ GABP beta____ NFkB____ NFkB____ GC Box____ AARE____ DMP1____ V$AHRARNT.03____ YY1____ ARE____ GC Box__GABP beta____ DMP1____ HRE____ ARE____ NFkB____ AARE____ HRE_

> NFEL2_3' Weighted_Promoter without spacer

D*esign: A promoter with NFE2l2 weighting on the 5' end and strong TFRE weighting on the 3' end*

NFe2l2_NRF1_NFe2l2_NFe2l2_AARE_AARE_NRF1_NFe2l2_NFkB_GC Box_DMP1_ARE_ARE_NFkB_GC Box_DMP1_GABP beta_V$AHRARNT.03_NFkB_HRE_GABP beta_NFkB_HRE_GABP beta_YY1_V$AHRARNT.03_GABP beta_

> 5' homeotypic NFE2l2 on 100bp Promoter

D*esign: The 100bp promoter with the homotypic promoter attached on the 5' end.*

___ NFe2l2____ ___ NFe2l2____ ___ NFe2l2____ ___ NFe2l2____ ___ NFe2l2____ ___ NFe2l2____ NFkB_AAHRE_AAARE_AADMP1_AAGABP beta_AAV$AHRARNT.03_AANRF1_AAAARE_

> Even Promoter without spacer

D*esign: The even promoter with spacers removed*

GC Box_AARE_HRE_YY1_NFkB_NFe2l2_GABP beta_GC Box_DMP1_NFe2l2_YY1_ARE_GC Box_NFe2l2_HRE_YY1_DMP1_NFkB_V$AHRARNT.03_AARE_DMP1_V$AHRARNT.03_NFkB_GABP beta_NRF1_ARE_HRE_ARE_GABP beta_AARE_NRF1_NRF1_V$AHRARNT.03_

> Balanced_Promoter(Weak Sequences Replaced with new ones

D*esign: Attempt to improve the balanced promoter sequence from library 2*

NFkB____ NFe2l2____ GABP beta____ ARE____ NFkB____ NFkB____ ARE____ ClockBMAL____ NFkB____ ARE____ GABP beta____ GABP beta____ DMP1____ EGR1____ ARE____ NFkB____ DMP1____ HRE____ DMP1____ DMP1____ GABP beta____ NFkB____ HRE____ NFe2l2____ GABP beta____ EGR1____ HRE____ ClockBMAL____ ClockBMAL____ GABP beta____ ClockBMAL____ ClockBMAL____ NFe2l2____ NFe2l2____ HRE_

> Reduced Promoter without spacer (weak sequences replaced with new)

D*esign: Attempted to improve reduced promoter without spacer from library 2*

GABP beta_ClockBMAL_DMP1_NFe2l2_HRE_ClockBMAL_DMP1_DMP1_ARE_ARE_GABP beta_GABP beta_NFe2l2_NFkB_HRE_ClockBMAL_NFkB_EGR1_NFkB_

> New_Sequence_Promoter_200bp

D*esign: 200bp promoter completely made of new sequences.*

v$gabpb1.01___ NFE2l2___ EGR1___ v$gabpb1.01___ ClockBMAL____ NFE2l2___ EGR1___ v$gabpb1.01___ V$ETV4.01___ NFE2l2___ V$ETV4.01___ NFE2l2___ v$gabpb1.01___ ClockBMAL_

> New_Sequence_Promoter_300bp

D*esign: 300bp promoter completely made of new sequences.*

V$ETV4.01___ ClockBMAL____ EGR1___ v$gabpb1.01___ V$ETV4.01___ ClockBMAL____ EGR1___ v$gabpb1.01___ v$gabpb1.01___ ClockBMAL____ V$ETV4.01___ ClockBMAL____ EGR1___ NFE2l2___ v$gabpb1.01___ EGR1___ NFE2l2___ NFE2l2___ v$gabpb1.01___ NFE2l2___ v$gabpb1.01___ NFE2l2___ V$ETV4.01___ NFE2l2

> New_Sequence_Promoter_Max

D*esign: 400bp promoter completely made of new sequences.*

NFE2l2___ V$ETV4.01___ v$gabpb1.01___ EGR1___ ClockBMAL____ NFE2l2___ ClockBMAL____ v$gabpb1.01___ V$ETV4.01___ V$ETV4.01___ NFE2l2___ NFE2l2___ v$gabpb1.01___ v$gabpb1.01___ ClockBMAL____ V$ETV4.01___ ClockBMAL____ EGR1___ NFE2l2___ V$ETV4.01___ EGR1___ ClockBMAL____ EGR1___ NFE2l2___ v$gabpb1.01___ v$gabpb1.01___ EGR1___ NFE2l2___ v$gabpb1.01___ EGR1___ V$ETV4.01___ ClockBMAL_

Bidirectional Promoter Designs

_ → A SINGLE UNDERSCORE INDICATES NO SPACER

____ → Indicates spacers

> Bidirectional_Using_NFkB_as_Spacer_reverse_Complement

D*esign: Bidirectional Promoter with reverse complement NFkB as spacer*

ARE_NFkBRC_ GC Box_NFkBRC_ NFE2L2_NFkBRC_ ARE_NFkBRC_ V$DMTF_NFkBRC_ HRE_NFkBRC_ NRF1_NFkBRC_ V$AHRR_NFkBRC_ V$ETSF_NFkBRC_ NFE2L2_NFkBRC_ YY1_NFkBRC_ V$ETSF_NFkBRC_ V$ETSF_NFkBRC_ NFE2L2

> Bidirectional_Using_NFkB_as_Spacer _Complement

D*esign: Bidirectional Promoter with complement NFkB as spacer*

ARE_NFkBC_ GC Box_NFkBC_ NFE2L2_NFkBC_ ARE_NFkBC_ V$DMTF_NFkBC_ HRE_NFkBC_ NRF1_NFkBC_ V$AHRR_NFkBC_ V$ETSF_NFkBC_ NFE2L2_NFkBC_ YY1_NFkBC_ V$ETSF_NFkBC_ V$ETSF_NFkBC_ NFE2L2

> Bidirectional_Balanced_Promoter

D*esign: Balanced promoter in bidirectional context*

V$NFKB___V$AHRR___V$ETSF___ARE___V$NFKB___V$NFKB___ARE___NRF1___V$NFKB___ARE__ _V$ETSF___V$ETSF___V$DMTF___ARE___ARE___V$NFKB___V$DMTF___HRE___V$DMTF___V$DM TF___V$ETSF___V$NFKB___HRE___V$AHRR___V$ETSF___ARE___HRE___NRF1___NRF1___V$ETSF ___NRF1___NRF1___V$AHRR___V$AHRR___HRE

> Bidirectional_Super_Prom_W/O_spacer

D*esign: Super Prom w/0 spacer in bidirectional context*

V$ETSF_NRF1_V$DMTF_HRE_GC Box_NRF1_ARE_NRF1_V$AHRR_ARE_V$ETSF_ARE _GC Box_HRE_V$DMTF_HRE_V$NFKB_ARE_V$NFKB_V$NFKB_ARE_V$ETSF_ V$AHRR_ GC Box_V$ETSF_V$NFKB_GC Box_ V$NFKB_ ARE_ NRF1_ V$NFKB_ V$DMTF_ V$DMTF_ V$AHRR_ HRE_ V$ETSF_ V$AHRR_ NRF1_V$ETSF_

> One_Block_OF_Each_TFRE

D*esign: Trying out bioinformatic blocks*

NRF1_V$DMTF_YY1_V$NFKB_ARE_HRE_GC Box_SP1_NFE2L2_V$EGR1.01_V$ETSF_V$AHRR_ARE_V$CTCF.04_

> NFKB_NFE2l2_GABPA_Skewed_Promoter

D*esign: Trying out bioinformatic blocks in different numbers*

V$NFKB_ NFE2L2_ ARE_ V$EGR1.01_ NFE2L2_ V$CTCF.04_ V$ETSF_ ARE_ V$ETSF_ HRE_ V$ETSF_ V$ETSF_ V$NFKB_ V$AHRR_ NRF1_ NFE2L2_ SP1_ V$NFKB_ V$DMTF_ GC Box_ YY1_ V$NFKB_ NFE2L2_

> Increased_SP1_Promoter

D*esign:* Increased_SP1_Promoter

V$DMTF_ V$NFKB_ SP1_ V$CTCF.04_ SP1_ ARE_ V$NFKB_ NFE2L2_ GC Box_ V$EGR1.01_ V$NFKB_ NFE2L2_ YY1_ ARE_ NFE2L2_ V$ETSF_ HRE_ NFE2L2_ V$ETSF_ V$NFKB_ V$AHRR_ NRF1_ V$ETSF_ SP1_ V$ETSF_

> Increased_SP1_and_CTCF_Promoter

D*esign:* Increased_SP1_and_CTCF_Promoter

NFE2L2_ SP1_ NRF1_ V$AHRR_ V$DMTF_ V$CTCF.04_ V$NFKB_ V$ETSF_ V$NFKB_ SP1_ ARE_ V$CTCF.04_ V$EGR1.01_ V$CTCF.04_ V$ETSF_ V$NFKB_ ARE_ NFE2L2_ SP1_ V$ETSF_ V$NFKB_ NFE2L2_ YY1_ NFE2L2_ GC Box_ V$ETSF_ HRE_

> Complicated_Promoter_With_elevated_everything

D*esign:* Complicated_Promoter_With_elevated_everything

YY1_ NRF1_ V$ETSF_ V$NFKB_ V$EGR1.01_ YY1_ ARE_ V$ETSF_ HRE_ V$DMTF_ ARE_ V$CTCF.04_ YY1_ ARE_ V$AHRR_ HRE_ V$EGR1.01_ SP1_ GC Box_ NRF1_ ARE_ NFE2L2_ V$AHRR_ GC Box_ NFE2L2_ SP1_ V$NFKB_ SP1_ GC Box_ NRF1_ V$DMTF_ NFE2L2_ V$AHRR_ V$ETSF_ V$CTCF.04_ NFE2L2_ V$NFKB_ V$EGR1.01_ ARE_ V$CTCF.04_ V$DMTF_ ARE_ V$NFKB_ V$ETSF_ HRE_

Appendix B

Code Used Throughout the Project

B.1 Bash Script Used for RNA-seq Alignment

```
#!/bin/bash
# Request 16 gigabytes of real memory (RAM)
#$ -l rmem=30G
# Request 4 cores in an OpenMP environment
#$ -pe openmp 4
# Email notifications to abourke1@sheffield.ac.uk
#$ -M abourke1@sheffield.ac.uk
# Email notifications if the job aborts
#$ -m abe

# Set the OPENMP_NUM_THREADS environment variable to 4
export OMP_NUM_THREADS=4
# Run the program foo with input foo.dat
# and output foo.res

#To extract splice sites from the gtf file you use HISAT. "
    hisat2_extract_splice_sites.py genes.gtf > splicesites.txt,
     where hisat2_extract_splice_sites.py is included in the
    HISAT2 package, genes.gtf is a gene annotation file, and
    splicesites.txt is a list of splice sites with which you
    provide HISAT2 in this mode. Note that it is better to use
    indexes built using annotated transcripts (such as
    genome_tran or genome_snp_tran), which works better than
    using this option. It has no effect to provide splice sites
     that are already included in the indexes."

#Running FastQC to generate a report on my reads

module load apps/java

cd /home/fcr18ab/FastQC/
./fastqc /fastdata/fcr18ab/Adrian\ RNA-seq\ Data/Loop1/*.gz -o
    /data/fcr18ab/FastQC_Reports/Adrian/run1/

./fastqc /fastdata/fcr18ab/Adrian\ RNA-seq\ Data/Loop2/*.gz -o
    /data/fcr18ab/FastQC_Reports/Adrian/run2/


#Trimming my reads to ensure they no longer have adapter
    contamination
```

```
cd /home/fcr18ab/Trimmomatic-0.39
for i in {1..27}
do echo Cutting Sample ${i}
java -jar trimmomatic-0.39.jar PE -phred33 -threads 4 \
/fastdata/fcr18ab/Adrian_RNA-seq_Data/Loop1/${i}.fastq.gz /
   fastdata/fcr18ab/Adrian_RNA-seq_Data/Loop2/${i}.fastq.gz \
/fastdata/fcr18ab/Adrian_RNA-seq_Data/TLoop1/p${i}.fastq.gz /
   fastdata/fcr18ab/Adrian_RNA-seq_Data/ULoop1/${i}.fastq.gz /
   fastdata/fcr18ab/Adrian_RNA-seq_Data/TLoop2/p${i}.fastq.gz
   /fastdata/fcr18ab/Adrian_RNA-seq_Data/ULoop2/${i}.fastq.gz
   \
ILLUMINACLIP:/home/fcr18ab/Trimmomatic-0.39/adapters/TruSeq3-
   SE.fa:2:30:10 SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3
   MINLEN:36
done

#Doing another quality control

cd /home/fcr18ab/FastQC/
./fastqc /fastdata/fcr18ab/Adrian_RNA-seq_Data/TLoop1/*.gz -o
   /data/fcr18ab/FastQC_Reports/Adrian/trun1/

./fastqc /fastdata/fcr18ab/Adrian_RNA-seq_Data/TLoop2/*.gz -o
   /data/fcr18ab/FastQC_Reports/Adrian/trun2/

#There is no need to create a genome for this anymore. I have
   already made it.
cd /home/fcr18ab/STAR-master/bin/Linux_x86_64
./STAR --runThreadN 4 \
--runMode genomeGenerate \
--genomeDir /fastdata/fcr18ab/Adrian_star_index \
--sjdbOverhang 149 \
--genomeFastaFiles /data/fcr18ab/Refrence/New_Refrence_Files/
   Cricetulus_griseus_picr.CriGri-PICR.dna_sm.nonchromosomal.
   fa /data/fcr18ab/Refrence/New_genes.fa \
--sjdbGTFfile /data/fcr18ab/Refrence/New_Refrence_Files/1
   Cricetulus_griseus_picr.CriGri-PICR.98.gtf \
--sjdbFileChrStartEnd /data/fcr18ab/Refrence/
   New_Refrence_Files/splice_sites.txt


#Aligning the genome
cd /home/fcr18ab/STAR-master/bin/Linux_x86_64
```

```
for i in {1..27}
do echo Processing Sample ${i}
./STAR --runThreadN 4 \
--twopassMode Basic \
--readFilesCommand zcat \
--readFilesIn /fastdata/fcr18ab/Adrian_RNA-seq_Data/TLoop1/p${
    i}.fastq.gz /fastdata/fcr18ab/Adrian_RNA-seq_Data/TLoop2/p$
    {i}.fastq.gz \
--outFileNamePrefix /fastdata/fcr18ab/A_aligned_reads/${i} \
--sjdbGTFfile /data/fcr18ab/Refrence/New_Refrence_Files/1
    Cricetulus_griseus_picr.CriGri-PICR.98.gtf \
--sjdbFileChrStartEnd /data/fcr18ab/Refrence/
    New_Refrence_Files/splice_sites.txt \
--outSAMtype BAM SortedByCoordinate \
--genomeDir /fastdata/fcr18ab/Adrian_star_index/ \
--outSAMunmapped Within \
--outSAMattributes Standard \
--quantMode TranscriptomeSAM GeneCounts \
--sjdbOverhang 149
done

#Counting my files
cd /home/fcr18ab/subread-1.6.4-source/bin/
./featureCounts -s 0 -t gene -g gene_id -T 4 -p \
-a /data/fcr18ab/Refrence/New_Refrence_Files/1
    Cricetulus_griseus_picr.CriGri-PICR.98.gtf \
-o /data/fcr18ab/A_Count/Star.counts \
/fastdata/fcr18ab/A_aligned_reads/1Aligned.sortedByCoord.out.
    bam \
/fastdata/fcr18ab/A_aligned_reads/2Aligned.sortedByCoord.out.
    bam \
/fastdata/fcr18ab/A_aligned_reads/3Aligned.sortedByCoord.out.
    bam \
/fastdata/fcr18ab/A_aligned_reads/4Aligned.sortedByCoord.out.
    bam \
/fastdata/fcr18ab/A_aligned_reads/5Aligned.sortedByCoord.out.
    bam \
/fastdata/fcr18ab/A_aligned_reads/6Aligned.sortedByCoord.out.
    bam \
/fastdata/fcr18ab/A_aligned_reads/7Aligned.sortedByCoord.out.
    bam \
/fastdata/fcr18ab/A_aligned_reads/8Aligned.sortedByCoord.out.
    bam \
```

```
/fastdata/fcr18ab/A_aligned_reads/9Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/10Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/11Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/12Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/13Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/14Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/15Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/16Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/17Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/18Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/19Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/20Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/21Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/22Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/23Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/24Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/25Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/26Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/27Aligned.sortedByCoord.out.
   bam

./featureCounts -s 0 -t exon -g gene_id -T 4 -p -M \
-a /data/fcr18ab/Refrence/New_Refrence_Files/1
   Cricetulus_griseus_picr.CriGri-PICR.98.gtf \
-o /data/fcr18ab/A_Count/MStar.counts \
```

```
/fastdata/fcr18ab/A_aligned_reads/1Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/2Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/3Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/4Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/5Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/6Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/7Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/8Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/9Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/10Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/11Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/12Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/13Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/14Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/15Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/16Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/17Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/18Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/19Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/20Aligned.sortedByCoord.out.
   bam \
/fastdata/fcr18ab/A_aligned_reads/21Aligned.sortedByCoord.out.
   bam \
```

B.2 Count and Quality Control R Code

```
---
title: "My_RNA_Seq_analysis"
author: "Adrian Bourke"
date: "29/11/2019"
output:
  html_document: default
  pdf_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_knit$set(root.dir="F:/Adrian RNA-seq Data/
   RNA_Count/Trying the antibody/A_Count/All Counts/")
```



```{r, include=FALSE}
#Packages Required
library(DESeq2)
library("dplyr")
library(tidyr)
```

```{r DeSeq2 Setup}
#Importing my data from the FeatureCounts Files
setwd("F:/Adrian RNA-seq Data/RNA_Count/Trying the antibody/
   A_Count/")
data <- read.table("Star.counts",header=T)
coldata <- read.csv("coldata.txt.csv")

#Altering the dataset for my analysis
samples <- data[,-c(2:6)]
names(samples)<- c("GeneID","sample1","sample2","sample3","
   sample4","sample5","sample6","sample7","sample8","sample9
   ","sample10","sample11","sample12","sample13","sample14","
   sample15","sample16",
                   "sample17","sample18","sample19","sample20
                      ","sample21","sample22","sample23","
                      sample24","sample25","sample26","
                      sample27")
coldata <- coldata[,-1]
```
```

```
rownames(samples) <-samples[,1]
samples = samples[,-c(1)]

#Creating the differential data set (DE has not been preformed
    yet)
dds <- DESeqDataSetFromMatrix(countData = samples,colData =
   coldata,design = ~Day)
colData(dds)
```

```{r Calcuation of FPKM}
#Changing my working directory to where I want my files to go
setwd("F:/Adrian RNA-seq Data/RNA_Count/Trying the antibody/
   A_Count/All Counts/")

#Getting the Raw Counts from the data set used for this
   analysis
foo <-counts(dds,normalized = FALSE)
write.csv(foo, file="Counts_Non_Normalised_Before_DE_nofltr.
   csv")

#Merging lenght column with the data for TPM calculation
mcols(dds)$basepairs <- as.numeric(data[,6])

#Using DESeq2 to calculate the FPKM for me. These files arent
   filtered
res <-DESeq(dds)
foo2 <- fpkm(res)
write.csv(foo2, file = "FPKM_for_samples_nofltr.csv")

#Checking and visualizing my library size.
colSums(samples)

#This command checks how many reads there are for a sample. i.
   e. the sum
#sum(assay(dds)[,1])
#colSums(assay(dds))
```

```{r Barchart of Millions of reads per sample}
#Creating agraph of my read counts. I need to change the axsis
    on this
library(ggplot2)
```

```
d <-colSums(assay(dds))
d <- as.table(d)
d<- as.data.frame(d)
d$z <- (d[,2]/1000000)
colnames(d) <- c("Samples","Millions_of_Reads","z")
g<-ggplot(d,aes(x = Samples, y = z, fill = Samples))
g+geom_col()+labs(x = "Sample", y = "Millions of Reads")
```

```{r Histogram}
#Filtering my non-expressed Genes
is_expressed <- assay(dds) >= 5
head(is_expressed)
sum(is_expressed[,1])

#Creating a graph of genes expressed
hist(rowSums(is_expressed),
     main="Number Of Samples a Gene is Expresssed In",xlab="
        Sample Count")
```

```{r Count Distribution Analysis}
#Keeping genes that are present in at least 9 samples with a
   greater frequency of 10
keep <- rowSums(assay(dds) >=5) >=6
table(keep)
dds <- dds[keep,]

#Visualising count distributions. This is usally achieved
   through boplots.
boxplot(assay(dds))
boxplot(log10(assay(dds)))

#To remove library size dependencys one must use the vst or
   rlog function
#to compensate for the library sizes and put data on log2
   scale
#The aim of this is to remove the dependence on the varience
   on the mean
vsd <- vst(dds,blind=TRUE)

#With this i have used avgTXlenght from assay dds to correct
   for the library size
```

```
#The next step is to once again use box plots to check for
    count distribution
boxplot(assay(vsd), xlab = "", ylab ="Log2 counts per million
    ", las=2,main="Normalised Distributions")
#Addring a horizontal line that corresponds to median logcpm
abline(h=median(assay(vsd)),col="blue")
```

```{r Sample HeatMap}
#What this tells us is the sample to sample distances
sampleDists <- dist(t(assay(vsd)))
library("RColorBrewer")
library("pheatmap")
sampleDistMatrix <- as.matrix(sampleDists)
rownames(sampleDistMatrix) <- paste(colData(dds)$Day,colData(
    dds)$Clone,sep="-")
colnames(sampleDistMatrix) <- colData(dds)$X
colors <- colorRampPalette(rev(brewer.pal(9,"Blues")) )(255)
pheatmap(sampleDistMatrix,col=colors)
```

```{r Sample_PCA_Plot}
#Preforming principle component analysis
pca <- c("Clone","Day")
plotPCA(vsd,intgroup = c("Day","Clone"),ntop=50000,)
plot_data <-plotPCA(vsd,intgroup = c("Day","Clone"),ntop
    =50000,returnData=T)
ggplot(plot_data,aes(x=PC1,y=PC2,col=Clone,pch=Day,label=name)
    ) +geom_point() + geom_text(alpha=1)
```

```{r TPM_Setup}
#Getting the data required to get TPM values
tpmdata <- assay(dds)
tpmdata <- as.data.frame(tpmdata)
tpmdata <- tibble::rownames_to_column(tpmdata, "Gene ID")
colnames(data) <- c("Gene ID","2","3","4","5","Lenght")
df2 <- data[,c(1,6)]
tpmdata <- merge(tpmdata,df2,by="Gene ID")
```

```{r TPM_Calculation}

#Function for the calculation of TPM

tpm <- function(counts, lengths) {
  rate <- counts / lengths
  rate / sum(rate) * 1e6
}

#Putting the data into a data table for use of the equation
ftr.cnt <- read.table("F:/Adrian RNA-seq Data/RNA_Count/Trying
    the antibody/A_Count/Star.counts", sep="\t",
   stringsAsFactors=FALSE,
                         header=TRUE)

names(ftr.cnt)<- c("GeneID","Chr","Start","End","Strand","
   Length","sample1","sample2","sample3","sample4","sample5","
   sample6","sample7","sample8","sample9","sample10","sample11
   ","sample12","sample13","sample14","sample15","sample16",
                    "sample17","sample18","sample19","sample20
                       ","sample21","sample22","sample23","
                       sample24","sample25","sample26","
                       sample27")
#Formula for the calculation of tpm
ftr.tpm <- ftr.cnt %>%
  gather(sample, cnt, 7:ncol(ftr.cnt)) %>%
  group_by(sample) %>%
  mutate(tpm=tpm(cnt, Length)) %>%
  select(-cnt) %>%
  spread(sample, tpm)



#Manipulating the table into a nice format and editing it.

write.csv(ftr.tpm, "TPM_Values.csv")
```


```{r Ceating a file with annotated TPM}
#Carrying in the files with annotation of genes
annotation <- read.csv("C:/Users/Adrian/Google Drive/Code with
    specific functions/RNA-seq Coding/Stuff I need for
   enrichment analysis/New_Alignment_Stuff/mart_export (3).txt
```

```
    ")
ENTZID <- read.csv("C:/Users/Adrian/Google Drive/Code with
    specific functions/RNA-seq Coding/Stuff I need for
    enrichment analysis/New_Alignment_Stuff/mart_export (5).txt
    ")
annotation <- merge(annotation,ENTZID,by="Gene.stable.ID")

##########################Got my csv created of all my
    differentially expressed genes##########################
TPM = ftr.tpm[,-c(2:6)]
names(TPM)[1] <-"ENSEMBL"
names(annotation)[1] <- "ENSEMBL"
Results_comb2 <- merge(annotation,TPM,by="ENSEMBL",all=T)
Results_comb2 <- Results_comb2[!duplicated(
    Results_comb2$ENSEMBL),]
write.csv(Results_comb2,"Annotated_tpm.csv")

#Creating a fpkm annotated to crsos check
FPKM <- read.csv("F:/Adrian RNA-seq Data/RNA_Count/Trying the
    antibody/A_Count/All Counts/FPKM_for_samples_nofltr.csv")
names(FPKM)<- c("ENSEMBL","sample1","sample2","sample3","
    sample4","sample5","sample6","sample7","sample8","sample9
    ","sample10","sample11","sample12","sample13","sample14","
    sample15","sample16",
                "sample17","sample18","sample19","sample20","
                   sample21","sample22","sample23","sample24
                   ","sample25","sample26","sample27")
#Formula for the calculation of tpm

Results_comb2 <- merge(annotation,FPKM,by="ENSEMBL",all=T)
Results_comb2 <- Results_comb2[!duplicated(
    Results_comb2$ENSEMBL),]
write.csv(Results_comb2,"FPKM.csv")
```

B.3 Differential Expression Code using DESEQ2

```
---
title: "DESeq2_Analysis_Producer_Vs_no"
author: "Adrian Bourke"
date: "29/11/2019"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(root.dir = "F:/Adrian RNA-seq Data/
   RNA_Count/Files with Antibody Counts in/A_Count/Clonal
   Comparision/Producer_Vs_Non/")
knitr::knit_meta(class=NULL, clean = TRUE)
```

```{r Packages Needed, echo=FALSE, message=FALSE}
#Packages Required
library(DESeq2)
library("dplyr")
library(tidyr)
```

```{r DESEQ2 Setup, warning=FALSE}
#Setting my Working Directory
setwd("F:/Adrian RNA-seq Data/RNA_Count/Files with Antibody
   Counts in/A_Count/")
#Importing my data from the FeatureCounts Files
data <- read.csv("Star.csv",header=T)
coldata <- read.csv("coldata.txt.csv")
samples <- data[,-c(2:6)]
names(samples)<- c("GeneID","sample1","sample2","sample3","
   sample4","sample5","sample6","sample7","sample8","sample9
   ","sample10","sample11","sample12","sample13","sample14","
   sample15","sample16",
                   "sample17","sample18","sample19","sample20
                      ","sample21","sample22","sample23","
                      sample24","sample25","sample26","
                      sample27")
rownames(samples) <-samples[,1]
samples <- samples[,-c(1)]
coldata <- coldata
```
```

```
#Setting up the Design Conditions
dds <- DESeqDataSetFromMatrix(countData = samples,colData =
    coldata,design = ~Producer)

#Checking the imported Design
colData(dds)
```

```{r Filtering and Setting Up Counts}
#Filtering the data for amount of genes per sample
keep <- rowSums(assay(dds) >=5) >=6
table(keep)
dds <- dds[keep,]
```

```{r Setting up DESeq2, message=T, warning=T}
#Looking at my factor levels and deciding what the refrence
    level is going to be.
dds$Producer <- relevel(dds$Producer,"No")

#Carrying out the Differential Gene Expression Analysis
dds <- DESeq(dds)
res <- results(dds)
res

#Writing the results to a csv file
write.csv(res,"F:/Adrian RNA-seq Data/RNA_Count/Files with
    Antibody Counts in/A_Count/Clonal Comparision/
    Producer_Vs_Non/Producer_Vs_Non.csv")

#To better visualise the data a log format is usally used. To
    find the object we want to shrink use:
resultsNames(dds)
```

```{r Creating results ordered and graphs of normalization}
#Putting the comparision that you want to analyse into the log
     format. I can only use apeglm if I dont need stat.
resLFC <- lfcShrink(dds, coef = "Producer_Yes_vs_No",type='
    apeglm')
resLFC1 <- lfcShrink(dds, coef = "Producer_Yes_vs_No")
resLFC
```

```
#Ordering my data based on its p values
resordered <- res[order(res$pvalue),]
summary(res)

#How many pvalues are less than 0.1
sum(res$padj <0.05,na.rm=TRUE)

#Plotting the results of my differential gene expression
   analysis#########################
plotMA(res,ylim=c(-10,10))
plotMA(resLFC,ylim=c(-10,10))

#Plotting the dispersion estimate to give an idea of how well
   the DEseq2 model fits
plotDispEsts( dds, ylim = c(1e-6, 1e1) )
hist( res$pvalue, breaks=20, col="grey" )

#Continuing the analysis to further the data needed
library(tibble)
results.ordered <- as.data.frame(resLFC) %>%
  rownames_to_column("ENSEMBL") %>%
  arrange(padj)

head(results.ordered)

write.csv (results.ordered,"F:/Adrian RNA-seq Data/RNA_Count/
   Files with Antibody Counts in/A_Count/Clonal Comparision/
   Producer_Vs_Non/DEProducer Vs Non_apgelm.csv")
```

```{r Volacano Plot}
#Plotting a volcano plot to look at my data#####################
plot(results.ordered$log2FoldChange,-log10(results.
   ordered$padj),
      pch=16,
      xlab="Log Fold Change",
      ylab="Signifigance")
degenes <- which(results.ordered$padj <0.05)
points(results.ordered$log2FoldChange[degenes],-log10(results.
   ordered$padj)[degenes],
        col="red",
        pch=16)
```

```
```{r Creating Annotated Differential Expression, echo=FALSE,
    message= FALSE}

#Potential to graph the fold change of certain genes to check
    the changes
plotCounts(dds,"ENSCGRG00015001096",intgroup = c("Producer"))

#Annotating the Results to continue with the analysis.
annotation <- read.csv("C:/Users/Adrian/Google Drive/
    Bioinformatic Data/Code with specific functions/RNA-seq
    Coding/Stuff I need for enrichment analysis/
    New_Alignment_Stuff/mart_export (3).txt")
ENTZID <- read.csv("C:/Users/Adrian/Google Drive/Bioinformatic
     Data/Code with specific functions/RNA-seq Coding/Stuff I
    need for enrichment analysis/New_Alignment_Stuff/
    mart_export (5).txt")

#Merging the two files together and naming the columns
annotation <- merge(annotation,ENTZID,by="Gene.stable.ID")
colnames(annotation) <- c("ENSEMBL","CHO K1","CHO CriGri","
    Human","Gene Name","Description","ENTREZID")

#The annotation data set contains duplicates. The first step
    in the process is to check how many duplicates it may have
library(tidyverse)
x <- annotation[,1]
duplicated(x)

#This command extracts all the duplicated elements
x[duplicated(x)]

#removal of the duplicated elements.#
filtered_annotation <- annotation[!duplicated(
    annotation$ENSEMBL),]
check <- annotation[!duplicated(annotation$ENSEMBL),]

#Putting together my files
results.annotated <- merge(results.ordered,filtered_annotation
    ,by="ENSEMBL",)
results.annotated<- as.data.frame(results.annotated) %>%
  arrange(padj)
write.csv(results.annotated,"F:/Adrian RNA-seq Data/RNA_Count/
```
```

```
   Files with Antibody Counts in/A_Count/Clonal Comparision/
   Producer_Vs_Non/DEProducer Vs Non_Annotated.csv")
```

```{r Creating a heatmap of the data}
#Analysis has been carried out and labbelled time to visualise
    the data
library(pheatmap)
topgenes <- results.annotated$ENSEMBL[1:10]
vsd <- vst(dds)
pheatmap(assay(vsd)[topgenes,])

#Adding more information to the heat map
sampleinfo <- as.data.frame(colData(dds)[,c("Producer","Clone
   ")])
pheatmap(assay(vsd)[topgenes,],
         annotation_col = sampleinfo)
```

```{r Sample and gene clusters}
#we need to  manually cluster the gene samples for this method
mate <- assay(vsd)[topgenes,]

#Calculate the distance between matrix samples.
d_samples <- dist(t(mate))
plot(hclust(d_samples))
rect.hclust(hclust(d_samples),k=2)

#We can cut the denogram to give a set number of clusters.
   Each cluster can be given a label of 1 or 2.
clusters <- cutree(hclust(d_samples),k=2)
clusters

#These groupings could then be tabulated against the samples
   metadata to see if particular biological groups are
   associated with the new clusters we have identified.
table(clusters,colData(dds)$Producer)

#A similar approach as this would work fo genes. Instead of
   transposing before calculating the distance matrix we jsut
   leave
#it as it is. For a small nuber of genes a heatmap such as a
   distance matrix can be computed.
```

```
#It may be computationally expensive for a large number of
   genes.
d_genes <- dist(mate)
plot(hclust(d_genes))
rect.hclust(hclust(d_genes),k=2)
cutree(hclust(d_genes),k=2)
```

```{r Creating interactive diagrams}
setwd("F:/Adrian RNA-seq Data/RNA_Count/Files with Antibody
   Counts in/A_Count/Clonal Comparision/Producer_Vs_Non/")
#Creating interactive diagrams of the plots
dds.mf <- dds
design(dds.mf) <- ~Producer
de.mf <- DESeq(dds.mf)
resultsNames(de.mf)
results <- results(de.mf,contrast=c("Producer","Yes","No"))
results

#Creating an online interactive graph that allows you to look
   at the DE of each gene
#Creating an online interactive graph that allows you to look
   at the DE of each gene
library("Glimma")
results <- as.data.frame(results.ordered)
results$log10MeannormCount <- log10(results$baseMean)
idx <- rowSums(counts(dds)) > 0
results <- results[idx,]
results$padj[is.na(results$padj)] <- 1
glMDPlot(results,
         xval="log10MeannormCount",
         yval="log2FoldChange",
         counts = counts(dds)[idx,],
         anno = data.frame(GeneID=rownames(dds)[idx]),
         groups = dds$Producer,
         samples = colnames(dds))
```

```{r Visualizing the Data with ggplot}
library(ggplot2)
plot_data <- plotCounts(dds,"ENSCGRG00015020631",intgroup=c("
   Producer","Clone"),returnData = T)
plot_data
```

```
ggplot(plot_data,aes(x=Producer,y=log2(count)))+geom_point()+
    facet_wrap(~Producer)
df <- results.annotated
ggplot(df,aes(x=log2(baseMean),y=log2FoldChange))+geom_point()
ggplot(df,aes(x=log2(baseMean),y=log2FoldChange,col=padj<0.05)
    )+geom_point()
ggplot(df,aes(x=log2(baseMean),y=log2FoldChange)) + geom_point
    (aes(color=padj<0.05),alpha=0.4) + scale_colour_manual(
    values=c("black","red"))
```

```{r Creating data files needed for FGSEA}
#CHO dosn't have any entrz ids or that so that I have to use
    mouse yet.
data1 <- results.annotated[,-c(12)]
data2 <- read.csv("C:/Users/Adrian/Google Drive/Bioinformatic
    Data/Code with specific functions/RNA-seq Coding/Stuff I
    need for enrichment analysis/ENTZ_id_for_cho.txt")
data3 <- read.csv("C:/Users/Adrian/Google Drive/Bioinformatic
    Data/Code with specific functions/RNA-seq Coding/Stuff I
    need for enrichment analysis/New_Alignment_Stuff/
    mart_export (6).txt")
data2 <- merge(data2,data3,by="Gene.stable.ID")

#Renaming the columns of merge data sets and
colnames(data2) <- c("Mouse ENSEMBL","ENTREZID","ENSEMBL")
results.annotated2 <- merge(data1,data2,by="ENSEMBL")
results.annotated2 <- results.annotated2[!duplicated(results.
    annotated2$ENSEMBL),]
```

```{r FGSEA Analysis Setup}
library(fgsea)
gseaInput <-filter(results.annotated2,!is.na(ENTREZID))
gseaInput <-gseaInput[!duplicated(gseaInput$ENTREZID),]
ranks <- -log10(gseaInput$padj)* gseaInput$log2FoldChange
names(ranks) <- gseaInput$ENTREZID
barplot(sort(ranks,decreasing = T))
```

```{r, FGSEA Pathway analysis}
load("C:/Users/Adrian/Google Drive/Bioinformatic Data/Code
    with specific functions/RNA-seq Coding/Stuff I need for
```

```
    enrichment analysis/mouse_H_v5p2.rdata")
#Insert the file that you want to analyze against above and
    below here.
pathways <- Mm.H
fgseaRes <- fgsea(pathways,ranks,nperm=1000)
dim(fgseaRes)
head(fgseaRes)

####Creating a table that gives the name of each pathway
    tested and the stats from doing the test#####
fgseaResTidy <- fgseaRes %>%
  as_tibble() %>%
  arrange(desc(NES))

#Show the results in a nice table.
fgseaResTidy
fgseaResTidy <- as.data.frame(fgseaResTidy)

#Creating a ggplot of the table above
ggplot(fgseaResTidy,aes(reorder(pathway,NES),NES)) +
  geom_col(aes(fill=pval<0.05)) +
  coord_flip() +
  labs(x = "Pathway", y = "normalized Enrichment Score",
       title = "Hallmark pathways NES from GSEA")

#The enrichment plot will show where the genes belong to in a
    particular gene set are towards the bottom or top of the
    gene list and how it was calculated.
#plotEnrichment(pathways[["HALLMARK_UNFOLDED_PROTEIN_RESPONSE
    "]],ranks)

#Create a table that shows the results for multiple pathsways.
topup <- fgseaRes %>% filter (ES>0) %>% top_n(40,wt=-padj)
topDown <- fgseaRes %>%
  filter(ES < 0) %>%
  top_n(10, wt=-padj)
topPathways <- bind_rows(topup, topDown) %>%
  arrange(-ES)
#plotGseaTable(pathways[topPathways$pathway],
#              ranks,
#              fgseaRes,
#              gseaParam = 0.5)
topPathways[sapply(topPathways,is.list)] <- apply(topPathways[
```

```
    sapply(topPathways, is.list)],
                                                    1,function(x
                                                      )
                                                     paste(
                                                        unlist(
                                                        x),
                                                            sep
                                                               =",

                                                               ",
                                                               collapse

                                                              =

                                                               ",

                                                               ")
                                                               )

#Writeing the data set to a file
write.csv(topPathways, file="F:/Adrian RNA-seq Data/RNA_Count/
    Files with Antibody Counts in/A_Count/Clonal Comparision/
    Producer_Vs_Non/FGSEA_all enriched pathways.csv")
```` ``` ````

```{r Extracting genes involed in particular Pathways}
######################Extracting the names involved in particular
     pathways###################
my_genes <- filter(results.annotated2,ENTREZID %in% pathways
    [["HALLMARK_UNFOLDED_PROTEIN_RESPONSE"]])%>%
   pull(ENSEMBL)
vsd <- vst(dds)
mat <- assay(vsd)[my_genes,]
mat <- mat - rowMeans(mat)
dim(mat)
pheatmap(mat,
          annotation_col = sampleinfo[,c("Producer","Producer")
             ])
```

```{r GOSEQ Analysis}
#########################GO SEQ ANALYSIS
    ##############################
```

```
#GoSeq is a method to study gene oncology analysis for rna-seq
    . It trys to account for
#gene lenght bias in detection of over representation

#From the GOseq vignette:
#GOseq first needs to quantify the length bias present in the
    dataset under consideration.
#This is done by calculating a Probability Weighting Function
    or PWF which can be thought of as a function which gives
    the probability that a gene will be differentially
    expressed (DE), based on its length alone.
#The PWF is calculated by fitting a monotonic spline to the
    binary data series of differential expression (1=DE, 0=not
    DE) as a function of gene length.
#The PWF is used to weight the chance of selecting each gene
    when forming a null distribution for GO category membership
    .
#The fact that the PWF is calculated directly from the dataset
     under consideration makes this approach robust, only
    correcting for the length bias present in the data.
#This whole segment is related to fitting the probability
    weight function to my genes

library(goseq)
isgene <- results.annotated2[!duplicated(results.annotated2$`
    Mouse ENSEMBL`),]
namesy <- results.annotated2[!duplicated(results.annotated2$`
    Mouse ENSEMBL`),]
isgene <- isgene$padj <0.05 & !is.na(isgene$padj)
genes <- as.integer(isgene)
names(genes)<- namesy$`Mouse ENSEMBL`
newdata <- data[,c(1,6)]

#Nameing the new coloumns
colnames(newdata) = c("ENSEMBL","GENELenght")
filtereddata <- annotation[!duplicated(annotation$ENSEMBL),]

#Merging the data
newdata <- merge(newdata,filtereddata,by="ENSEMBL")
newdata <- newdata[,1:2]
results.annotated3 <- merge(namesy,newdata,by="ENSEMBL")
pwf <- nullp(genes, "mm10","ensGene",bias.data = results.
    annotated3$GENELenght)
```

```
#####################Conducting gene enrichment analysis######
goResults <- goseq(pwf, "mm10","ensGene")
goResults %>%
  top_n(10, wt=-over_represented_pvalue) %>%
  mutate(hitsPerc=numDEInCat*100/numInCat) %>%
  ggplot(aes(x=hitsPerc,
             y=category,
             colour=over_represented_pvalue,
             size=numDEInCat)) +
  geom_point() +
  expand_limits(x=0) +
  labs(x="Hits (%)", y="GO term", colour="p value", size="
     Count")

write.csv(goResults,"F:/Adrian RNA-seq Data/RNA_Count/Files
   with Antibody Counts in/A_Count/Clonal Comparision/
   Producer_Vs_Non/goseq Results.csv")
```

```{r ClusterProfiler}
##Analysis with clusterprofiler#####
library(clusterProfiler)
universe <- results.annotated3 %>%
  pull(`Mouse ENSEMBL`)
sigGenes <- results.annotated3 %>%
  filter(padj < 0.05, !is.na(`Mouse ENSEMBL`)) %>% pull(`Mouse
     ENSEMBL`)

enrich_go <- enrichGO(
  gene= sigGenes,
  OrgDb = org.Mm.eg.db,
  keyType = "ENSEMBL",
  ont = "ALL",
  universe = universe,
  qvalueCutoff = 0.05,
  readable=TRUE
)

enrich_go1 <- enrichGO(
  gene= sigGenes,
  OrgDb = org.Mm.eg.db,
```

```
  keyType = "ENSEMBL",
  ont = "BP",
  universe = universe,
  qvalueCutoff = 0.05,
  readable=TRUE
)

enrich_go_tidy <- enrich_go %>%
  slot("result") %>%
  tibble::as.tibble()
enrich_go_tidy

#Writeing the file to a csv
write.csv(enrich_go_tidy,"F:/Adrian RNA-seq Data/RNA_Count/
   Files with Antibody Counts in/A_Count/Clonal Comparision/
   Producer_Vs_Non/cluster_profiler_analysis.csv")



############Creating a dot plot from the data that was obtained
    from the cluster profiler ###########
dotplot(enrich_go)

emapplot(enrich_go)
emapplot(enrich_go1)

#Trying a barplot to see if I can better represent the data.
barplot(enrich_go)
```

```{r KEGG Analysis}
sigGenes2 <- results.annotated3 %>%
  filter(padj < 0.05, !is.na(ENTREZID)) %>% pull(ENTREZID)
search_kegg_organism('hamster',by='common_name')
keg_res <- enrichKEGG(gene=sigGenes2,organism="mmu")
head(keg_res,n=10)
write.csv(keg_res,"F:/Adrian RNA-seq Data/RNA_Count/Files with
    Antibody Counts in/A_Count/Clonal Comparision/
   Producer_Vs_Non/Pathways_For_KEGG_Analysis.csv")

#If you want to look up a pathway online use this code below
   with the pathway code
#browseKEGG(keg_res, 'mmu03013')
```

```
########Creating an interactive plot of my data#############
library(pathview)
logFC <- results.annotated3$log2FoldChange
names(logFC) <- results.annotated3$ENTREZID

pathview(gene.data = logFC,
         pathway.id = "mmu04141",
         species = "mmu",
         limit = list(gene=2.5, cpd=1))

sigGenes3 <- results.annotated3 %>%
  filter(!is.na(ENTREZID)) %>% pull(ENTREZID)
```

```{r eval=FALSE}
##Trying out a regulatory gene analysis###
library(SPIA)

#Checking if I have the ENTZ ID type data that I need for the
    analysis.
results.annotated4 <- results.annotated3[!is.na(results.
    annotated3$ENTREZID),]
results.annotated4 <- results.annotated4[!duplicated(results.
    annotated4$ENTREZID),]

#Manipulating the data frame so that it works for my data
tg1 <- results.annotated4[results.annotated4$padj <0.2,]
DE <- tg1$log2FoldChange
names(DE) <- as.vector(tg1$ENTREZID)
all <- results.annotated4$ENTREZID

#Carrying out the analysis.

# pathway analysis based on combined evidence; # use nB=2000
    or more for more accurate results
res1=spia(de=DE, all=all, organism="mmu", nB=2000, plots=FALSE,
    beta=NULL, combine="fisher", verbose=FALSE)

#make the output fit this screen
res1$Name=substr(res1$Name,1,10)

#show first 15 pathways, omit KEGG links
```

```
res1[1:20,]

#Plotting the graph from the analysis
#plotP(res1,threshold=0.05)
#points(I(-log(pPERT))~I(-log(pNDE)),data=res1[res1$ID
    =="05210",],col="green",pch=19,cex=1.5)

#Furthering the graph analysis
res1$pG=combfunc(res1$pNDE,res1$pPERT,combine="norminv")
res1$pGFdr=p.adjust(res1$pG,"fdr")
res1$pGFWER=p.adjust(res1$pG,"bonferroni")
#plotP(res1,threshold=0.05)
#points(I(-log(pPERT))~I(-log(pNDE)),data=res1[res1$ID
    =="04612",],col="green",pch=19,cex=1.5)

write.csv(res1,"F:/Adrian RNA-seq Data/RNA_Count/Files with
    Antibody Counts in/A_Count/Clonal Comparision/
    Producer_Vs_Non/SPIA analysis.csv")
```

B.4 Promoter over-representation code for 3600bp

```
mypath <- setwd("C:/Users/Adrian/Google Drive/Bioinformatic
    Data/Genomatix_Large_Data_Flowthrough/Results/New analysis
    CHO specific/UTR_matinspector_Results/")

multmerge = function(mypath){
  filenames=list.files(path=mypath, full.names=TRUE)
  datalist = lapply(filenames, function(x){read.delim(file=x,
      header=T)})
  Reduce(function(x,y) {rbind(x,y)}, datalist)
}

mymergedata <- multmerge("C:/Users/Adrian/Google Drive/
    Bioinformatic Data/Genomatix_Large_Data_Flowthrough/Results
    /New analysis CHO specific/UTR_matinspector_Results/")

Merged_Data <- read.table("F:/Genomatix_Large_Data_Flowthrough
    /Merged_data.txt",header=T)
Neede_Data <- read.table("F:/Genomatix_Large_Data_Flowthrough/
    Neede_data.txt",header=T)
colnames(mymergedata)[1] <- c("ENSEMBL")

Filtered_Need <- Neede_Data %>%
  filter(Average_TPM >5)
Filtered_Need <- Filtered_Need[!duplicated(Filtered_Need$Gene.
    Symbol),]

median(Filtered_Need[,3])

#For characterisation purposes will take the top tpm high =
    1000
library("dplyr")
High_Genes <- Filtered_Need %>%
  filter(Average_TPM > 1000) %>%
  mutate(Gene.Symbol=tolower(Gene.Symbol))

All_the_res <- Filtered_Need%>%
  filter(Average_TPM < 1000) %>%
  mutate(Gene.Symbol=tolower(Gene.Symbol))

#High Genes merged
High_genes <- merge(High_Genes,mymergedata,by="ENSEMBL")
```

```
#Frequency code
Frequency_High <- as.data.frame(table(High_genes[,8]))
colnames(Frequency_High) <- c("Family","FreqH")



#Normalization
chunk <- 1826
n <- nrow(All_the_res)
r  <- rep(1:ceiling(n/chunk),each=chunk)[1:n]
d <- split(All_the_res,r)


Giant_Data2 <- Frequency_High
c <- data.frame()
for(i in 1:4){
  s=i
  b <- as.data.frame((d[[i]]))
  c <- merge(b,mymergedata,by="ENSEMBL")
  e <- c[sample(nrow(c), 603442), ]
  a <- as.data.frame(table(e[,8]))
  r <- c[sample(nrow(c), 603442), ]
  q <- as.data.frame(table(r[,8]))
  t <- c[sample(nrow(c), 603442), ]
  w <- as.data.frame(table(t[,8]))
  colnames(a) <- c("Family",paste("Dataset",i,"",sep=""))
  colnames(q) <- c("Family",paste("Dataset",i,"",sep=""))
  colnames(w) <- c("Family",paste("Dataset",i,"",sep=""))
  l <- merge(a,q,by="Family")
  k <- merge(l,w,by="Family")
  k$i <- (rowSums(k[2:4])/3)
  k <- k[,c(1,5)]
  Giant_Data2 <- merge(Giant_Data2,k,by="Family")
}


Giant_Data2$Total_lower <- rowSums(Giant_Data2[3:6])
Giant_Data2$Average <- Giant_Data2$Total_lower/4
Giant_Data2$NormFreq <- Giant_Data2$FreqH -
   Giant_Data2$Average
colnames(Giant_Data2) <- c("Family","FreqH","1","2","3","4","
```

```
    Total_lower","Average","NormFreq")

library("ggplot2")
c<-ggplot(data=Giant_Data2,aes(x=reorder(Family,-NormFreq),y=
    NormFreq)) + geom_bar(stat="identity") +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
c

Normalization_table2 <- Giant_Data2[,c(1,2,7,8,9)]


write.csv(Normalization_table2,"C:/Users/Adrian/Google Drive/
    Bioinformatic Data/Genomatix_Large_Data_Flowthrough/Results
    /New analysis CHO specific/CHO_Analysis/Average_TPM/Family
    Normalization/CHO_Family.csv")
```

B.5 Promoter over-representation code for 1200bp

```
---
title: "Creating the masterfile of TFREs"
output: html_notebook
---

```{r}
Genomatix_Matrix_info <- read.csv("
    GeneIDs_Of_Matrices_Genomatix.csv")

Necessary_Info <- Genomatix_Matrix_info[,c(1,2,5,6)]
human <- read.csv("Human_Gene_ID_To_ENSEMBL.txt")
colnames(human) <- c("Human_Enesmbl","Gene.ID","Gene.name")
human_Gen <- merge(Necessary_Info,human,by="Gene.ID")
write.csv(human_Gen,"Human_Genomatix_Merge.csv")

#human_Gen <- read.csv("Human_Genomatix_Merge.csv")
Mouse <- read.csv("Mouse_Gene_ID_To_ENSEMBL.txt")

colnames(Mouse) <- c("Mouse_Enesmbl","Gene.name_mouse","Gene.
    ID")

Gen_mouse <- merge(Necessary_Info,Mouse,by="Gene.ID")


```

```{r Converting Human to CHO}
Human_To_CHo <- read.csv("Human_To_CHO.txt")
colnames(Human_To_CHo) <- c("Human_Enesmbl","ENSEMBL")
CHO_ENSEMBL_ID_Transcription_Factors_Human <- merge(human_Gen,
    Human_To_CHo,by="Human_Enesmbl")

Mouse_To_Human <- read.csv("Mouse_To_CHO.txt")
colnames(Mouse_To_Human) <- c("Mouse_Enesmbl","ENSEMBL")
CHO_ENSEMBL_ID_Transcription_Factors_Mouse <- merge(Gen_mouse,
    Mouse_To_Human,by="Mouse_Enesmbl")
```

```{r}
#Writing the files of what I need.
CHO_Transcription_Genomatix <-
```
```

```
    CHO_ENSEMBL_ID_Transcription_Factors_Human[,c(7,3,5)]
write.csv(CHO_Transcription_Genomatix,"CHO_TF_Information.csv
    ")
```

```{r}
library("dplyr")
RNA_Seq_Data <- read.csv("F:/Adrian RNA-seq Data/RNA_Count/
    Files with Antibody Counts in/A_Count/Merck File of
    Analysis/TPM of Samples/Annotated_TPM_Filled_in_Genename.
    csv")
RNA_Seq_Data <- RNA_Seq_Data[,c(2,36)]
matrix <- read.csv("Matrix.csv")
Transcription_Factors_In_My_Data2 <- matrix[,c(1,15)]

Transcription_Factors_In_My_Data2 <-
    Transcription_Factors_In_My_Data2 %>%
  group_by(Gene.name) %>%
  summarise_all(funs(sum))
#Transcription_Factors_In_My_Data2 <-
    Transcription_Factors_In_My_Data2[!duplicated(
    Transcription_Factors_In_My_Data2$Gene.name),]



#Transcription_Factors_In_My_Data2$
```

```{r Transcription Factors in my data set,fig.fullwidth=TRUE,
    fig.fullheight=TRUE}
library(RColorBrewer)
library(ggplot2)
library(randomcoloR)
mycolors <- distinctColorPalette(824)
p <- ggplot(data=Transcription_Factors_In_My_Data2,aes(x=
    reorder(Gene.name,-Z.Score..user.defined.),y=Z.Score..user.
    defined., fill = Gene.name))+ geom_bar(stat="identity",
    position=position_dodge()) + scale_fill_manual(values =
    mycolors) +theme_classic() + theme(plot.title =
    element_text(hjust = 0.5),axis.text.x=element_blank(),axis.
    ticks.x=element_blank()) + xlab("TF") + ylab("Z Score") +
  labs(x="Transcription Factors",fill="Gene.name")
p + guides(fill=FALSE)
```

```
ggsave("Log2FoldChangeTranscription Factors.jpeg")
```

```
```{r}
a <- read.csv("CHO_TF_Information.csv")
matrix <- merge(matrix,a,by="Gene.name")
library(dplyr)
Data1 <- RNA_Seq_Data
Data2 <- matrix
Fusion <- merge(Data1,Data2,by="ENSEMBL")
write.csv(Fusion, "Fusion.csv")
Fusion <- read.csv("Fusion.csv")
```

```
```{r Writing the next figure Had to do it in excel}
```

```
```{r}
TF_activators <- read.csv("TFs across Days.csv")

TF_activators <- TF_activators[,c(5,7)]

colnames(TF_activators) <- c("Gene.name","ENSEMBL")
TF_activators$Gene.name <- tolower(TF_activators$Gene.name)

Fusion_Of_activators <- merge(Fusion,TF_activators,by="Gene.
    name")
Fusion_Of_activators <- Fusion_Of_activators[,c(1,3,17)]
Fusion_Of_activators <-  Fusion_Of_activators %>%
  group_by(Gene.name) %>%
  summarise_all(funs(sum))


write.csv(Fusion_Of_activators,"Activators_Only.csv")
```
```

B.6 Code for Bidirectional CHO Promoter Analysis

```
---
title: "Overrepresentation analysis -Yusuf-"
author: "Adrian Bourke"
date: "10/12/2020"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r Creating the sequence file}
library("seqinr")
RNA_seq_Data <- read.csv("F:/Adrian RNA-seq Data/RNA_Count/
   Files with Antibody Counts in/A_Count/All Counts/
   Annotated_avg_tpm_Filled_in_genename.csv")
input <- readLines("Promoter_Sequences_1000up_200_down.txt")
output <- file("5UTR analysis.txt","w")

currentSeq <- 0
newLine <- 0

for(i in 1:length(input)) {
  if(strtrim(input[i], 1) == ">") {
    if(currentSeq == 0) {
      writeLines(paste(input[i],"\t"), output, sep="")
      currentSeq <- currentSeq + 1
    } else {
      writeLines(paste("\n",input[i],"\t", sep =""), output,
         sep="")
    }
  } else {
    writeLines(paste(input[i]), output, sep="")
  }
}

close(output)
```
```

```
```

```{r CHOK1GS Keeping the files together}
CHOK1GS_Genes <- read.delim("C:/Users/Adrian/Google Drive/
    Bioinformatic Data/Genomatix_Large_Data_Flowthrough/Results
    /New analysis CHO specific/1200_bp_results/
    Raw_Files_USC_Promoters/CHOK1GS_Genes.txt")
colnames(CHOK1GS_Genes) <- c("CHOK1 ENSEMBL","Transcript.
    stable.ID.version")
PICR <- read.delim("PICR to CHOK1GS.txt")
colnames(PICR) <- c("ENSEMBL","CHOK1 ENSEMBL")
conversion <- merge(CHOK1GS_Genes,PICR,by="CHOK1 ENSEMBL")
Sequences <- read.csv("usc_Promoters.csv")
Sequences <- merge(conversion,Sequences,by="Transcript.stable.
    ID.version")
Sequences <- Sequences[,c(3,4)]
Bidirectional <- read.csv("Bidirectional_Humanpromoters_in_cho
    .csv")
Bidirectional <- Bidirectional[,c(6,5)]
colnames(Bidirectional) <- c("ENSEMBL","Gene description H")
Sequences <- merge(Sequences,Bidirectional,by="ENSEMBL")
```


```{r Downliading the file}
RNA_Seq_Data <- read.csv("F:/Adrian RNA-seq Data/RNA_Count/
    Files with Antibody Counts in/A_Count/Merck File of
    Analysis/TPM of Samples/Annotated_TPM_Filled_in_Genename.
    csv")
merge <- merge(Sequences,RNA_Seq_Data,by="ENSEMBL")
merge <- merge[!duplicated(merge$ENSEMBL),]
merge <- merge[!grepl("Sequence unavailable",merge$Sequence),]
```


```{r Splitting the data up}
library(dplyr)
High <- merge %>%
   filter(Average_TPM > 1000) %>%
   mutate(Gene.name = tolower(Gene.name))
High <- High[,c(1,2)]

#Medium <- merge %>%
  #filter(Average_TPM >2) %>%
```

```
 # mutate(Gene.name = tolower(Gene.name))
Medium <- merge[,c(1,2)]
# Medium <- Medium[sample(nrow(Medium),131,replace=F),]
```

```

```{r Creating the fasta files}
library(seqRFLP)
dataframe2fas(High,"Bidirectional_Genes.fasta")
dataframe2fas(Medium,"All_Genes_DF.fasta")


```
```

B.7 Code for Mouse Versus CHO Promoter Analysis

```
---
title: "CHO_Vs_Mouse_Results"
author: "Adrian"
date: "2022-09-13"
output: html_document
---

```{r }
setwd("3600bp promoters/")
allfiles=list.files()
allfiles2=allfiles[grep(".tsv",allfiles)]
f.names=gsub('.{4}$', '', allfiles2)

mat.fam=c();mat1=c()
for(i in 1:length(allfiles2)){
  assign(f.names[i],read.delim(allfiles2[i]))
  a=get(f.names[i])
  a[,5]=as.character(a[,5])
  a[,7]=as.character(a[,7])
  assign(f.names[i],a)
  mat.fam=c(mat.fam,get(f.names[i])[,5])
  mat1=c(mat1,get(f.names[i])[,7])
}

matrix.families=as.data.frame(table(mat.fam))
check = as.data.frame(mat.fam)
matrix1=as.data.frame(table(mat1))
```



```{r Checking the Mouse data}
library(dplyr)
setwd("C:/Users/Adrian/Google Drive/Bioinformatic Data/
   Genomatix_Large_Data_Flowthrough/")
Neede_Data <- read.table("Neede_data.txt",header=T)
Merged_Data <- read.table("Merged_data.txt",header=T)

```
```

```
```{r The rest of the code}
#Trying my merge with upper and lower case changes
Neede_Data <- Neede_Data %>%
  mutate(Gene.Symbol=tolower(Gene.Symbol))
Merged_Data <- Merged_Data %>%
  mutate(Gene.Symbol=tolower(Gene.Symbol))
Filtered_Need <- Neede_Data %>%
  filter(Average_TPM >3)
Filtered_Need <- Filtered_Need[!duplicated(Filtered_Need$Gene.
   Symbol),]

library("dplyr")
High_Genes <- Filtered_Need %>%
  filter(Average_TPM > 1000) %>%
  head(100) %>%
  mutate(Gene.Symbol=tolower(Gene.Symbol))



High_genes <- merge(High_Genes,Merged_Data,by="Gene.Symbol")
High_genes <- High_genes[sample(nrow(High_genes),80514),]
Matrix_Frequency_High <- as.data.frame(table(High_genes[,7]))
Frequency_High <- as.data.frame(table(High_genes[,6]))

matrix.families <- matrix.families[order(matrix.families$Freq,
   decreasing = T),]
Frequency_High <- Frequency_High[order(Frequency_High$Freq,
   decreasing = T),]
Matrix_Frequency_High <- Matrix_Frequency_High[order(
   Matrix_Frequency_High$Freq,decreasing = T),]
matrix1 <- matrix1[order(matrix1$Freq,decreasing = T),]

matrix.families$ID <- seq.int (nrow(matrix.families))
Frequency_High$ID <- seq.int (nrow(Frequency_High))
Matrix_Frequency_High$ID <- seq.int (nrow(
   Matrix_Frequency_High))
matrix1$ID <- seq.int (nrow(matrix1))

colnames(matrix.families) <- c("Family","Freq in CHO","
   Position in CHO")
colnames(Frequency_High) <- c("Family","Freq in Mouse Ortho","
   Position in Mouse")
colnames(Matrix_Frequency_High) <- c("Matrix","Freq in Mouse
```
```

```
    Ortho","Position in Mouse")
colnames(matrix1) <- c("Matrix","Freq in CHO","Position in CHO
    ")
```

```

```{r}
#Sample from mouse
Family_Comparision <- merge(matrix.families,Frequency_High,by
    ="Family")
Family_Comparision$Normalised_Position <- Family_Comparision$`
    Position in CHO`- Family_Comparision$`Position in Mouse`
Family_Comparision <- Family_Comparision[-68,]
Family_Comparision$SD <- sd(
    Family_Comparision$Normalised_Position)
Family_Comparision$TF_Percentage <- (Family_Comparision$'Freq
    in CHO'/sum(Family_Comparision$'Freq in CHO'))*100
Family_Comparision$Mouse_TF_Percentage <- (Family_Comparision$
    'Freq in Mouse Ortho'/sum(Family_Comparision$'Freq in Mouse
     Ortho'))*100
Family_Comparision$Difference <-
    Family_Comparision$TF_Percentage -
    Family_Comparision$Mouse_TF_Percentage
mean(Family_Comparision$Difference)
Family_Comparision$Normalized_Frequency <- Family_Comparision$
    `Freq in CHO` - Family_Comparision$`Freq in Mouse Ortho`
Family_Comparision$Normalized_Frequency <- abs(
    Family_Comparision$Normalized_Frequency)
Family_Comparision$Perc_Difference <- (sum(
    Family_Comparision$Normalized_Frequency)/80514)*100
Family_Comparision$sumdif =  abs(Family_Comparision$Difference
    )
Family_Comparision$Difference_2 =sum(Family_Comparision$sumdif
    )
write.csv(Family_Comparision,"Family_comparision.csv")
```

```