The Dissertation Committee for Joseph Kit Cheng Certifies that this is the approved version of the following dissertation:

Beyond protein factories: Expanding the synthetic biology toolkit for engineering mammalian hosts

Committee:

Hal Alper, Supervisor

Lydia Contreras

George Georgiou

Christopher Sullivan

Beyond protein factories: Expanding the synthetic biology toolkit for engineering mammalian hosts

by

Joseph Kit Cheng, M.E., B.S.Ch.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2016

Dedication

For my parents, who relinquished their opportunities so that I may have them.

Acknowledgements

Most importantly, I would like to acknowledge my advisor, Dr. Hal Alper. Words simply cannot express my gratitude for Hal's support and mentorship during my graduate studies. The support, advice, and opportunities I received from Hal from the moment I stepped onto campus were instrumental to my successes throughout my graduate career. I am particularly grateful that Hal took a chance on an incoming graduate student who stepped away from academia for several years, and fostered the scientific curiosity and rigor that brought me to the University of Texas. Despite his ever-growing responsibilities and busy schedule, there was always time to discuss ongoing projects and he was always supportive of how project directions evolved with incoming data. I could not envision a better experience as a graduate student than being an alum of the Alper lab.

Next, I would like to thank my dissertation committee members Dr. Lydia Contreras, Dr. George Georgiou, and Dr. Christopher Sullivan for taking the time out of their already hectic schedules to provide feedback and suggestions towards this dissertation work. Their insight and recommendations from the preliminary oral examination astutely challenged the proposed work and my thought processes for these projects. With additional feedback following the examination, those projects culminated in the work described here. Again, I am very grateful for their suggestions and feedback, which ultimately pushed my growth as a researcher.

Additionally, the past and present members of the Alper lab provided crucial intellectual discussions and essential camaraderie to make my graduate studies an unforgettable adventure. The evolution of a nascent lab is closely tied to the demeanor of the graduate and undergraduate students in the lab, and the opportunities for professional

discussions and mentorship with common banter set the stage for a thoroughly enjoyable experience. I would like to particularly acknowledge the graduate students before me (Amanda Lewis, Eric Young, John Blazeck, Leqian Liu, Kate Curran, Nathan Crook, Sun-mi Lee), and our first postdoctoral researcher (Dr. Jie Sun) who established many lab processes and protocols that made routine tasks as painless as possible. Going forward, this collaborative environment was maintained by fostering discussions with current graduate students Kelly Markham, James Wagner, Haibo Li, and Nick Morse.

The work described in this dissertation would not be made possible without the assistance of several talented undergraduates from the McKetta Department of Chemical Engineering: Do Soon Kim, Aakash Jani, and Annabel Wang. Specifically, Do Soon Kim made significant contributions to the projects described in Chapters 2, 4, and 5; meanwhile, Aakash Jani contributed to the work found in Chapters 4 and 5. Annabel Wang strongly demonstrated her competency and capacity during her time in the lab, making contributions to the work found in Chapters 3-6. Furthermore, Scott Tucker, a Cellular and Molecular Biology graduate student, contributed significantly to the successes and is a co-author to the work described in Chapter 6, and assisted in the groundwork underlying the pivotal experiments described in Chapter 3.

Without a doubt, the support for instruments, materials, and resources required for the work described in this dissertation greatly contributed to the successes. Richard Salinas and Charlie Gonzales from the Institute for Cellular and Molecular Biology's core facilities provided critical assistance to the flow cytometry and Sanger sequencing work required in the entire dissertation. Within the McKetta Deparment of Chemical Engineering, present and former staff members Jim Smitherman, Shallaco McDonald, Butch Cunningham, Kevin Haynes, Eddie Ibarra, Tammy McDade, Teri Sahba, and Ben Hester ensured that instruments were readily maintained and resources were promptly procured. Additional

support from departmental staff Kate West Baird, Randy Rife, Jason Barborka, Carrie Brown, and Courtney Hazlett facilitated many of the daily administrative tasks.

Lastly, those who comprise my Austin family made meaningful impact throughout my graduate studies. Of course, many of the individuals from our lab, with their cohorts and partners-in-crime, created an important network with whom I shared wonderful exploratory and dining adventures around Austin. Gregory Blachut, Sunshine Zhou, Jon Bollinger, Ellen Line, Nathan Crook, Jie Sun, Brent Sherman, Lynn Li, Kate Curran, Sun-mi Lee, Leqian Liu, Wenzong Li, Greg Mullen, Alex Pak, John Leavitt, Kelly Markham, Jovan Kamcev, Josh Laber, Amanda Paine, Matt Beaudry, and Kevin Baldridge most notably helped maintain a steady-state of happiness amidst the tumultuous ride that is exploratory research and graduate study. Finally, Steven Blackwell, the biggest contributor to this stability, provided the much-needed support during my critical trials and tribulations as a graduate student.

Beyond protein factories: Expanding the synthetic biology toolkit for engineering mammalian hosts

Joseph Kit Cheng, Ph.D.

The University of Texas at Austin, 2016

Supervisor: Hal S. Alper

The incredible clinical and commercial successes of recombinant protein therapeutics cemented the use of mammalian cells as the premier production hosts for these products. However, we can further exploit these cells to harness their potential for addressing current and future medical needs through metabolic and advanced engineering of these cells. To do so, we need a deeper understanding of the intricate gene regulation network that governs these cells and the ability to attain precise control of gene expression levels. In addition, some of these applications, such as gene therapy and immunotherapy, could benefit greatly by refraining from using viral-derived genetic elements. Therefore, this work seeks to establish additional transcriptional control elements to improve our ability to regulate expression with generalizable approaches and methods, facilitating the adaptation of these techniques for any mammalian cell type of interest.

Here, we successfully demonstrated three key genetic elements can be utilized to tune gene expression in a rational manner. First, we conducted a genome-wide screen to survey genomic integration sites that support high transcriptional activity. We showed that CRISPR/Cas9-mediated *de novo* integration into one of these transcriptional hot-spots at the *GRIK1* locus resulted in a 2.4-fold increase in heterologous gene expression over random integration. Subsequently, we set the groundwork necessary to evaluate a cell line

development strategy that aims to increase the frequency of successful *de novo* targeted integrations. Second, we utilized two approaches for rational promoter engineering. We established a transcriptomics-guided workflow for *de novo* synthetic promoter design based on the Design-Build-Test paradigm. By using this workflow, we generated two synthetic designs that were comparable to a strong viral promoter and a strong endogenous promoter. We also employed an alternative approach by creating hybrid promoters, which resulted in a hybrid promoter variant that was also comparable to the same viral and endogenous promoters. Third, we exploited the general mammalian terminator structure and created a synthetic terminator that was comparable to a strong viral terminator. We evaluated 12 endogenous and 30 synthetic terminators for heterologous gene expression and revealed interactions between several key components of the terminator. Critically, we showed that transgene expression was 1.9x higher with endogenous and synthetic elements when compared with strong viral-derived elements. Ultimately, we showed that transgene expression can be finely adjusted by the approaches and methods described in this dissertation, and that viral-derived elements can be readily substituted by our synthetic designs.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction[1]

The increase in quality, quantity, and complexity of recombinant products heavily drives the need to predictably engineer model and complex mammalian cells for their bioproduction. However, until recently, limited tools offered the ability to precisely manipulate their genomes, thus impeding the full potential of rational cell line development processes. Furthermore, the issues of cell productivity, cell stability, cost of goods and services, and speed of development have put new demands on the field[1, 2]. Synthetic biology tools, which have long been applied to microbial cell systems, can improve the speed of R&D and reduce cost of goods. Recent advances in site-specific genome editing techniques[3, 4], genetic regulatory elements[5], and metabolic and pathway engineering[6] of mammalian cell systems can ultimately facilitate faster and more flexible cell line development. In particular, targeted genome editing can combine the advances in synthetic and systems biology with current cellular hosts to further push productivity and expand the product repertoire. Moreover, many of these advances collectively enable the precise and gradated expression levels required for other metabolite production and immune cell engineering applications.

1.1. CELLULAR ENGINEERING OF MAMMALIAN PRODUCTION HOSTS

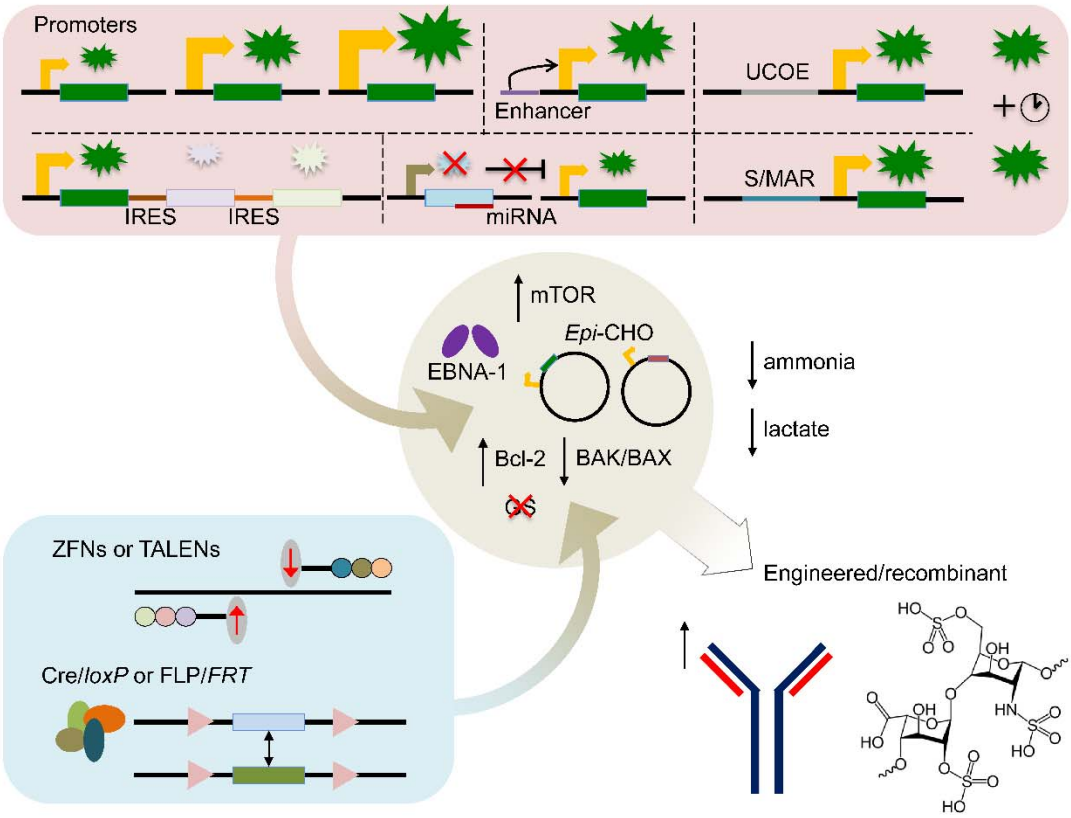
Our capacity to culture and engineer mammalian cell systems for protein production has rapidly expanded in past decades and has raised the importance of mammalian bioprocess engineering efforts. These advancements have led to increases in

[1] Part of the content in this chapter adapted from two previously authored publications. JKC equally contributed to the writing of the 2012 manuscript and wrote the 2014 manuscript, and incorporated additional information that resulted in this chapter.

the quality, quantity and complexity of recombinant products. This improvement is most apparent in the ever-increasing titers of monoclonal antibodies that have gone from 50 mg/L to upwards of 5 g/L in just over two decades[7]. Mammalian cells remain the predominant host for producing antibodies and other protein therapeutics based on advantageous post-translational modifications, reduced immunogenicity, and the establishment of an infrastructure of mammalian cell cultivation and bioprocess engineering at pharmaceutical companies. Yet, issues of cell productivity, cell stability, cost of goods and services, and speed of development have put new demands on the field.

In general, the cost of bringing a drug to the market is quite high[2] as a result of significant R&D, clinical testing, and failure rates. While tools of metabolic engineering and synthetic biology cannot solve clinical testing and failure rates, they can improve the speed of R&D as well as reduce cost of goods. To this end, new synthetic approaches are becoming available to improve the speed, accuracy, and yield of cell culture systems. These advances will continue to solidify the need for the cell-based bioprocess engineering that lies at the heart of most protein-based pharmaceutical companies. Recent advancements in site-specific genome editing techniques, genetic regulatory elements, and metabolic and pathway engineering of mammalian cell systems improve mammalian host cell engineering (**Figure 1-1**). These advancements, coupled with a better understanding of cell systems captured through -omics approaches[8, 9], facilitate faster and more flexible cell line development, ultimately reducing cost and time.

Figure 1-1: Genome editing and genetic elements enable metabolic and pathway engineering in mammalian cell line development.



Targeted genome editing tools permit manipulation of mammalian cells with increased specificity over homologous recombination and illegitimate integration, and facilitate the introduction of synthetic parts such as enhancers, promoters, miRNAs, and other regulatory elements into mammalian cells. The combination of these tools expands metabolic engineering in these cells to improve production and hasten cell line development.

Metabolic engineering serves as the discipline that combines genome editing techniques and genetic control elements to manufacture products of interest. While metabolic engineering has long been applied to microbial organisms[10, 11], limitations in the ability to make precise genetic manipulations delayed complete metabolic engineering of mammalian cell systems. Utilizing many of the tools described below, researchers are now beginning to rationally engineer the metabolism and pathways of mammalian cells for

enhanced product formation, higher cell densities, and decreased byproduct formation, as depicted in **Figure 1-1**. This is a critical advancement that could expand the types of compounds produced in mammalian cells, as well as improve titers and productivity. Apoptosis, for example, is a normal cellular phenomena but is detrimental in a cell culture process [12]. Recently, researchers have used ZFNs to generate CHO cell lines lacking pro-apoptotic genes and these cell lines have shown resistance to apoptotic inducers[12]. This study utilized both genomic information and a site-specific genome editing technique to modify the intrinsic apoptosis pathway. Alternatively, up-regulation of miRNA was shown to inhibit anti-apoptotic genes[13] and microarray profiling efforts in HEK293 cells identified several miRNAs as up-regulated when cells undergo apoptosis[14]. These efforts will lead to more robust host cells.

Similarly, reduced formation of lactate, a metabolic byproduct, is known to improve cell growth and product formation[15, 16]. An apoptosis-resistant, lower lactate producing CHO cell line was recently developed by over-expressing an anti-apoptotic gene and suppressing a lactate-converting enzyme[17]. Over-expression of other anti-apoptotic genes resulted in shifted nutrient consumption profiles and decreased lactate and ammonia accumulation[18]. Improvements in efficiency and cellular metabolism have also been achieved by generating glutamine synthetase null strains using ZFNs[19].

Recently, CHO-derived cell lines were engineered to constitutively express a mammalian global sensor (mTOR)[20] that is responsible for controlling several metabolic activities. The ectopic expression of mTOR in CHO-K1 cells resulted in several profound effects including higher specific productivity[20]. Ultimately, IgG production was increased nearly four-fold over non-engineered parental cells, and these benefits were translated in bioreactors[20].

The first study using metabolic engineering to produce a non-protein product, heparan sulfate (HS), in CHO cells both demonstrated the feasibility and highlighted the limitations of such an approach in mammalian systems[6]. Although the activity of the engineered HS is significantly less than current pharmaceutical grade heparin, this work illustrates the potential to metabolically engineer mammalian cells for complex products. Furthermore, disaccharide analysis of the engineered HS suggests that tuning enzyme expression (possibly through the genetic elements previously described) could lead to pharmaceutical grade HS. Mammalian cell systems also aid vaccine development by producing glycoproteins and virus-like particles. Recently, two studies demonstrated that recombinant hemagglutinin produced from HEK cells confer protection against pathogenic influenza viruses[21, 22]. Further improvements have been achieved by engineering cellular glycosylation and protein secretion, and these findings have previously been summarized[23, 24].

1.2. TRANSITIONING TOWARDS SITE-SPECIFIC TRANSGENE INTEGRATION IN MAMMALIAN HOSTS

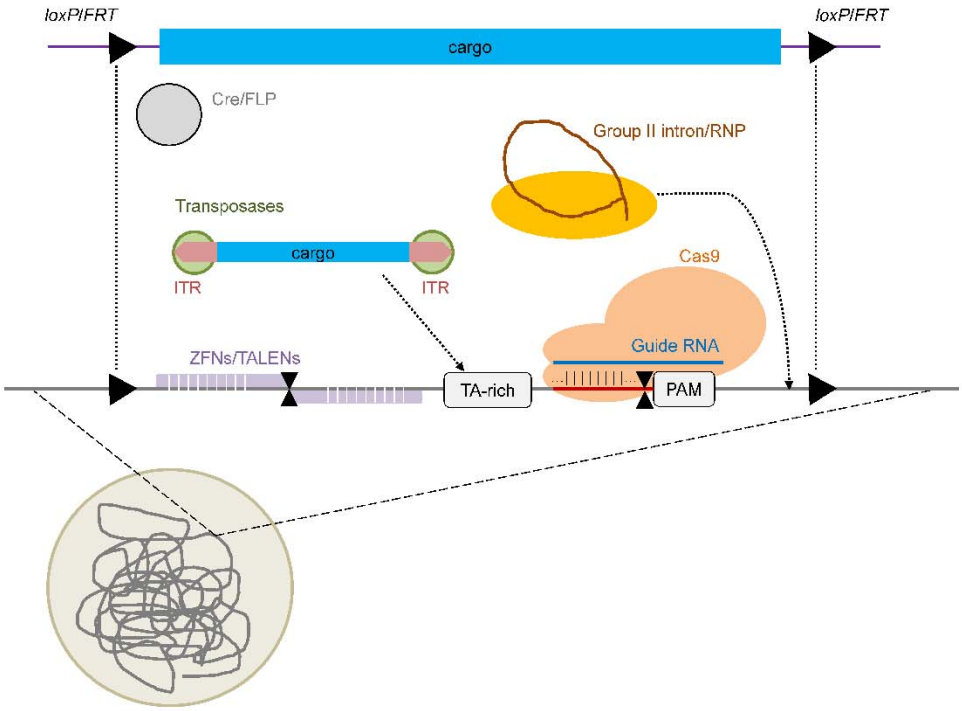
Achieving mg/L quantities of an antibody can easily be achieved by introducing a transgene into a host cell. However, the production of gram quantities (upwards to tens of grams per liter) requires a combination of cell line engineering, expression optimization, and advanced culturing control. As we continue to push the envelope for titer and to expand our scope of products, we require increasingly precise methods to modify the genome. Yet, until recently, the genetic tools available to make targeted edits exhibited rather coarse resolution even for model organisms.

Enabling targeted editing of eukaryotic and mammalian genomes diversifies the biopharmaceutical repertoire and offers design of selection systems beyond current schemes (*e.g.* dihydrofolate reductase-deficient CHO cell lines or antibiotic resistance). As

described above, metabolic engineering of mammalian cells to produce the first bioengineered heparin is imminent[6]. Furthermore, combining pathway understanding with well-characterized synthetic genetic parts will enable additional complex products to be produced. These efforts will be greatly expedited by targeted genomic edits. Moreover, these technologies open up the possibility to engineer hosts cell lines without selection markers (or in some cases, for marker-less genome editing), thus bypassing antibiotic selection schemes that can be suboptimal depending on the selecting agent used[25].

With the increasing synthetic toolkit for genetic editing, cell line development is about to experience a renewal at the intersection of systems biology and synthetic biology. Recent advancements in a spectrum of targeted genome editing tools (**Figure 1-2**) ranging from coarse, loci-level resolution to precise, base-pair specific modifications can be readily adapted to each application. High-throughput analytical methods and cheap genome sequencing enable a more precise linkage between genotype and phenotype. These approaches could theoretically allow researchers to make concerted and rational decisions about the genotype of a cell line. This capacity to move beyond random libraries and integrations (or at the very least, understand why and how a cell line performs as it does in an effort to recapture high productivity for another molecule), require a sophisticated suite of genome editing tools. Moreover, the complexity across different cell types and cell lines makes such a vision challenging.

Figure 1-2: Targeted genome editing spans coarse, regional locus-level recognition to nucleotide-level specificity.



Targeted genome editing tools permit manipulation of a variety of host cells relevant to the biopharmaceutical industry. The combination of the tools described here improves prospects for engineering in these cells. Transposases recognize ITR sequences flanking the DNA cargo and mediate movement of the cargo elsewhere in the genome (*e.g.* TA-rich regions). Recognition sites for recombinases such as loxP/FRT are specified as DNA sequences. Group II introns require additional proteins (*e.g.* reverse transcriptase) to fully carry out their targeted integration into the genome. The target DNA sequence is sufficient for inducing specific cleavage by nucleases; however, the guide RNA (gRNA) is required to direct Cas9 nuclease to the target site including the PAM. Abbreviations and symbols: denote the targeted cleavage site triggered by the nuclease. ZFN, zinc-finger nuclease. TALEN, transcription activator-like effector nuclease. PAM, protospacer adjacent motif. ITR, inverted terminal repeats. RNP, ribonucleoprotein.

A variety of techniques (**Figure 1-2**) have been developed to edit genomes including adapting components from other species (*e.g.* transposases, recombinases, group II introns, and RNA-directed nucleases) and creating synthetic approaches comprised of discrete building blocks (*e.g.* zinc-finger and transcription activator-like effector nucleases). This spectrum of targeted genome editing tools has a myriad of capabilities

and applications, but each approach has distinct drawbacks and limitations (see **Table 1** from *Curr. Opin. Biotechnol.* 2014 publication[26]).

1.2.1. Exploiting established genetic elements

Genetic engineering is required to transform mammalian host cells into super-producers of proteins. Specifically, efficient mammalian cell engineering requires precise, reliable genome editing techniques to enable the expression of heterologous genes and deletion of unwanted genes. In contrast to the ease of genome editing in microbial systems, mammalian cell engineering still relies heavily on semi-random integration[8, 27] and low probability homologous recombination events[28-30] coupled with laborious screening[8, 31-33]. However, classes of enzymes naturally exist which can recognize specific DNA sequences and modify those genomic loci. These enzymes are being engineered as synthetic tools to recognize new sequences and perform precise genomic modifications, such as double strand breaks (DSBs) that facilitate gene deletions, insertions and replacements.

Transposable elements, such as transposons and group II introns, and recombinases were among the first genome editing tools applied to mammalian cell lines. Transposons, group II introns, and recombinase-mediated cassette exchanges link desired genetic cargo as part of the mobile element with recognition sites flanking this cargo (**Figure 1-2**). Transposon-aided insertion is mediated by sequence recognition rules (*e.g.* TA-rich region) that vary depending on the transposable element. Self-splicing group II introns target DNA with the same or similar sequences to the donor sequence containing the intron and incorporate the intron sequence through reverse transcription. Lastly, recombinases recognize specific DNA sequences (*e.g. loxP* and *FRT*) between which exchanges or rearrangements occur.

In the context of pharmaceutical applications, transposable elements (for an active list in vertebrates, see review by Ivics[34]) such as *Sleeping Beauty* (*SB*) and *PiggyBAC* (*PB*) can serve as a safe alternative to viral vectors for gene delivery[34-39]. Such systems require expression of the transposase and can deliver genetic cargo to pre-set loci due to particular sequence preferences and genomic integration patterns of the transposases[34]. As an example, functional wild-type heme oxygenase-1 was delivered to the livers of mice inflicted with sickle cell disease using a *SB* transposase plasmid[36] and this *SB* system recently showed a positive outlook for a human clinical study[37]. Moreover, using *SB* to generate fluorescent reporter clones in human embryonic stem cells (hESCs) preserved the undifferentiated state and differentiation pattern, facilitating the study of a complex cell type[38], thus indicating a low level of genetic and epigenetic disruption by this system. Since different transposons have varying cargo size limitations, local transposition tendencies, and integration site preferences, *PB* was used instead of *SB* in a mice genetic screen for oncogene discovery[39].

Similar to transposases, recombinases such as Cre and FLP require flanking recognition sites at the target genomic locus, but can enable insertions, deletions, replacements, and rearrangements. Pre-integrated sites (potentially at transcriptional hot-spots in the genome) can enable re-targeting[40]. Such re-targeting in human cells is achievable at high rates (over 10%) as demonstrated by swapping of fluorescent markers[41]. Even in hESCs and mouse fibroblasts (NIH 3T3), Cre and FLP recombinases were used in conjunction with high recombination activity[42]. When coupled with inducible or tissue-specific promoters, controlled, targeted expression in model organisms with limited tools is possible[43]. Recently, the Cre/*loxP* system was used to construct human artificial chromosomes that are stably maintained, serving as another potential alternative to viral vectors[44]. Despite promise and versatility of this genome editing tool[41-43], recombination

efficiencies depend on the cell type and genomic locus, potentially resulting in a labor-intensive endeavor (especially when screening for desired clone).

Recombinase and transposon based genome editing shows great promise in non-mammalian hosts. For example, by combining the Cre recombinase with the mobile group II intron (Targetron) systems, the Genome Editing via Targetrons and Recombinases (GETR) platform mediated large genome edits in bacteria with high efficiencies[45]. In this platform, careful design of recombination sites and selection of introns can mitigate potentially deleterious events such as scarring and unplanned homologous recombination between introns[45]. Previously, highly specific gene disruption was achieved in *E. coli*, a common biopharmaceutical production host, with frequencies up to 22%[46]. Group II intron activity was also demonstrated in eukaryotes, and the resulting site-specific integration or double-strand break (DSB) induced homology directed repair (HDR) was dictated by Mg^{2+} concentration[46].

Artificial chromosomes present an alternative technology that does not require integration into the host genome. This technology is used in some bacterial and fungal applications and can support large quantities of recombinant DNA. Recently, artificial chromosome expression (ACE) technology was used to generate monoclonal antibody expressing CHO cells exhibiting high productivity[31]. The recent discovery of small, circular microDNAs in mammalian tissues represents another genetic avenue that could be engineered to complement and extend existing transgene expression technologies[47]. Improved transient expression was demonstrated in modified HEK293 cells expressing Epstein-Barr virus nuclear antigen 1[48]. A similar approach in CHO cells using the *Epi*-CHO transient gene expression (TGE) system showed up to a 64% increase in monoclonal antibody titer when compared to previously reported systems[49, 50]. These tools collectively provide significant flexibility and precision in genome editing. These represent significant

improvements over standard practices such as homologous recombination or illegitimate integration, resulting in faster cell line development and higher titers for better bioprocesses and potentially easier downstream product separation. Thus, many of these approaches could be further adapted and optimized to achieve genome editing in mammalian hosts.

1.2.2. Customizable genomic editing via targeted nucleases

The ability to precisely cleave DNA (through restriction enzymes) was a critical turning point that led to the establishment of recombinant DNA technology and the growth of the biopharmaceutical industry. Recently, there has been great effort in establishing these techniques *in vivo*. By inducing a double-stranded break (DSB) at the cleavage site, the cell's repair pathways are triggered and the break is generally resolved with non-homologous end joining (NHEJ) or HDR. NHEJ results in nucleotide insertions or deletions (indels), which is ideal for interrupting gene function at the specific locus and introducing heterologous gene expression.

Zinc-finger nucleases (ZFNs) can be obtained by fusing a zinc finger DNA-binding domain to a DNA-cleavage domain (typically derived from *FokI* Type IIS endonuclease). In this manner, custom editing targets can be specified at the sequence level. ZFNs do not require generic targeting sequences and are modular in assembly, allowing great flexibility in their targeting[29]. ZFNs facilitate both genomic integrations and gene knockouts[29]. Custom ZFNs can be ordered through companies such as Sangamo BioSciences and have been demonstrated in a variety of cell types and applications[51], including the rapid and efficient deletion of genes[52, 53]. Recently, zinc-finger recombinases (ZFR) were developed by fusing zinc finger domains and serine recombinases, and utilized in human cells to deliver reporter genes at specific loci[27]. Although this method requires pre-insertion of ZFR recognition sites, DNA damage responses are circumvented and thus higher levels of

specificity are achieved[27]. Recently, this technology enabled improved cell line development for monoclonal antibody production in Chinese hamster ovary cells[19]. Despite different methods used to construct ZFNs[54, 55], off-target cleavage was detectable in many applications[54, 56, 57]. In particular, ZFNs targeting human *CCR5* often showed cleavage activity at the highly homologous *CCR2* locus[54], similar to CRISPR/Cas9[58]. Moreover, ZFNs can be costly and time-consuming to produce, with mutation efficiencies only up to 18.8% in certain applications[55].

Like ZFNs, transcription activator-like effector nucleases (TALENs) contain a DNA-binding domain and cleavage domain. Transcription activator-like effector nucleases (TALENs) are also modular in nature and can be built to recognize any DNA sequence[59, 60]. Efficient endogenous deletions[61] and gene insertions[62] were recently demonstrated in human cells using TALEN architecture. Comparable efficiencies to ZFNs were seen when five distinct genomic loci were targeted in both hESCs and induced pluripotent stem cells (iPSCs)[62]. However, a *CCR5*-specific TALEN showed less off-targeting activity at the *CCR2* locus compared to its ZFN counterpart, suggesting better specificity[63]. Comparable efficiency was observed between CRISPR/Cas9 and TALENs when modifying human pluripotent stem cells[64]. Furthermore, TALEN-mediated HDR in β-thalassemia iPSCs successfully corrected disease-causing mutations without transgene integration and while maintaining pluripotency[65]. In another study, hESCs and iPSCs were edited by TALENs at four genes to produce representative human disease-related models[66]. Even heritable mutations eliminating *IgM* function can be established using TALENs[67]. Miller, *et al.* and Cermak, *et al.* described strategies for TALEN designs[68, 69], though production capacity was rather limited. To address this limitation, recently, the FLASH method for automated medium- to high-throughput TALEN production can generate

>7,200 arrays per year[70]. As a result, this technology is becoming more accessible for genome editing[71].

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR associated (Cas) genes natively function together as a prokaryotic immune system. In recent years, this system has been adapted as a potent targeted genome editing tool with broad host system capabilities[58, 64, 72-82]. The recognition sequence for Cas nuclease cleavage is a short protospacer adjacent motif (PAM) sequence (**Figure 1-2**), enabling targeted DSBs across many cell types[72-76]. The most common Cas9 nuclease is derived from *S. pyogenes* (SpCas9), yet orthologs show orthogonal function, enabling simultaneous and independent targeted gene regulation in bacteria and human cells[75]. While heralded as a new, powerful approach for genome editing, significant off-targeting (up to 50%) was observed in human cells[72]. This off-targeting primarily triggers NHEJ, often with deleterious consequences at unintended genomic loci, and modifications to Cas9 converting it from a nuclease to a nickase (or using paired Cas9 nickases) can reduce this promiscuous activity[74, 77, 83, 84].

Aside from targeted gene disruption, the targeted recognition properties of these nucleases enable novel applications as synthetic markers and actuators[58, 64, 72-82]. As an alternative to fluorescence *in-situ* hybridization, chromatin conformation and dynamics in live human cells can be studied using a fluorescently tagged, nuclease-deactivated Cas9[78]. Even heavily repeated regions, such as telomeres, can be imaged with this repurposed Cas9[78]. With comparable performance to RNA-interference, genome-scale knockdown screening is possible with CRISPR/Cas9 in *E.coli*[79] and in human cells[85], evaluating >18,000 genes with >64,000 unique guide RNA sequences[80]. Activating and repressive domains can be fused to catalytically inactive Cas9 to enhance or decrease endogenous gene expression in eukaryotes[76, 81], while similar strategies can be combined with TAL

effector arrays[86] and zinc finger proteins[87]. Even targets with small loci, such as microRNA, can be disrupted[88]. Furthermore, targeted editing of histone modifications can be achieved by combining histone demethlyase activity with TAL effector arrays[89] or nuclease-deficient Cas9[90], thus modulating gene expression at an epigenomic level. These quick adaptations of the nuclease systems certainly do not entirely encompass their potential as they have yet to be fully characterized. Moreover, these approaches demonstrate a newfound way to modify more than just genome sequences in cell lines of interest.

While there are still challenges to these recent editing techniques, methods are established to better probe and address their limitations, particularly for off-targeting with nuclease directed editing[56, 57, 72, 77, 82-84, 91, 92]. With this improved understanding of the Cas9 nuclease activity, researchers developed high fidelity variants the common *S. pyogenes*-derived Cas9 in recent years[93, 94] and a small-molecule inducible variant[95] to combat the propensity for off-target activity. Such mutagenic concerns preclude their use in the clinic currently, but their potential use for personalized medicine and gene therapy is certainly becoming more realistic. By combining several of these techniques, such as stable expression of a transgene via an episomal vector[96], certain diseases can be treated with nuclease-induced HDR of mutant/diseased alleles. Alternatively, the delivery of the Cas9 became a focus of attention to address their potential applications in conjunction with adeno-associated virus, leading to novel methods to introduce this nuclease activity[97-99]. Recent efforts also demonstrated editing of the mouse and human immunoglobulin genes[100]. Thus, in addition to their application in engineering host cell lines for high productivity and amenable development, these editing tools are poised to transform therapies as their technologies mature.

1.3. *CIS*- AND *TRANS*-ACTING GENETIC ELEMENTS

Genome editing alone is not sufficient to provide the gene expression regulation required of these mammalian hosts. To address this issue, genetic elements can be used to synthetically regulate gene expression and facilitate protein production and metabolic engineering efforts. Some of these tools include genetic elements such as enhancers, promoters, internal ribosome entry sites (IRES), ubiquitous chromatin opening elements (UCOEs), scaffold/matrix attachment regions (S/MARs), and micro RNAs (miRNAs), which can be combined with both recombinant and native genes to enhance genetic engineering efforts in these cells. The utility of related genetic elements has been demonstrated in microbes[101-105]; however, adaptation and adoption across mammalian genomes is in its early stages[106].

Enhancers are *cis*-acting genetic elements that help regulate transcriptional activity. Recently, a short enhancer sequence was paired with a minimal promoter, yielding tunable expression and levels exceeding that of the strongest known native mammalian promoters[107]. Such enhancer elements can be systematically dissected using high-throughput techniques recently developed and applied to mammalian regulatory elements[106, 108]. In a similar manner, these techniques can be adapted for the systematic analysis of other elements, such as promoters and silencers and can also enable the development of novel regulatory elements. These efforts can be expanded through bioinformatics mining of recently published sequences such as the CHO genome[109].

Promoters with well-characterized transcription levels are critical genetic parts that enable heterologous gene expression and construction of advanced genetic circuits. While tools are widely available for tuning gene expression in microorganisms[110-112], the counterparts in mammalian cell systems are just emerging. Two types of promoter elements are useful for recombinant cell lines: constitutive (constant expression) and

inducible (modulated expression by a small molecule or other trigger). By modifying the TATA and CAAT box elements, synthetic constitutive promoters capable of 40-fold range expression levels were recently developed and implemented in several cell types[113, 114]. These promoters allow for better, higher-resolution optimization of gene expression in the context of metabolic pathways and heterologous transgene expression. Furthermore, sequence shuffling was used to generate fully constitutive synthetic promoters in mammalian cells[115]. In some applications, inducibility is desired to enable temporal bioprocessing control. To this end, several inducible promoters have been well-established in mammalian cell systems[116] and have spurred the development of synthetic circuits capable of modulating multiple genes. As examples, the cumate gene switch was adapted from microbes and successfully implemented in mammalian systems[117]. Tetracycline-responsive promoters were recently expanded to include aptamer-based control, mediating transgene expression in a small molecule concentration-dependent manner[118, 119]. Furthermore, a biotin-inducible expression system demonstrated direct correlation between biotin concentration and reporter gene expression in both CHO and HEK cells, and gene expression can be triggered even at large scales[120]. In each of these instances, gene expression control (whether constitutive or inducible) is critical for optimizing cellular hosts for protein production.

An IRES permits mRNA translation initiation and increases the flexibility afforded by synthetic and native promoters by allowing for expression of multiple genes using a single promoter and controlling the ratio of protein expression. These elements are particularly critical since the light chain (LC) and heavy chain (HC) of antibodies are expressed as separate genes. This approach was recently demonstrated using IRES-mediated tricistronic vectors, resulting in more positive clones and higher clonal productivity by controlling the expression ratio of LC to HC[121]. The use of an IRES

demonstrates the additional layer of complexity needed for optimizing cellular hosts for antibody production.

Finally, there are additional elements that help expand the toolkit for engineering mammalian cells, including post-transcriptional regulatory elements[122, 123], UCOEs[124-126], and S/MAR[127] elements. UCOEs can help regulate transgene expression by modifying chromatin state and overcoming transgene silencing[124]. The impact of these elements is evinced by increased protein titers[125] and improved transgene stability[126] in cell culture. S/MARs similarly modify chromatin state and were recently used to prolong transgene expression[127] and initiate gene amplification in animal cells[128]. In addition, miRNAs can play a significant role in engineering gene networks and protein production by binding complementary mRNA sequence resulting in reduced translation. For example, by profiling highly conserved miRNA expression in IgG-producing CHO cells, several differentially expressed miRNAs were identified to potentially regulate gene expression, cellular growth and proliferation, and the overall cell cycle[129]. Productivity of CHO cells was increased by exogenously increasing cellular miR-7 levels[130]. As a high level of miRNA sequences are conserved between the Chinese hamster and other mammals[131], this work can likely be extended to other cellular systems. Many of these elements may need to be combined for high and stable expression levels required to further optimize and improve protein production, while they are also additional orthogonal elements to complement the common *cis*-acting genetic elements.

1.4. THE PATH FORWARD

Despite many recent advances in improved synthetic biology techniques, more improvements are needed before cellular engineering of mammalian cells occurs at a level on par with microbial industrial hosts. The growing need for cell culture systems and cell

lines that facilitate rapid screening of pharmaceutical candidates in pre-clinical phases is strongly driven by the desire to reduce both the cost and time required for developing a pharmaceutical. Pivotal studies implementing the tools described above illustrate progress in the field to quickly produce molecules of increasing complexity. Moreover, these genetic tools create precise and predictable changes within a cell and aid in efforts of Quality by Design (QbD) throughout the cell line development process. Chromosomal context and epigenetic mechanisms should also be considered, as they greatly influence genomic architecture[132]. Some of the tools described cause modifications to epigenetic structure, such as S/MARs, IRESs and UCOEs.

In particular, the explosion of targeted genome editing techniques now available can certainly benefit the biopharmaceutical industry from candidate discovery through licensed commercialization. The ability to perform genetic screens in model organisms and even representative human cell lines should improve the chances of developing a successful therapeutic. Identifying and targeting loci with high transcriptional activity for transgene expression in production hosts can improve product yields, ultimately reducing manufacturing costs. Another laborious and time-consuming step in developing mammalian cell lines is isolating highly productive clones through selection, and adapting the work by Jiang, *et al*.[133] could expedite this process.

By adapting advances in synthetic biology from microorganisms to complex mammalian cell systems and developing additional pioneering tools, it is possible to produce a wider range of specialty compounds beyond recombinant proteins. Advanced genome editing tools, in conjunction with novel, synthetic genetic elements, can be used to predictably design mammalian cell systems for industrial biotechnology. However, some of the genetic elements described above, particularly synthetic promoters, could exhibit distinctly different behavior depending on the cell type under investigation[113, 126].

Even combining these elements may not successfully mitigate this challenge[126]. Yet, these combinations can bestow beneficial effects upon other important considerations in mammalian cell engineering, such as transgene silencing[125]. As a result, more effort is required in the field to further expand and adapt these synthetic parts for broad utility and application. Collectively, these prospective advances help mitigate the great complexity and lack of tools currently impeding cell line development processes. High productivity, high stability, and the ability to adapt cells for continuous culturing will continue to drive goals and targets for future cell lines. Advances in the area of synthetic biology outlined here will ultimately minimize the cost burden associated with pharmaceutical development of both novel therapies and biosimilars of the future.

Therefore, the work presented in this dissertation represents a significant step forward in developing and characterizing the tools to precisely regulate gene expression in mammalian cells. These discrete and predictable levels of gene expression are critical in complex applications beyond simple gene overexpression for production, such as metabolic engineering immune cell engineering applications with intricate gene expression networks. As such, mammalian cells will continue to be a valuable cell factory and model organism for addressing medical needs in the future.

In particular, this dissertation addresses how we can precisely regulate gene expression through directing transgene integration into prescribed loci, exploring a strategy to improve integration into target loci, and modulating mRNA abundance. To combat the issue of random integration of transgenes, Chapter 2 describes an approach to pre-catalogue preferential integration loci for high transcriptional activity. With these established loci, Chapter 3 explores a strategy to improve successful integration events into the target loci. Chapters 4 and 5 evaluate orthogonal approaches to promoter engineering for driving

transcription in mammalian hosts, while Chapter 6 exploits the conserved 3' UTR structure to create a set of endogenous and synthetic terminators for tuning expression.

Chapter 2: Placing the message – establishing transgene integration loci[2]

2.1. CHAPTER SUMMARY

Mammalian cell line development requires streamlined methodologies that will reduce both the cost and time to identify candidate cell lines. Improvements in site-specific genomic editing techniques can result in flexible, predictable, and robust cell line engineering. However, an outstanding question in the field is the specific site of integration. Here, we seek to identify productive loci within the human genome that will result in stable, high expression of heterologous DNA. Using an unbiased, random integration approach and a green fluorescent reporter construct, we identified ten single-integrant, recombinant human cell lines that exhibit stable, high-level expression. From these cell lines, eight unique corresponding integration loci were identified. These loci are concentrated in non-protein coding regions or intronic regions of protein coding genes. Expression mapping of the surrounding genes revealed minimal disruption of endogenous gene expression. Finally, we demonstrated that targeted *de novo* integration at one of the identified loci, the 12th exon-intron region of the *GRIK1* gene on chromosome 21, resulted in superior expression and stability compared to the standard, illegitimate integration approach at levels approaching 4-fold. The information identified here along with recent advances in site-specific genomic editing techniques can lead to expedited cell line development.

[2] The content in this chapter can be found in a previously authored publication. JKC and AML equally contributed to the experiments and analyses, and collectively wrote the manuscript comprising this chapter. Reprinted with permission from Cheng, J. K., Lewis, A. M., Kim, D. S., Dyess, T., & Alper, H. S. (2016). Identifying and retargeting transcriptional hot spots in the human genome. *Biotechnol J*, 11(8), 1100–1109. DOI: 10.1002/biot.201600015.

2.2. INTRODUCTION

Cellular hosts can produce a wide range of useful products including therapeutics, antibiotics, biofuels, specialty chemicals and other small molecules[6, 11, 134]. In particular, mammalian cell lines such as CHO, HeLa, HEK293 and HT1080 are important industrial hosts commonly used to produce protein therapeutics, such as insulin, antibodies, cytokines, enzyme replacement therapies, and growth factors[1]. Specifically, three globally, commercially available therapeutics are produced in the human-derived cell line HT1080[135], with additional in the pipeline. However, these cell lines remain harder to engineer than bacterial and fungal counterparts. Precision genome engineering tools together with a synthetic biology paradigm have fueled a renaissance for strain engineering[102, 136], yet, these tools inherently require knowledge of the genomic architecture. As a result, most current strain development programs that rely on stable cell line development (CLD) often include time-consuming, labor-intensive, repetitive and expensive screening of thousands of potential cell lines[7, 137-139]. Consequently, site-specific integration of transgenes into pre-characterized loci would speed the CLD process and yield significant cost and time-savings.

Over the past decade, the genome editing toolbox for mammalian cells has rapidly expanded[26]. Technologies including Cre and FLP recombinase[41, 140], ΦC31 integrase[141, 142], zinc-finger nucleases[27, 143], transcription activator-like effector nucleases[59, 68], and more recently, clustered regularly interspaced short palindromic repeats (CRISPR) systems[144] have enabled new ways to integrate transgenes. However, two commonalities arise for each of these methods: (1) the need to create double-strand breaks in DNA to mediate non-homologous end joining and homology-directed DNA repair and (2) the need to pre-determine high expression loci. Moreover, commercialization and continued improvement

of these methods demonstrates their importance for varied applications including CLD[145, 146].

While efficiency of integration (both re-integration and *de novo* targeting) has increased, it is still unclear what genomic loci are best for high expression. Specifically, not all genomic loci are equal with respect to their capacity to facilitate and stably maintain high levels of transgene expression. The importance of integration sites has been well established[7, 137, 139, 147, 148], however, limited information is available about desirable sites with only a few characterized in particular cell types with interesting characteristics[149, 150]. Variations in expression across the genome has been demonstrated in other model organisms including *E. coli*[151], *S. cerevisiae*[152] and zebrafish[153] with upwards of 8-fold difference in expression levels for yeast[152]. In the absence of similar studies for mammalian cells, pre-determined criteria, such as 'Good Safe Harbours'[154] have been applied *a priori* to identify potentially useful integration sites[155]. Outside of such experimental genomic searches, global expression sets have limited utility since non-coding regions can serve as good sites for high-level transgene expression. Many commercial technologies for site-specific integration continually exploit a small number of integration loci. However, little consideration has been given to the nature of these sites and only a small number of exonic sequences are used[156, 157], thus ignoring a large portion of the protein coding and all non-coding regions.

The phenomenon of genomic hot spots is ubiquitous as cell biology processes including meiotic recombination[158], epigenetic modifications[159, 160], viral integration[161-164], chromatin structure[165, 166], and transcriptional capacity[165] exhibit a loci-dependent bias with a non-random distribution across the genome[167]. Additionally, retroviral integration occurs at a non-random frequency across the genome and exhibits an integration bias with defined motifs and preference for CpG islands, regions of high gene density, and regions near

transcription start sites and transcription factor binding sites[163, 168-171]. These sites are important since they indicate natural propensities for viral integration as well as provide a mechanism for viral-assisted gene therapy.

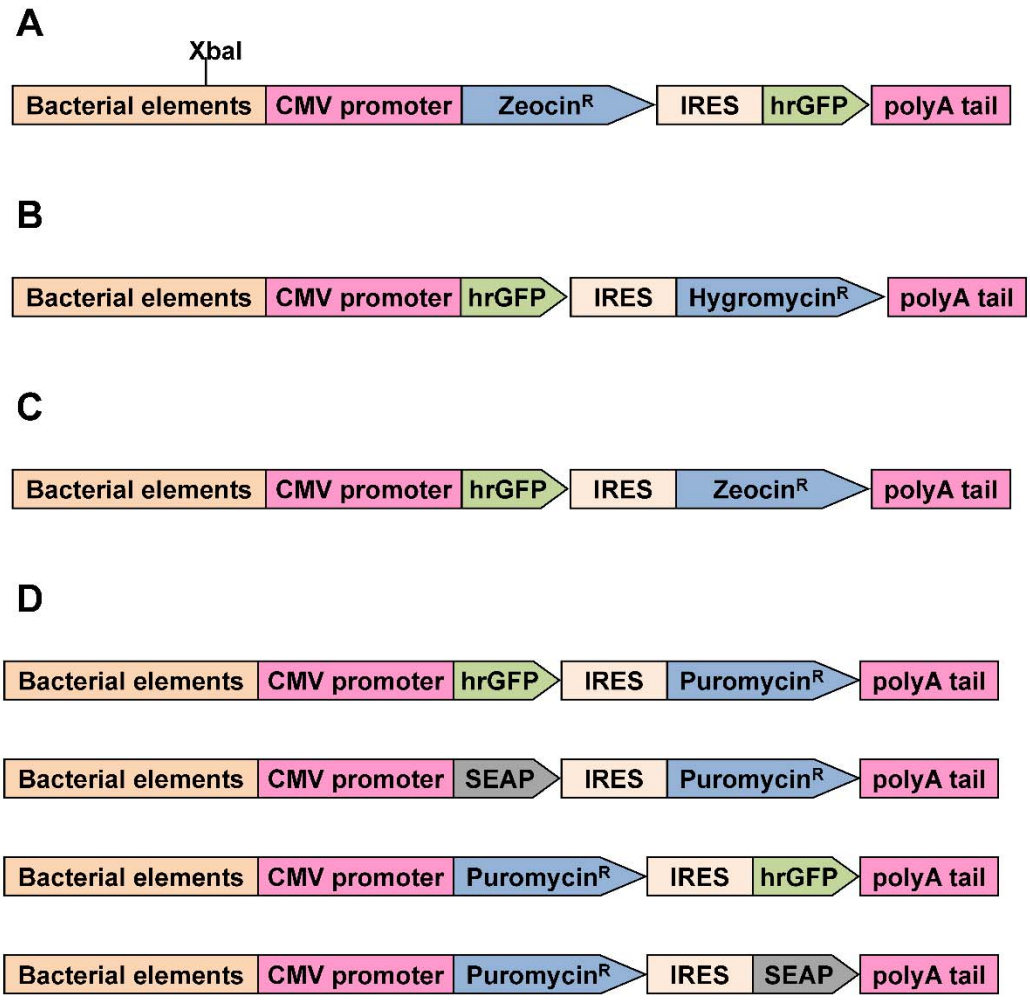
At first approximation, integration into euchromatin, lightly packed gene-rich regions, is most likely to favor expression whereas integration into heterochromatin is unlikely to confer transcription capacity, as these regions are often silenced by histone deacetylation, histone methylation and promoter methylation[137]. Recombinant protein production in the HEK293 and CHO Flp-In host cell lines highlights the importance of targeting an integration site within euchromatin and the benefit of site-specific integration[172, 173]. In addition, multiple integration copies can be specifically targeted into the same locus using Cre recombinase and mutant l*oxP* sites for gene amplification[174]. However, these case studies required the development of these specific host cell lines with FRT or *loxP* targeting sites in the desired, pre-described locus. Furthermore, the integration of a transgene may affect chromatin structure and thus change the transcriptional capacity for a given locus. Thus, a cataloging of regions within the genome that can enable and maintain stable, high-level transcription of a transgene (so called 'transcriptional hot spots') will serve as a great coupling to emerging genome editing technologies. To this end, we seek to identify transcriptionally active areas using a genome-wide screen in the HT1080 mammalian host, demonstrate improved expression and stability of these sites compared to illegitimate integration, map the surrounding expression landscape, and demonstrate the capacity for precision re-targeting to this loci. In doing so, the work in this chapter addresses an unmet need at the forefront of human genome research, cell line development, and biologics research by providing a catalogue of genomic hot spots supportive of high-level, stable transgene expression.

2.3. RESULTS AND DISCUSSION

2.3.1. Establishing single-cell clones with stable and high expression capacity

Initially, we conducted an unbiased survey of the human genome to identify genomic loci that afforded stable, high-level heterologous gene expression. To do so, a random integration strategy was used in conjunction with a transgenic reporter construct (**Figure 2-1**) to explore the entire genome. These constructs contained both antibiotic selection markers and fluorescent reporter genes (GFP) expressed with the CMV promoter. The human sarcoma cell line, HT1080, was transfected with a linearized pIRES-hrGFP reporter construct (**Figure 2-1A**) and subjected to a sequential selection first by Zeocin™ selection (previously demonstrated as superior in establishing recombinant human cell populations[25] followed by GFP expression. Expression of the GFP reporter gene was measured using flow cytometry. Bulk populations exhibited a broad range in expression, as evident by high coefficients of variance, with higher enrichment of mean GFP expression coinciding with more stringent antibiotic selection (**Supplemental Figure S1a** of the publication[175]). We pursued this work in the human cell line HT1080, as opposed to another common industrial host such as CHO, since the CHO genome was unavailable at the time of these experiments[176, 177] and the human genome was well-annotated, facilitating loci identification.

Figure 2-1: Dual-selection transgene constructs for high expression clones.



HT1080 cells were transfected with two heterologous constructs, each containing a single promoter and IRES to allow for simultaneous expression of two genes that enable dual expression. (**A**) the pIRES-hrGFP construct contains the Zeocin resistance gene in the first cistron and a human optimized GFP gene in the second cistron. (**B**) the pHL-GFP contruct contains a GFP gene in the first cistron and the puromycin resistance gene in the second position. (**C**) the pAML-Zeo construct was used for retargeting of the 12th exon of the GRIK1 gene on chromosome 21. (**D**) the hrGFP and SEAP expression cassettes used for de novo integrations.

Recombinant populations established using 100 and 250 ug/mL Zeocin™ were further enriched using FACS sorting to select the top 10-15% of GFP expression. The expression profiles of the resulting sorted populations (demonstrating enrichment) are

shown in **Supplemental Figure S1b** of the publication[175]. Using a similar approach, hygromycin-resistant single cell clones were established by Shire Human Genetic Therapies using the pHL-GFP reporter construct (**Figure 2-1B**) and cells were enriched using FACS. In both cases, dilution cloning was performed to isolate single cell clones and establish homogenous populations. In an effort to identify stable clone lines, the expansion of single cell clones, which took place over a period of 6-8 weeks, was performed without any antibiotic selection. Following expansion of both populations, stable GFP expression and transgene copy number was evaluated. This combined effort (both the Zeocin™ and hygromycin-based clones) resulted in a total of ten clones with single site integration (single transgene copy), high geometric mean GFP fluorescence (**Figure 2-2A**), stable expression, and high GFP RNA expression according to RT-PCR (**Figure 2-2B**). Each of these clones had a stable expression profile and mRNA levels that are very high relative to average expression of human genes (in particular, compare native gene expression with transgene in **Figure 2-3**). Even so, we do observe a difference in rank order for clone expression using these two quantitation methods of flow cytometry (geometric mean GFP) and RT-PCR (relative gene expression level). We noted that absolute fluorescence values did not always correlate with mRNA expression which indicates other influences in the cell. Flow cytometry was performed using live cells, and could be influenced by cell morphology, culture viability, and day-to-day instrument variability. Real-time PCR is a highly sensitive *in vitro* assay that measures transcript level (a level that does not always correlate precisely with protein abundance). Nevertheless, we do see a similar range of expression (10 log-fold) between the highest and lowest expressing clones across these two methods.

Figure 2-2: Isolated single cell clones exhibit high protein and mRNA expression.

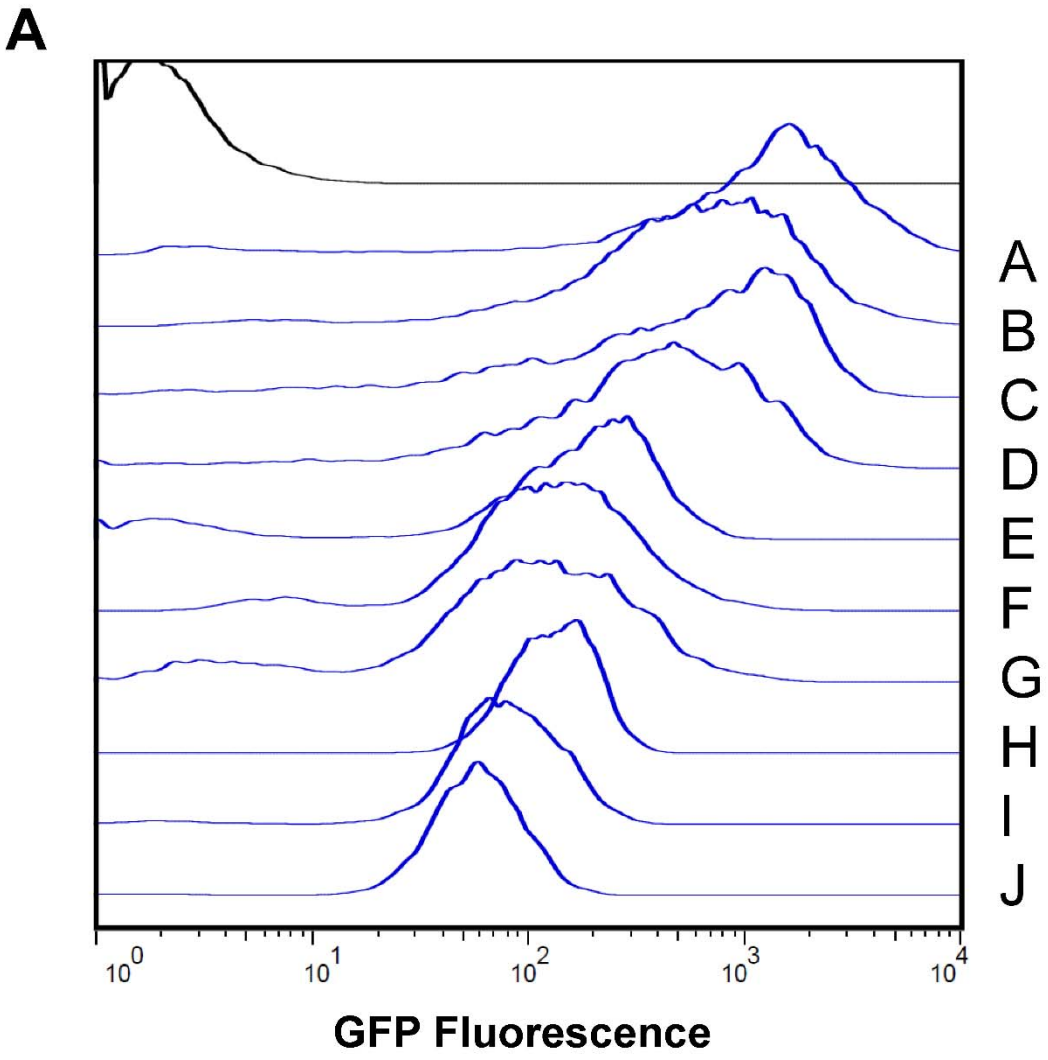
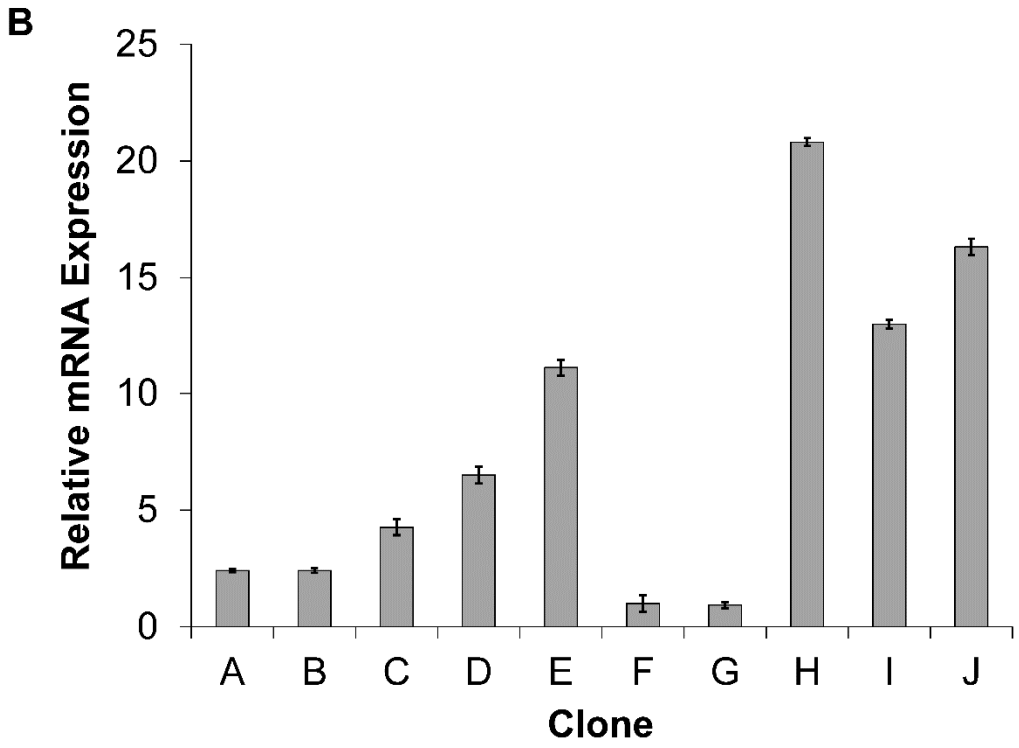


Figure 2-2, continued:



Ten single cell clonal populations were isolated from the recombinant populations and protein (**A**) and mRNA expression (**B**) were measured. **A.** GFP expression profiles for clonal populations (A-J, labeled based on approximately descending geometric mean fluorescence) were measured using flow cytometry and are shown in blue compared to untransfected HT1080 (black). **B.** Relative mRNA expression (±SD) of the clonal populations (A-J) was measured by RT-PCR for the first cistron, hrGFP. mRNA expression levels are normalized arbitrarily to clone F for comparison purposes.

Figure 2-3: mRNA expression maps for protein-coding sequences surrounding integration loci.

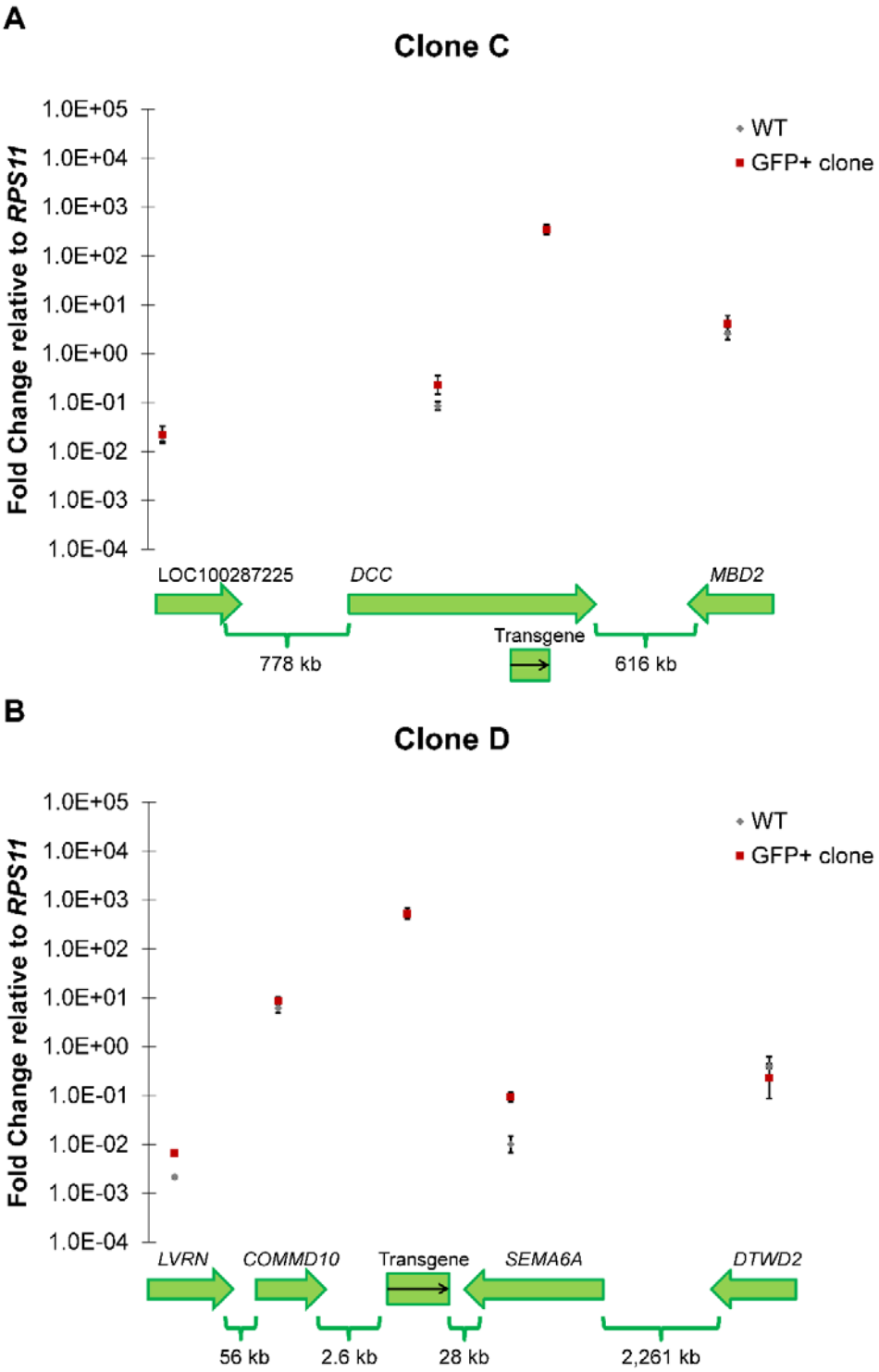


Figure 2-3, continued:

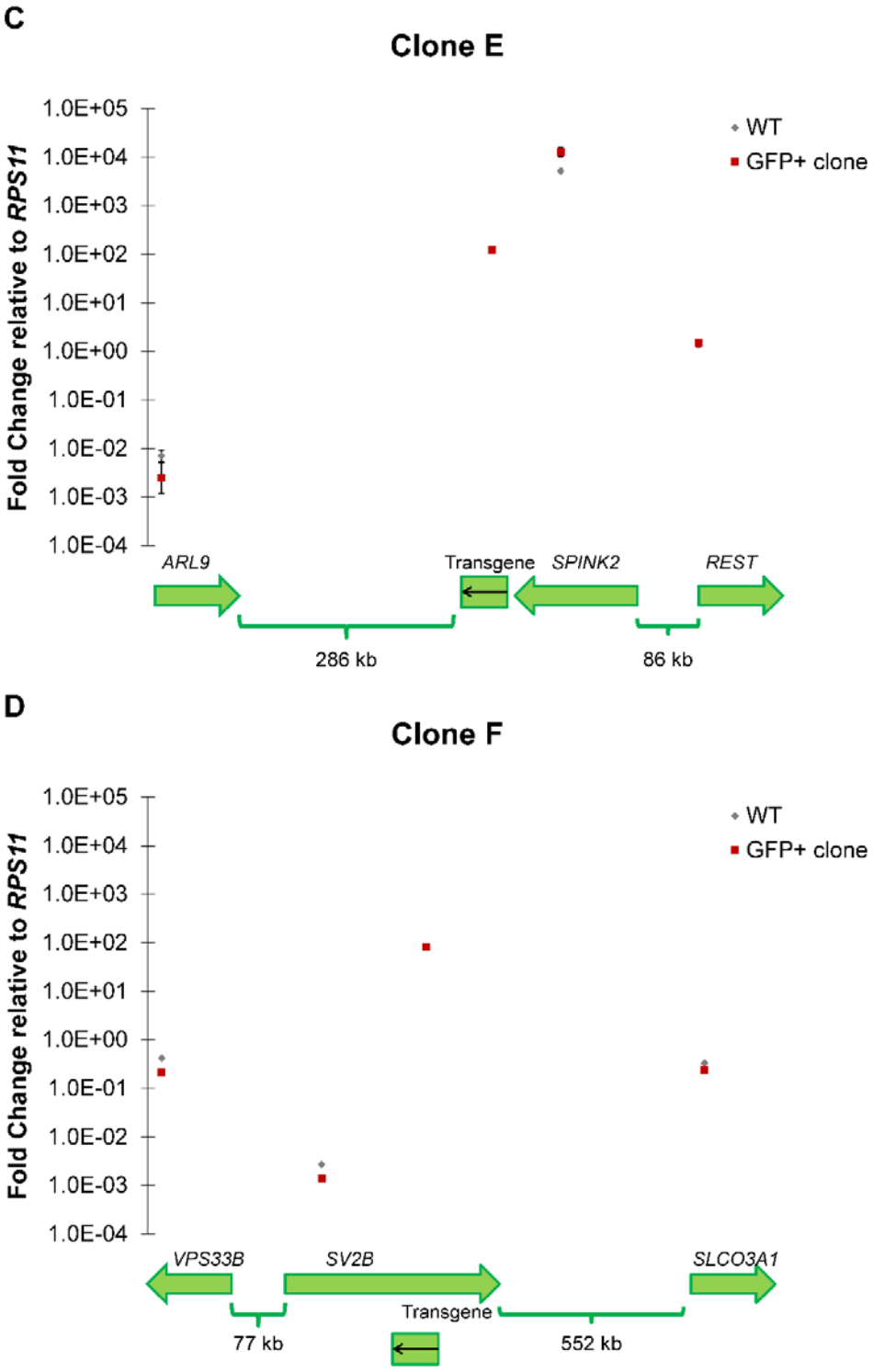


Figure 2-3, continued:

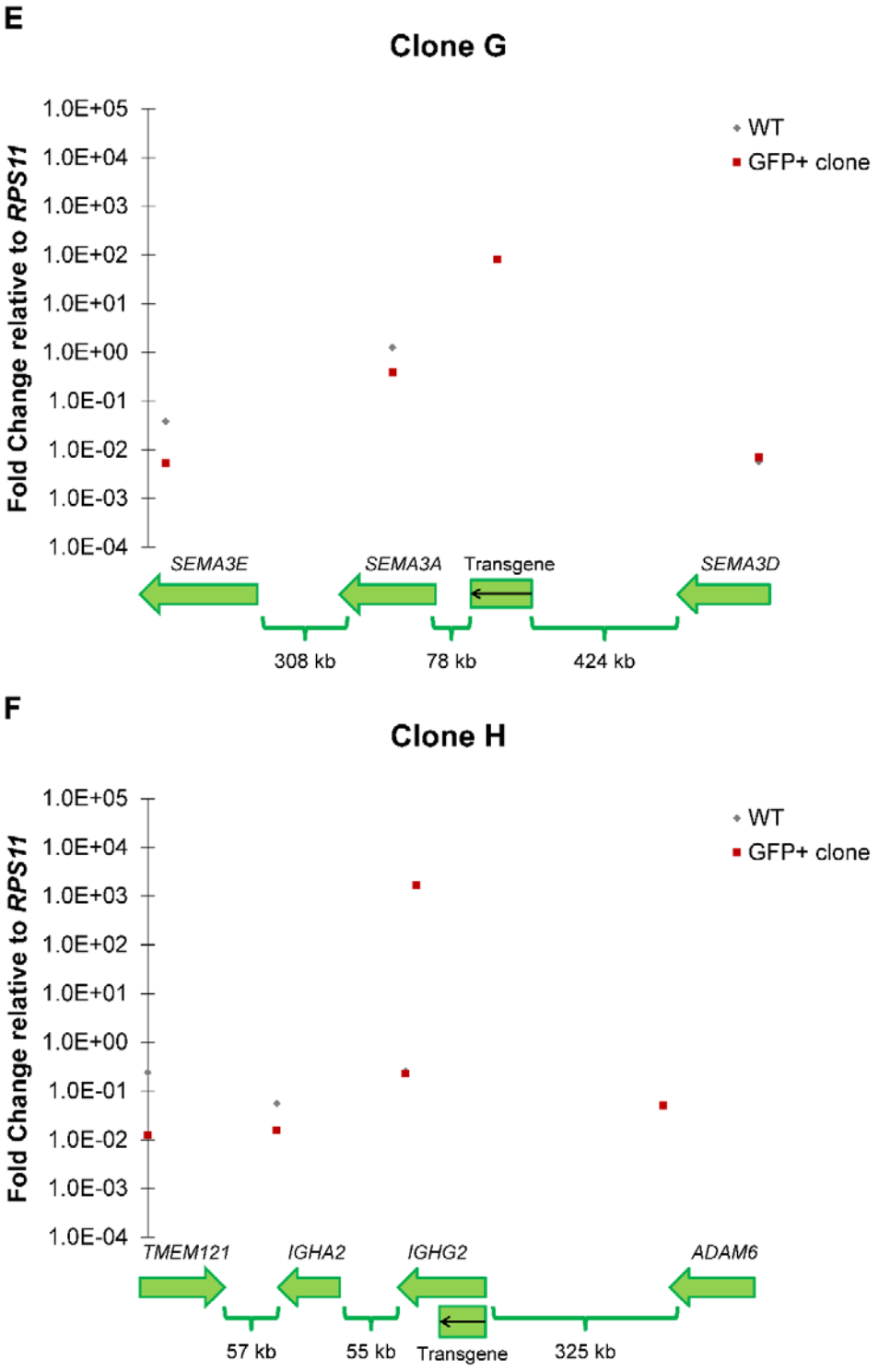
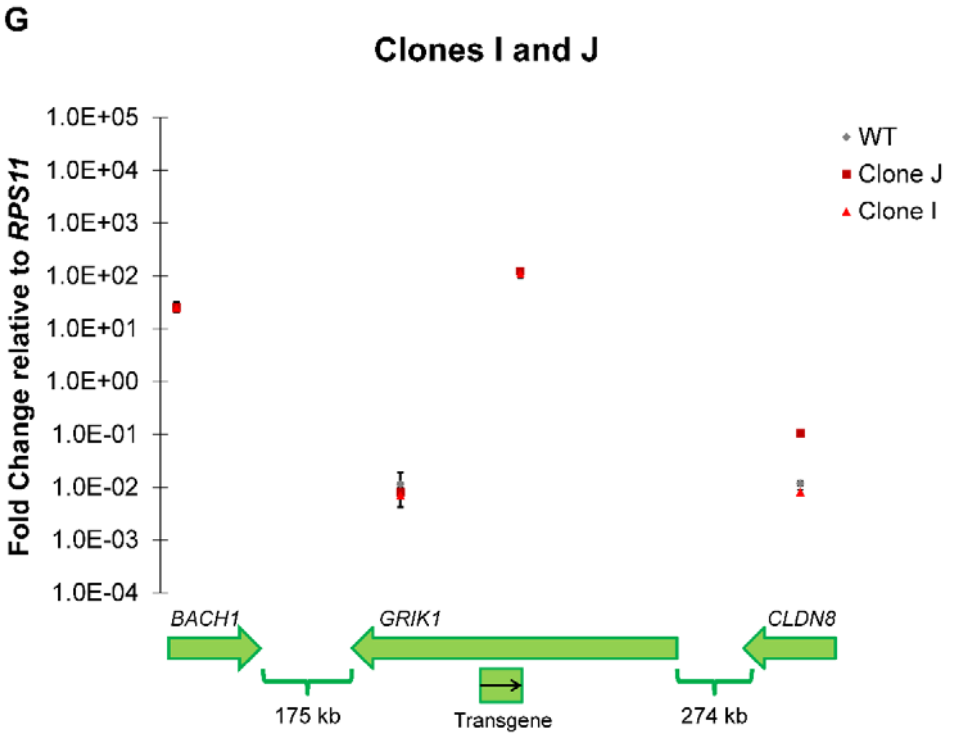


Figure 2-3, continued:



Fold change in mRNA expression (±SD) was measured in surrounding protein-coding genes relative to the house-keeping gene *RPS11* (an endogenous gene of high expression) for the wild-type cell line and the cell line with integrated transgene. **A.** mRNA expression profile for clone C on chromosome 18, including the transgene and endogenous genes *DCC*, *MBD2* and uncharacterized locus 100287225. **B.** mRNA expression profile for clone D, integrated on chromosome 5, including the transgene, *AQPEP*, *COMMD10*, *SEMA6A* and *DTWD3*. **C.** mRNA expression profiles for clone E, integrated on chromosome 4, including the transgene, *ARL9*, *SPINK2* and *REST*. **D.** mRNA expression profile for clone F, integrated on chromosome 15, including the transgene and endogenous genes *VPS33B*, *SV2B* and *SLCO3A1*. **E.** mRNA expression profile for clone G, integrated on chromosome 7, including the transgene, *SEMA3E*, *SEMA3A*, and *SEMA3D*. **F.** mRNA expression profile for clone H, integrated on chromosome 14, including the transgene, *TMEM121*, *IGHA2*, *IGHG2* and *ADAM6*. **G.** mRNA expression profiles for clones I and J integrated on chromosome 21, including the transgene and endogenous genes *BACH1*, *GRIK1* and *CLDN8*. The black arrow indicates the promoter direction for the transgene. Error bars indicate standard deviation from RT-PCR triplicates.

2.3.2. Identification of transcriptional hot-spots in isolated, high-expressing clones

Identifying the exact location of an integration event is greatly aided by high-throughput sequencing techniques. However, given the size of the human genome, this is an expensive approach. Therefore, we employed a variety of low-throughput PCR-based methodologies, including TAIL PCR, inverse PCR and plasmid recovery based on genomic DNA to identify the integration loci for each of these ten clones. The integration site was confirmed by PCR (further details are provided in the **Supplementary Material** of the publication[175] for each clone).

Using these PCR-based approaches, we were able to identify the integration loci of our ten GFP-expressing clone lines (**Table 2-1** and **Supplemental Table S1** of the publication[175]). This information showed that the integration loci are distributed throughout the human chromosomes with half the integration events occuring in gene intronic regions. The remainder of the integration events occurred further from the nearest protein-coding regions, with two integrations found in long non-coding RNA regions. This is a surprising result and clearly demonstrates that regions outside of protein-coding sequence are hospitable towards heterologous transgene expression. The clone with the highest mRNA expression levels had an integration in chromosome 14 in the *IGHG2* gene. This region of the genome is rich in immunoglobulin proteins, which are spaced close together, yet not a region of particularly high expression for this particular cell line. Finally, two integration sites were each identified from duplicate, independent integration events. Clones I and J both arose from integration into the *GRIK1* gene on chromosome 21 at the 13th intron and 12th intron respectively, and clones A and B both arose from integration into an unplaced genomic contig. Unfortunately, very little information is available about this genomic contig, including its chromosome, because this is a region of

high redundancy. Nevertheless, 8 unique genomic loci were identified from the 10 stable, high-expressing clones analyzed here.

Table 2-1: High transcription integration loci are distributed throughout the genome.

Clone	Chr.	Intron	Nearest Gene	Nearest Gene Function
A			Unplaced genomic contig (3980bp)	
B			Unplaced genomic contig (3980bp)	
C	18	26	*DCC*	Netrin 1 receptor
D	5		*SEMA6A,* 28kb downstream	Transmembrane domain
E	4		*SPINK2,* 9kb upstream (in LOC105377668)	Serine peptidase inhibitor
F	15	1	*SV2B*	Synaptic vesicle glycoprotein
G	7		*SEMA3A*, 78kb downstream (in LOC101927378)	Secreted neuronal protein
H	14	3	*IGHG2*	Immunoglobulin heavy constant
I	21	13	*GRIK1*	Glutamate receptor, neural
J	21	12	*GRIK1*	Glutamate receptor, neural

From ten stable, high expressing clones, we identified eight integration loci using PCR-based low-throughput methodologies. Each site was confirmed using primers matching the transgene and genomic locus, which produced a positive band but lack of band with wild-type gDNA. Each locus is discussed in detail in **Supplementary Material** of the publication[175].

2.3.3. Expression mapping of hot spot loci reveal influence and impact of surrounding genes

Next, we sought to evaluate the expression profile of surrounding protein coding regions in the various clones before and after transgene integration to determine both the benefits to transgene expression provided by the surrounding genomic DNA, as well as perturbations that may be caused by integration. Perturbations are specifically important in gene therapy applications, where 'harmless' integration loci must be chosen such that surrounding genes, especially oncogenes, are not inadvertently impacted. Previous studies have demonstrated both modes of action with transgene expression[155, 178].

In this experiment, expression levels of protein coding genes were determined using RT-PCR with whole cell RNA for both the GFP-positive clone and wild-type HT1080. Expression of each gene was compared to ribosomal protein 11 (RPS11), a common human

housekeeping gene that is highly expressed at levels 4000-fold higher than the average gene (ranking it as the 117th most highly expressed gene based on a microarray study conducted by Shire Human Genetic Therapies).

The resulting expression maps for all clones (excluding those integrated in the unplaced human genomic contig for which no information is available) are shown in **Figure 2-3**. Universally, it is seen that protein coding sequences distantly located from the site of integration exhibit little to no difference between wild-type and GFP-positive clone transcripts indicating minimal expression perturbation caused by transgene integration. Here, we define minimal expression perturbation as being less than a 2-fold change in gene expression level. Minimal, local transcriptional expression disruption is observed for clones C, E, F, and I (**Figures 2-3A, 2-3C, 2-3D** and **2-3G**). For the case of clone D (**Figure 2-3B**), expression in the GFP-positive clone of neighboring gene *SEMA6A* is elevated compared to the wild-type. Finally, with the exception of the integration site for clone D and E, we see that expression of the transgene is significantly elevated relative to the surrounding genes, which in most cases are lowly expressed. Thus, while these sites can enable high-level transcription, the transgene cassette is not simply hijacking a region of high transcription. Collectively, these results indicate that the identified hot spot loci are indeed good integration loci with minimal impact to the expression of local genes.

2.3.4. De novo targeting into *GRIK1* 12th exon/intron enables improved transgene expression

Finally, we sought to demonstrate the impact of combining these high-transcription loci with site-specific genomic targeting techniques to speed the process of CLD and serve as an alternative to clonal screening. We opted to use the CRISPR system which was recently demonstrated to be a flexible, highly efficient method for mammalian genome editing[144], although delivery of large constructs via this method have not been previously

shown in human cell lines. For this test, we selected the 12th exon-intron region of the *GRIK1* gene on chromosome 21 as the target of choice as it exhibited high expression and was in a region identified independently in two clones. Furthermore, site-specific integration at this locus was confirmed using primers found in **Supplementary Material, Table S2** (61-63, 84-87) of the publication[175].

Initial tests to reconstitute high GFP expression with different transgene arrangements were conducted in HT1080 cells by delivering the pPG and pGP expression cassettes (**Figure 2-1D**) to this locus. Targeted integrations to this *GRIK1* locus were performed by transfecting the guide RNA (gRNA) construct along with the hCas9 and hrGFP/SEAP cassettes. The gRNA in this case encodes a crRNA-tracrRNA fusion transcript driven by the U6 polymerase III promoter modified to include a specific 23-nucleotide region of homology to a distinct region in this locus. This gRNA design was selected to minimize off-targeting effects using the criteria outlined by previous researchers[144]. Seventy-two hours after the transfection, cells were subjected to selection at MIC_{75} levels until viability recovered to greater than 90%.

Compared to the controls with random integration, the targeted transfections to the *GRIK1* locus in HT1080 resulted in a roughly 1.3- to 1.4-fold increase in GFP expression levels as measured via flow cytometry, at the stable bulk level ($p < 0.001$ when comparing their geometric mean difference) (**Supplemental Figure S2a and S2b** of the publication[175]). Moreover, the population histograms are indicative of an overall shift in high expression, not a clear sub-population for these two transgene arrangements (**Supplemental Figure S2c** of the publication[175]). To demonstrate the true potential of this site-specific integration, we isolated single cell clones with confirmed integration into the loci using a PCR-based approach[179]. In this case, site-specific integration and isolation demonstrated a 3.1- to 3.9-fold improvement in geometric mean fluorescence (**Figure 2-**

4A) when measured between 72 days and 110 days post-selection ($p < 0.001$ when compared to several clones with random integration for both constructs). Based on the clones isolated from limited dilution cloning with our transgene integrated in the *GRIK1* locus, we estimate that the targeted integration efficiency is roughly <1% with the gRNA construct used. These efficiencies are lower than previously reported values for other human cell types, but this is indeed the largest construct (6-kb) attempted for integration into the human genome[144], despite recent work reporting a ~5-kb integration into HEK293 and CHO with comparable efficiencies[180] using the CRISPR/Cas system.

Finally, we sought to demonstrate the capacity of site-specific integration into hot-spot loci for a model secreted protein, SEAP, at the *GRIK1* locus to verify that the improvements observed were not dependent/associated with the gene expressed. Similar to the hrGFP study, HT1080 control cells were transfected with the pPS and pSP expression cassettes (**Figure 2-1D**) with similar controls. Cultures were maintained for up to 120 days to account for any silencing that may occur in both populations (**Supplemental Figures S2d-S2f** of the publication[175]). On average, the SEAP productivity was 1.3- to 1.4-fold higher for the targeted integration pools over that of random integration (**Supplemental Figure S2g** of the publication[175]) accounting for differences in heterogeneous bulk populations (**Supplemental Figure S2h** of the publication[175]). This improvement is similar to the benefit in hrGFP expression observed when using targeted integrations at the *GRIK1* locus. Clonal populations were once again isolated and confirmed using PCR, and the site-specifically integrated clone exhibited a 2.4-fold increase in productivity ($q_{p,random\ clones} = 0.00956 \pm 0.00298$ compared to $q_{p,specific\ clone} = 0.0226 \pm 0.0015$, 95% CI, $p < 0.001$, confirming the increase in heterologous gene expression by specifically integrating a transgene into the *GRIK1* locus (**Figure 2-4B**). It should be noted that while the total

production of SEAP is low in this cell line, this is consistent with previous expression of this protein in this cell line.

Figure 2-4: Targeted integration into the GRIK1 loci results in elevated hrGFP and SEAP expression.

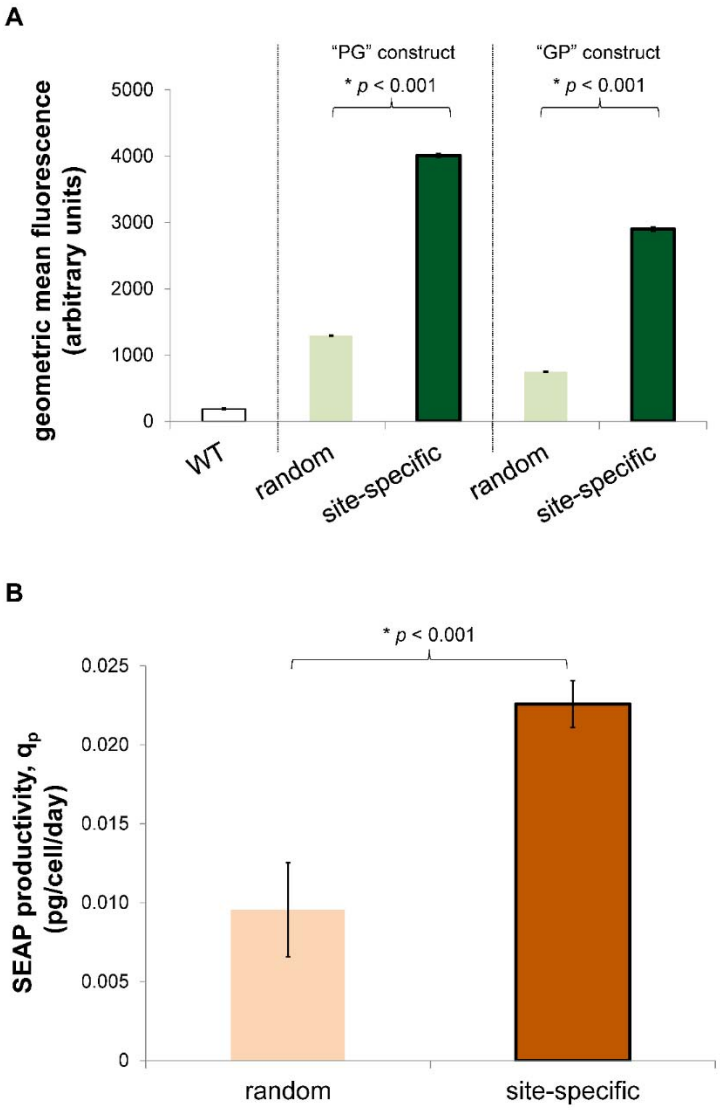


Figure 2-4, continued:
A mammalian expression cassette expressing hrGFP or SEAP and puromycin was transfected into HT1080 cells in a random (control) and targeted fashion (Grik1B) using the CRISPR/Cas system. Following antibiotic selection, heterogeneous populations were evaluated and subsequently derived clonal populations by limited dilution cloning. **A.** Flow cytometry was used to measure GFP expression. Geometric mean fluorescence values of isolated hrGFP-expressing clones show a clear increase in GFP expression upon targeted integration into a transcriptional hot-spot. Error bars represent 95% confidence interval of geometric mean from 3 clones (pPG) or 4 clones (pGP) isolated by limited dilution cloning. **B.** SEAP productivity (pg/cell/day) of clonal (isogenic) populations transfected with the pSP expression cassette. Site-specific integration into the transcriptional hot-spot in the *GRIK1* locus supports a 2.4-fold increase in productivity when compared to random integration of the transgene. Error bars represent 95% confidence interval of mean production from 3 clones (random, pSP) and 1 site-specifically integrated clone (pSP).

2.4. CONCLUSIONS

By using an unbiased genome integration approach, we identify ten recombinant human cell lines with stable, single-copy, high-level heterologous gene expression and subsequently identify the corresponding integration loci. These results indicate the importance of non-protein coding regions for heterologous gene expression, despite the fact that previous studies have focused exclusively on exonic regions. Expression maps for each of these loci demonstrate negligible perturbations caused to the surrounding genes. Finally, we demonstrate that *de novo,* targeted integration at one of the identified loci, the 12th exon of the *GRIK1* gene on chromosome 21, results in superior expression compared to the standard, illegitimate integration approach. This work provides a much needed cataloging of potential genomic hot spots that can be linked together with emerging genome editing tools. Targeting these advantageous integration loci can significantly reduce the time, labor and materials associated with CLD. Additionally, this approach (and specific targets identified) can be extended to other mammalian cell lines used for industrial protein production, including CHO and HEK293 and help speed CLD. Critically, this work

demonstrates that the integration locus is an important component of our cellular engineering toolbox that we can leverage for fine-tuning gene expression.

Chapter 3: Inserting the message – characterizing transgene integration selectivity

3.1. CHAPTER SUMMARY

Due to the low success rate of identifying clonal populations with the transgene integrated at the correct locus in the work described in Chapter 2, we explored a cell line development strategy to improve upon that rate. Our strategy will offer significant enrichment of the transgene containing population and simultaneously isolating integrations at the correct locus within that population by combining both positive and negative selection. The characterization of this strategy described here focuses on using the CRISPR/Cas9 system for directing specific double strand breaks (DSBs) to facilitate the targeted integrations in mammalian host cell lines. First, we confirmed the appropriate concentrations required for effective selection by determining MIC_{75} for each selection agent in each cell type. Subsequently, adopting a co-targeting approach[181] enabled us to validate gRNA designs for targeting a specific locus and characterize the frequency of DSB formation and repair at that locus by Sanger sequencing and quantitative PCR. The work described in this chapter sets the foundation for the full characterization of our cell line development strategy.

3.2. INTRODUCTION

Transgene incorporation into the mammalian genome is routinely achieved through random integration facilitated by innate DNA repair mechanisms[178]. This transgene integration process can be fundamentally divided into two steps: 1) the generation of a double strand break (DSB) and 2) the repair of the resulting break by homology-directed repair (HDR), microhomology-mediated end-joining (MMEJ), and/or non-homologous end-joining (NHEJ)[182-184]. Since these DBSs can occur randomly in the host cell genome, the random integrations would require extensive screening to identify suitable transgene-

expressing cell lines, increasing the time and material resources for this process[185]. One approach to expedite this process was to establish pre-engineered cell lines with recombination/recognition sites for future integrations[40, 41, 186-188]. Alternatively, integration vectors derived from lentivirus or adenovirus can offer configurable directed integration[189, 190] but with rather poor efficiency. However, these pre-established sites may not be ideal depending on the application as we have shown in Chapter 2 that the integration site itself can significantly impact gene expression, limiting the utility of these engineered cell lines.

We can exploit recent nuclease technologies such as ZFNs, TALENs, or the CRISPR/Cas9 system in order to programmatically introduce DSBs for transgene integration through the same innate DNA repair mechanisms[191-193]. These nucleases offer directed DSB generation based on the DNA target, although they can be either DNA-encoded (ZFN and TALEN) or RNA-guided (CRISPR/Cas9). Due to this mode of target recognition, it is plausible that minimal mismatches to the target sequence would signal for nuclease activity, leading to off-target DSBs and adverse consequences. Previous work indicated that the guide-RNA (gRNA) required for CRISPR/Cas9 functionality can influence DSB frequency at the target site, even though different gRNAs were designed based on the same protospacer motif (PAM) corresponding to a particular Cas9 enzyme[64, 144, 194]. Cleverly, by using a co-targeting approach to the hypoxanthine phosphoribosyltransferase (HPRT) gene and selection with 6-thioguanine (6-tG), the selected population was substantially enriched for DSB and repair activity at the target loci in a variety of mammalian cells[181]. Therefore, this co-targeting approach can be incorporated into a cell line development strategy and/or used to screen gRNA designs.

To further characterize the CRISPR/Cas9 system, the Cas9 nuclease derived from *S. pyogenes* (hereafter SpCas9) was critically evaluated for its on-target activity and off-

target effects[72, 146, 195-198] and subsequently, efforts to improve on-target activity were reported by engineering the enzyme itself[93, 94, 199]. Alternatively, functional Cas9 enzymes derived from other species, such as *N. meningitidis*, recognize different PAMs and can be used to reduce off-target effects[200]. Improvements to the Cas9 enzymes could certainly improve the generation of desired DSBs at the target locus with minimal cuts elsewhere in the genome, but the nuclease choice and its fidelity should not impact the innate DNA repair mechanism itself. However, it was recently reported that the repair of DSBs induced by the CRISPR/Cas9 system is highly dependent on the protospacer sequence in the target site and is not purely random[201]. Therefore, by shifting DNA repair towards HDR from the preferred NHEJ in mammalian cells[202], it would be possible to favor more integrations at the desired locus.

When the ability of the CRISPR/Cas9 system to mediate transgene integration was specifically investigated in CHO cells, the integration rate for a 3.7-kbp transgene varied between 7-28% depending on the target locus[179]. The integration of an ~5-kbp transgene by NHEJ were rather low in HEK293 cells (0.17%) and CHO cells (0.45%) however[180], indicating that these success rates are highly variable across target loci and even within the same cell types. Other efforts tested the coupling of ZFN mRNA delivery[146, 203] or CRISPR/Cas9 mRNA delivery with adeno-associated virus vectors[99] to achieve reasonable levels of genome editing in a variety of key mammalian cell types.

Ultimately, the interplay between the ability to generate precise on-target DSB and effective repair of the break with the transgene donor will dictate the integration process. Yet an efficient integration process still requires enrichment of the transgene containing and expressing population, and a stringent selection system could easily facilitate this enrichment[204]. Unfortunately, a highly selective cell line development strategy alone does not necessarily imply that transgene integrations are at the desired locus (unless the loss-

of-function due to insertion at the locus can be leveraged). Therefore, the selectivity of a particular cell line development strategy comprised of both the enrichment for transfected cells and the isolation of the cell population with transgene integration at the desired locus need to be addressed simultaneously.

To account for the selectivity in a cell line development strategy, we investigated options to facilitate our enrichment by coupling positive and negative selection and alternative nucleases with increased on-target activity. We suspected that the two-pronged improvements will significantly favor the transfected cell population with transgene integration at the desired locus. Following conventional practice, screening the transfected cell population with positive selection allows enrichment of the transgene containing population, but coupling this selection with negative selection could remove transgene integrations at undesired loci. By increasing the propensity of the CRISPR/Cas9 system to induce DSBs at the desired locus, more DNA repair events could occur at that site, thereby increasing the likelihood of integrations at that the target site.

3.3. RESULTS AND DISCUSSION

3.3.1. Characterizing basal HDR rate and transgene designs

For our cell line development strategy, combining positive and negative selection relies heavily on HDR and perhaps MMEJ to repair the DSB at the target site. Any resulting NHEJ at the target site would be eliminated due to the negative selection marker present in our transgene. Therefore, obtaining an estimate of the basal HDR rate without selection could serve as a baseline or a minimal expected rate of integration for our system into the desired locus. We evaluated this basal HDR rate in HT1080 model cells using a dual reporter (SEAP and hrGFP) linked by an EMCV IRES[205] transgene with 800-bp of flanking homology to the *GRIK1* locus identified in Chapter 2 (**Figure 3-1**).

Figure 3-1: Transgene design for evaluating basal HDR rate in HT1080 cells targeting the *GRIK1* locus.

Coding sequences for secreted alkaline phosphatase (SEAP) and humanized *Renilla* green fluorescent protein (hrGFP) are linked by the EMCV IRES to enable bicistronic expression.

We generated 3 heterogeneous bulk populations that expressed our dual reporter transgene without using selection pressure, and the resulting bulk populations contained a low proportion of cells expressing our transgene based on flow cytometry (**Figure 3-2, left**). From these populations, we seeded 576 wells for limited dilution cloning and isolated 5 populations that stably expressed hrGFP. However, from these 5 populations, only 3 were clonal populations as suggested by the histograms from flow cytometry (**Figure 3-2, right**); it is obvious that 2 of the populations are comprised of two distinct populations: untransfected and transgene expressing as represented by the bimodal distribution. Based on the frequency of isolating these clonal populations with confirmed integration at the *GRIK1* target site, the overall, basal HDR rate is ~0.5% in HT1080. This HDR rate matches more closely with the integration rates previously reported in HEK293 and CHO cells for a ~5-kbp transgene[180], and our transgene is ~5.6-kbp including the homology regions.

Figure 3-2: Histograms from flow cytometry analysis evaluating basal homologous recombination in HT1080.

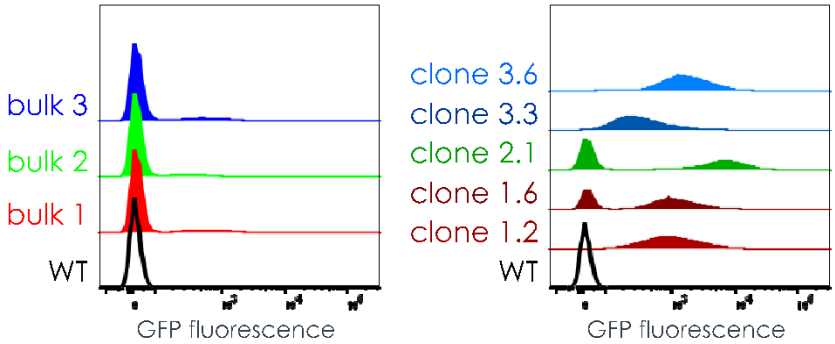
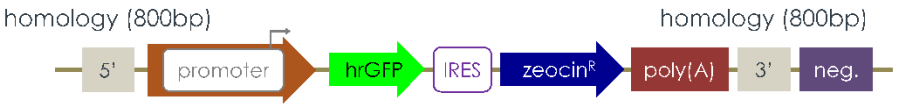


Figure 3-2, continued:
(left) Histograms from flow cytometry analysis of 3 HT1080 heterogeneous bulk populations containing our dual reporter transgene. (right) Histograms of 5 populations isolated by limited dilution cloning derived from the HT1080 bulk populations. The clone numbering "x.y" denotes that the "y" clone isolated from bulk population "x".

These low rates in HT1080, HEK293, and CHO suggest that an optimized strategy could drastically improve the frequency of successful transgene integration. In order to assess our combined positive and negative selection cell line development strategy, we created transgenes expressing the hrGFP reporter and Zeocin™ resistance[204] linked by the same IRES element that are flanked by 800-bp homology to *GRIK1*/*C. griseus Grik1* (hereafter *Cg.Grik1*) on each side. An additional expression cassette containing the negative selection marker based on the fusion of the *S. cerevisiae*-derived cytosine deaminase gene and the uracil phosphoribosyltransferase gene[206-208] (sequence derived from pSELECT-zeo-Fcy::fur, InvivoGen, San Diego, CA) located downstream of the 3' 800-bp locus homology region (**Figure 3-3**). Selection against the expression of this fusion gene with 5-fluorocytosine (5-fC) should eliminate cells with the transgene integrated anywhere besides the target locus.

Figure 3-3: Transgene design for evaluating positive and negative selection in HT1080, HEK293, and CHO cells targeting the *GRIK1*/*Cg.Grik1* locus.



Coding sequences for the humanized *Renilla* green fluorescent protein (hrGFP) is linked by the EMCV IRES to Zeocin™resistance, enabling bicistronic expression.

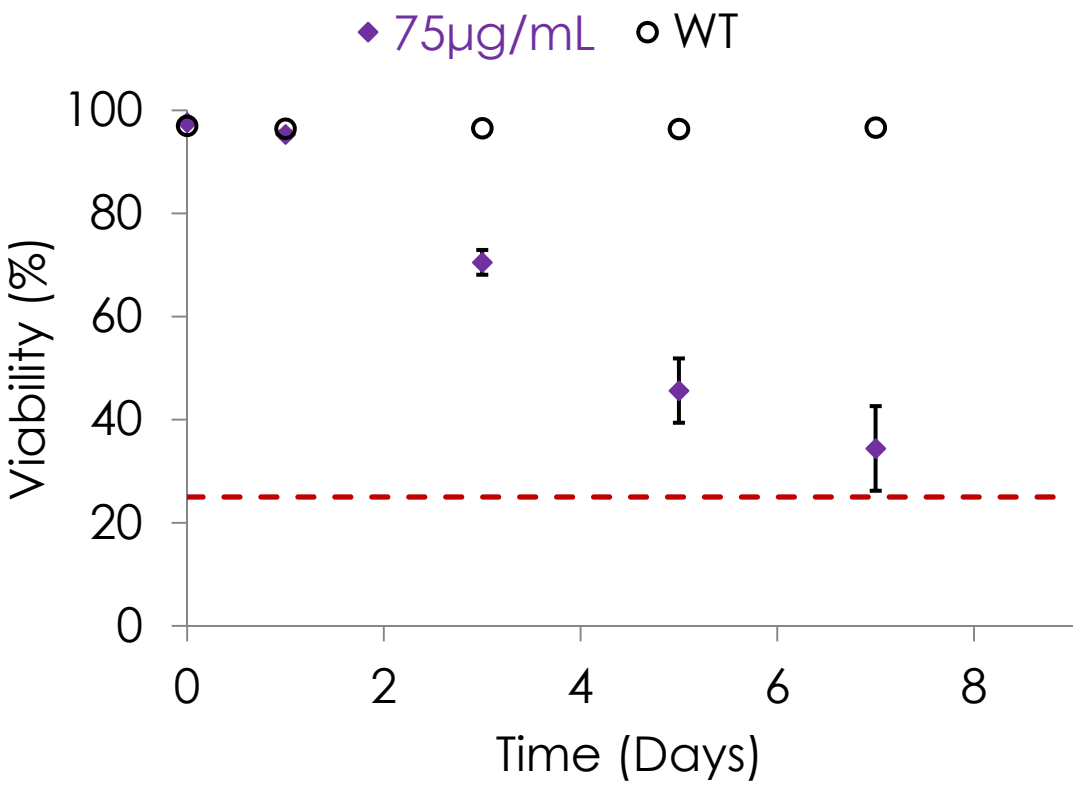
3.3.2. Determining effective concentrations of selection agents

The positive selection, negative selection, and co-targeting selection require Zeocin™, 5-fC, and 6-tG selection respectively. The effective concentration of these agents can vary across cell types, and we established working concentrations for our cell

line development strategy empirically by determining the maximum inhibitory concentration that kills 75% of the population (MIC_{75}). This determination for Zeocin™ and 6-tG was executed using wild-type (parental) host cells since Zeocin™ resistance or mutations to the *HPRT1*/*C. griseus Hprt1* (hereafter *Cg.Hprt1*) locus are required for cell survival against these selection agents.

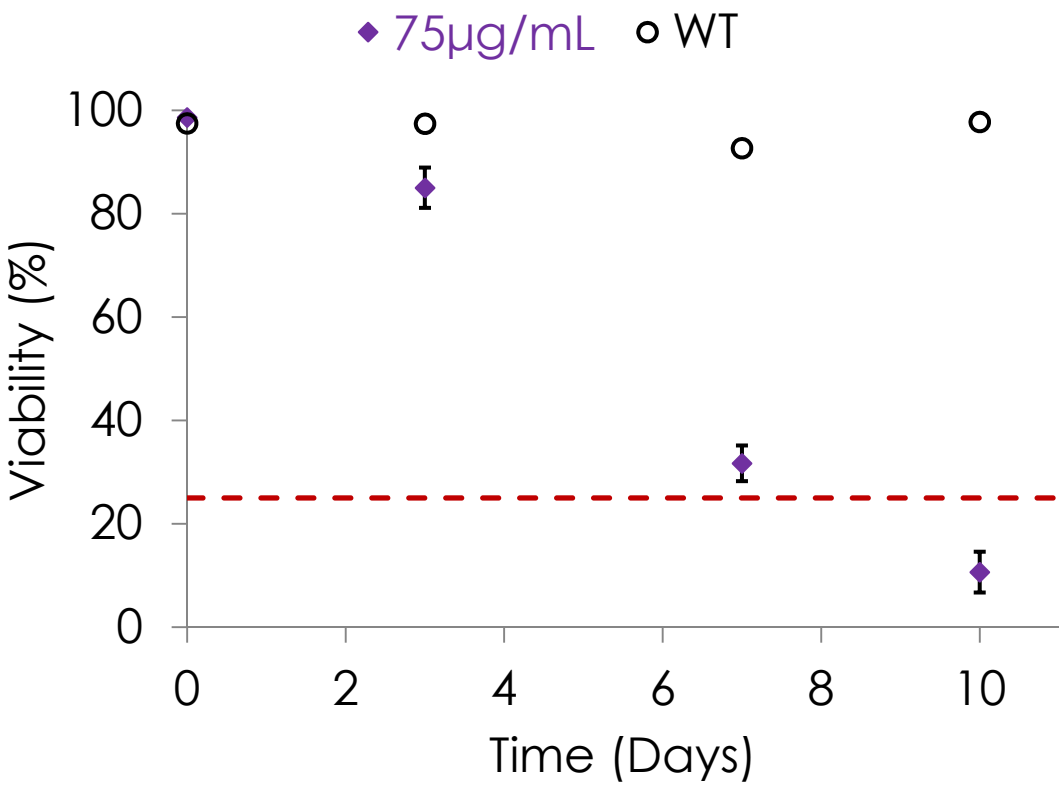
We confirmed that the previously reported approximate concentrations of Zeocin™ required are sufficient for inducing robust cell death in HT1080 and HEK293 cells[25]. We tracked the cell viability of 3 parental HT1080 (**Figure 3-4**) and HEK293 (**Figure 3-5**) cell lines subjected to 75 μg/mL Zeocin™ at day 0 and observed sufficient cell death in 7-10 days with media replacement every 2-3 days. Interestingly, a much higher selection concentration of Zeocin™ was required to achieve adequate killing of CHO cells within that same time period based on our data (**Figure 3-6**).

Figure 3-4: Confirmation of effective Zeocin™ selection in HT1080 cells



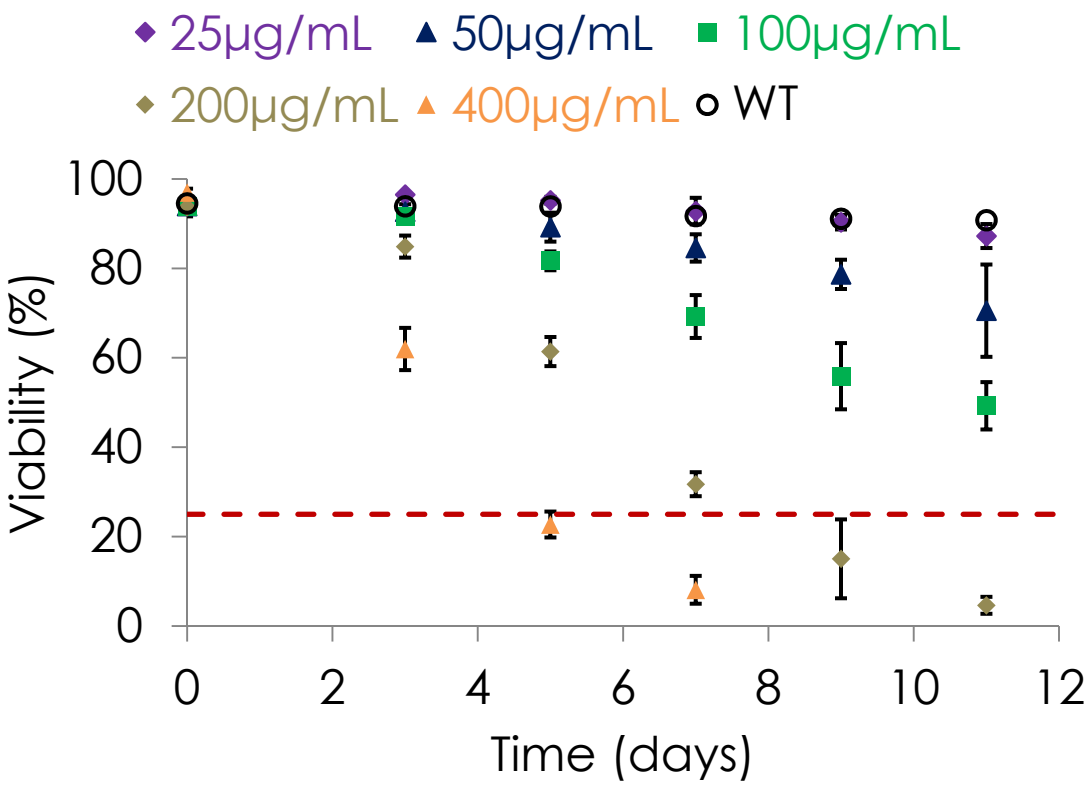
Culture media replaced every 2 days. Cell viability determined by ViCell XR (Beckman Coulter).

Figure 3-5: Confirmation of effective Zeocin™ selection in HEK293 cells



Media replaced every 3 days. Cell viability determined by ViCell XR (Beckman Coulter).

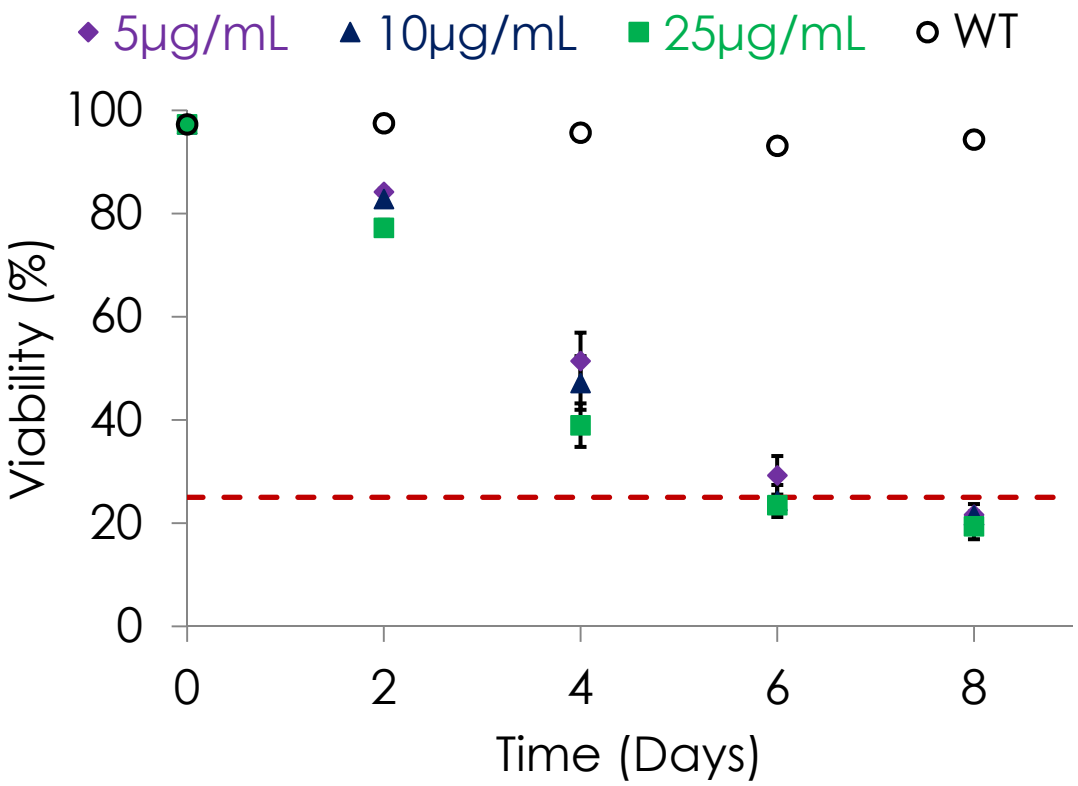
Figure 3-6: Confirmation of effective Zeocin™ selection in CHO cells



Media replaced every 2 days. Cell viability determined by ViCell XR (Beckman Coulter).

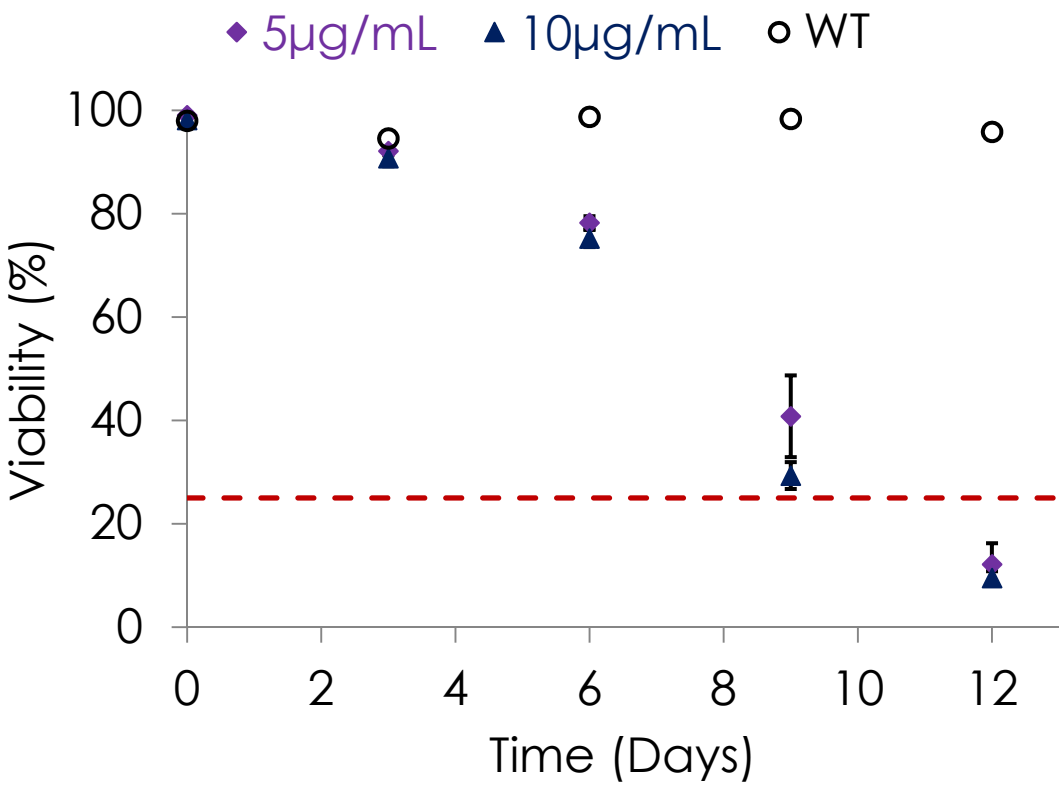
Based on previous work[181], we evaluated final concentrations of 6-tG at 5 µg/mL and 10 µg/mL in all 3 parental cell lines with the addition of 25 µg/mL for HT1080. All concentrations tested were effective in these cell lines, with 10 µg/mL only showing a marginal benefit in HEK293 cells (**Figures 3-7** to **3-9**). Therefore, we anticipated using the 5 µg/mL final concentration for 6-tG selection.

Figure 3-7: Confirmation of effective 6-tG selection in HT1080 cells



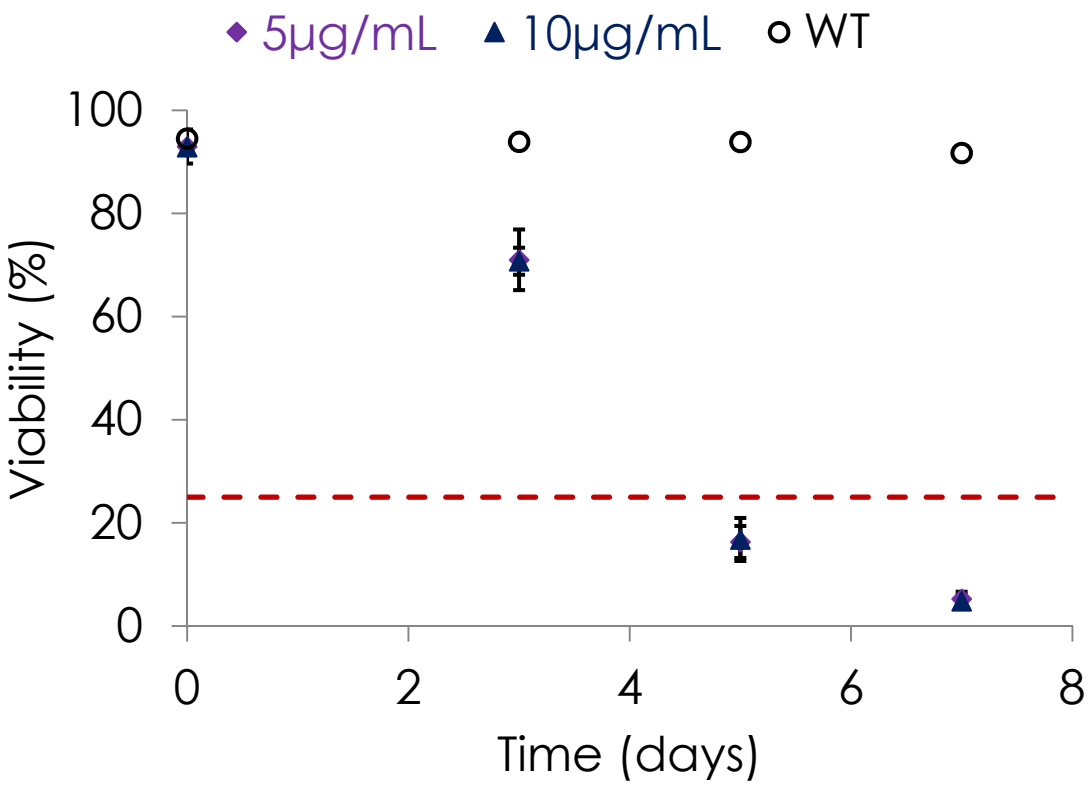
Media replaced every 2 days. Cell viability determined by ViCell XR (Beckman Coulter).

Figure 3-8: Confirmation of effective 6-tG selection in HEK293 cells



Media replaced every 3 days. Cell viability determined by ViCell XR (Beckman Coulter).

Figure 3-9: Confirmation of effective 6-tG selection in CHO cells



Media replaced every 2 days. Cell viability determined by ViCell XR (Beckman Coulter).

However, in order to evaluate various concentrations of 5-fC required for effective selection, we would first need to generate cell lines containing the negative selection fusion gene in HT1080, HEK293, and CHO cell lines. To obtain representative selection concentrations for 5-fC, we would need cell populations that contain a mixture of the entire transgene correctly integrated at the target site (removing the negative marker through HDR) and integrations due to NHEJ repair (retaining the negative marker), thus requiring the establishment of such cell lines prior to this analysis. Nonetheless, it is expected that 5-fC at 125-500 µg/mL is effective in mammalian cells expressing cytosine deaminase[206], although this concentration may be even lower in our system given the superiority of the

yeast-derived cytosine deaminase variant in our fusion gene[207] and the presence of both cytosine deaminase and uracil phosphoribosyltransferase can further increase 5-fC toxicity[209].

3.3.3. Qualifying gRNA designs in HT1080, HEK293F, and CHO-S cell lines

To effectively use the CRISPR/Cas9 system, the corresponding gRNAs designed for various loci must be able to properly interact with tbe Cas9 enzyme to induce the DSB. We can leverage the co-targeting approach previously established[181] in order to evaluate the extent at which a particular gRNA design would induce DSB activity at the desired locus (*GRIK1*/*Cg.Grik1*). Previously, without the co-targeted editing of *HPRT1*/*Cg.Hprt1* and selection with 6-tG, it would be difficult to ascertain whether DSBs did not occur at the target site due to poor gRNA design, defunct Cas9 activity, or poor Cas9 expression in the particular cell type, implying that integrations would not be directed towards the target site.

In order to estimate the editing frequency at the target locus using CRISPR/Cas9 and a particular gRNA, we can recover the sequence of target locus with PCR. Given the lower fidelity of *Taq* polymerase used to recover the sequence of the target locus, we first analyzed our recovery procedure using the WT/parental cell lines. This analysis of the WT sequence informs us of the false-positive rate of arbitrarily detecting editing either as a mutation or indel at the target region. We amplified the target regions of HT1080, HEK293, and CHO cells from their extracted genomic DNA. These PCR products were sub-cloned into pCR™4-TOPO® TA vectors (Thermo Fisher Scientific) for sequence recovery by Sanger sequencing. Based on this data (**Table 3-1**), the PCR/TOPO TA recovery method does not yield false estimates of indels, but introduces some mutations at the target regions (~3%).

Table 3-1: Estimated frequency of editing in *HPRT1*/*Cg.Hprt1*, *GRIK1*/*Cg.Grik1*, or AAVS1 locus from WT/parental HT1080, HEK293, and CHO cells.

cell line	target	total colonies	Indel count/freq		Mutation count/freq		indel+mutation count/freq		WT count/freq	
HT1080	HPRT1	11	0	0.0%	0	0.0%	0	0.0%	11	100.0%
HT1080	GRIK1	12	0	0.0%	1	8.3%	0	0.0%	11	91.7%
HT1080	AAVS1	11	0	0.0%	0	0.0%	0	0.0%	11	100.0%
HT1080	total	34	0	0.0%	1	2.9%	0	0.0%	33	97.1%
HEK293	HPRT1	11	0	0.0%	0	0.0%	0	0.0%	11	100.0%
HEK293	GRIK1	11	0	0.0%	0	0.0%	0	0.0%	11	100.0%
HEK293	AAVS1	11	0	0.0%	1	9.1%	0	0.0%	10	90.9%
HEK293	total	33	0	0.0%	1	3.0%	0	0.0%	32	97.0%
CHO	Grik1	11	0	0.0%	0	0.0%	0	0.0%	11	100.0%
CHO	Hprt1.1	10	0	0.0%	0	0.0%	0	0.0%	10	100.0%
CHO	total	21	0	0.0%	0	0.0%	0	0.0%	21	100.0%

Counts determined by Sanger sequencing of PCR amplified/TOPO TA recovery. The counts correspond to insertions/deletions (indel), mutations in the gRNA region, indel and mutations in the gRNA region, or the native wild-type (WT) sequence of the target region.

We then quantified the frequency of genome editing at the *GRIK1*/*Cg.Grik1* locus using our gRNA designs (GRIK1A, GRIK1B, GRIK1C, Cg.Grik1.1, Cg.Grik1.2) in HT1080, HEK293, and CHO cells, and as a control, at the AAVS1 locus using gRNAs previously described[144] (AAVS1.T1 and AAVS1.T2) in HT1080 and HEK293 cells. Stable cell lines of each cell type were generated using the SpCas9[144] in the Cas9 expression vector used by Slaymaker, *et al*[93] co-transfected with our target gRNA designs and *HPRT1*/*Cg.Hprt1.1* gRNA in plasmid vectors (Addgene 41824). We extracted gDNA

from each of these transformed populations after their recovery from selection with 5 μg/mL 6-tG. To estimate of the editing frequency at these loci in these Cas9-edited cell populations, we recovered the genomic sequence of the target region by PCR and confirmed the sequence by Sanger sequencing. We found that our GRIK1B gRNA was particularly effective in HT1080 although we detected lower activity using the same gRNA in HEK293 (**Table 3-2**). Likewise, we detected activity for both control gRNAs to AAVS1 in HT1080, and corroborated previous findings that the T2 target is more effective than the T1 target[144], yet neither targets were edited in HEK293 (**Table 3-2**). Most importantly, by using this *HPRT1*/*Cg.Hprt1* co-targeting approach, we confirmed that the Cg.Grik1.1 gRNA was effective while the Cg.Grik1.2 gRNA was unfortunately a poor design (**Table 3-2**). Through this approach, we can attribute the lack of target locus editing (represented as minimal indels and indels with mutations in the target locus) is directly related to the gRNA design instead of nuclease activity since a functional Cas9 is required to edit *HPRT1*/*Cg.Hprt1* to confer survival with 6-tG selection.

Table 3-2: Estimated frequency of editing in *GRIK1*/*Cg.Grik1* or AAVS1 from HT1080, HEK293, and CHO cells after 6-tG selection.

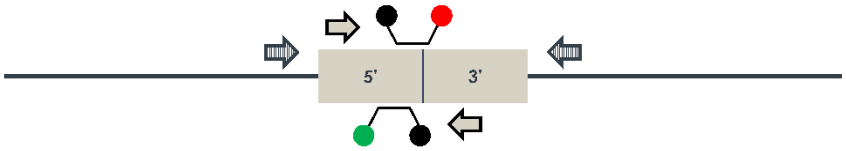
cell line	target	total colonies	Indel count/freq		Mutation count/freq		indel+mutation count/freq		WT count/freq	
HT1080	GRIK1A	21	8	38.1%	0	0.0%	0	0.0%	13	61.9%
HT1080	GRIK1B	20	15	75.0%	0	0.0%	0	0.0%	5	25.0%
HT1080	GRIK1C	8	4	50.0%	0	0.0%	2	25.0%	2	25.0%
HT1080	AAVS1.T2	22	21	95.5%	0	0.0%	0	0.0%	1	4.5%
HT1080	AAVS1.T1	22	14	63.6%	0	0.0%	0	0.0%	8	36.4%
HEK293	GRIK1B	24	8	33.3%	2	8.3%	0	0.0%	14	58.3%
HEK293	GRIK1C	23	3	13.0%	1	4.3%	0	0.0%	19	82.6%
HEK293	AAVS1.T2	23	0	0.0%	0	0.0%	0	0.0%	23	100.0%
HEK293	AAVS1.T1	20	0	0.0%	1	5.0%	0	0.0%	19	95.0%
CHO	Cg.Grik1.1	22	17	77.3%	1	4.5%	0	0.0%	4	18.2%
CHO	Cg.Grik1.2	24	2	8.3%	2	8.3%	1	4.2%	19	79.2%

Counts determined by Sanger sequencing of PCR amplified/TOPO TA recovery. The counts correspond to insertions/deletions (indel), mutations in the gRNA region, indel and mutations in the gRNA region, or the native wild-type (WT) sequence of the target region.

In addition to verifying the edited genomic sequence by PCR and Sanger sequencing, we can qualitatively confirm that the target locus was edited and estimate the editing frequency in the target locus with quantitative PCR[210]. Probes designed near and at the target site detect the abundance of the region (**Figure 3-10**), and the editing frequency can be estimated through a differential signal between the two probes. This difference is due to the loss of signal from the probe corresponding to the target site that is subject to editing and repair by NHEJ. Unlike the PCR/TOPO TA recovery method, this gene editing

frequency qPCR (GEF-qPCR[210]) assays the genomic DNA directly, avoiding any sampling bias selected for Sanger sequencing analysis.

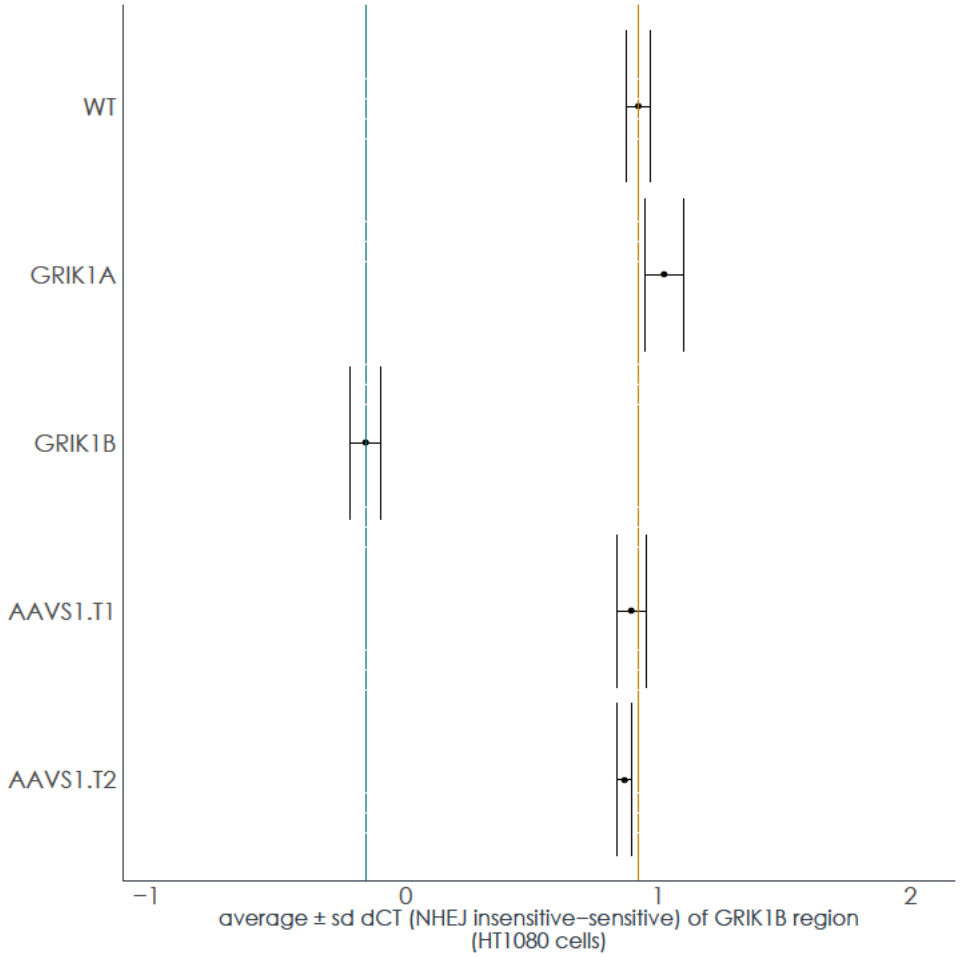
Figure 3-10: Depiction of qPCR probes near the target site (green) and at/spanning the target site (red).



Primers (solid arrows) amplify the target region for GEF-qPCR while the primers slightly outside of the target region (arrows with lined fill) recover the sequence for TOPO TA cloning and Sanger sequencing analysis.

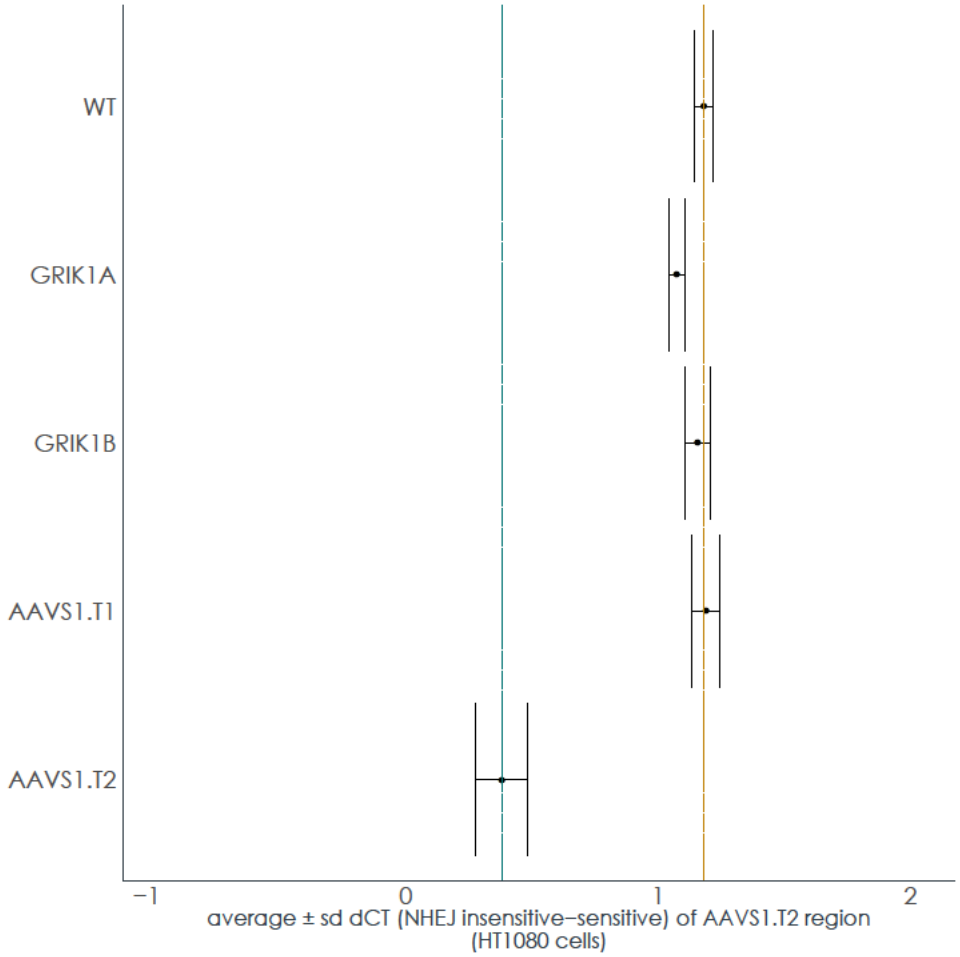
Using the same genomic DNA analyzed with the PCR/TOPO TA recovery method, we first estimated the editing of the *GRIK1* using the GRIK1B gRNA and AAVS1 locus using the AAVS1.T2 gRNA with GEF-qPCR as the threshold cycle (C_T, equivalent to quantification cycle C_q) difference relative to the WT sequence in HT1080. The measured change in threshold cycle (dC_T) was significantly different for the GRIK1B target site in *GRIK1* than the GRIK1A target site or the AAVS1 locus control, in addition to the parental WT sequence (**Figure 3-11**). Similarly, the same differential was observed at the AAVS1.T2 site using the AAVS1.T2 gRNA while the adjacent AAVS1.T1 site and *GRIK1* sites were not impacted (**Figure 3-12**). Similarly, the same measurements for the Cg.Grik1.1 gRNA target in *Cg.Grik1* indicated that editing was clearly detected at the target site in CHO cells (**Figure 3-13**).

Figure 3-11: Estimated editing of *GRIK1* at the GRIK1B target from HT1080 cell populations subjected to co-targeting and 6-tG selection based on dC_T from GEF-qPCR.



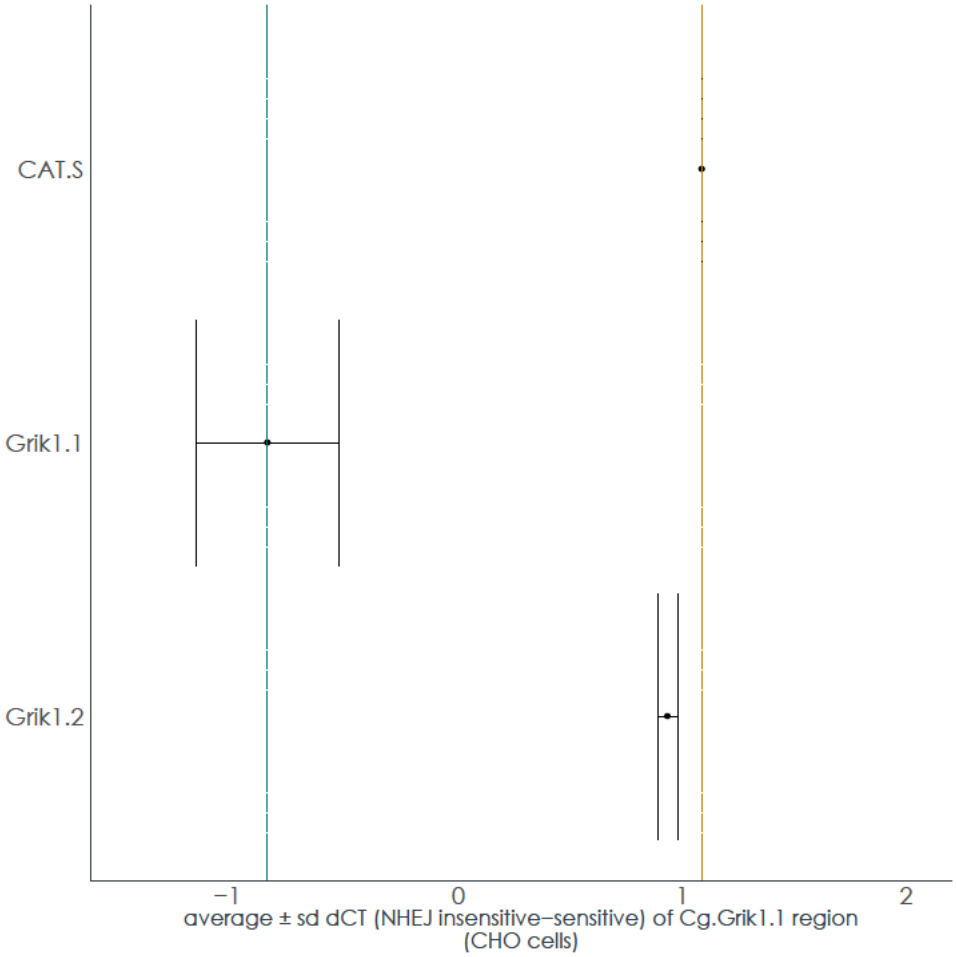
In comparison to the parental WT cells and other genomic DNA with edits at the AAVS1 locus, the genome was edited in *GRIK1* at the GRIK1B target site with the GRIK1B gRNA while no changes were detected using the GRIK1A gRNA.

Figure 3-12: Estimated editing of AAVS1 at the AAVS1.T2 target from HT1080 cell populations subjected to co-targeting and 6-tG selection based on dC_T from GEF-qPCR.



In comparison to the parental WT cells and other genomic DNA with edits at the *GRIK1* locus, the genome was edited in AAVS1 at the T2 target site with the AAVS1.T2 gRNA while no changes were detected using the AAVS1.T1 gRNA.

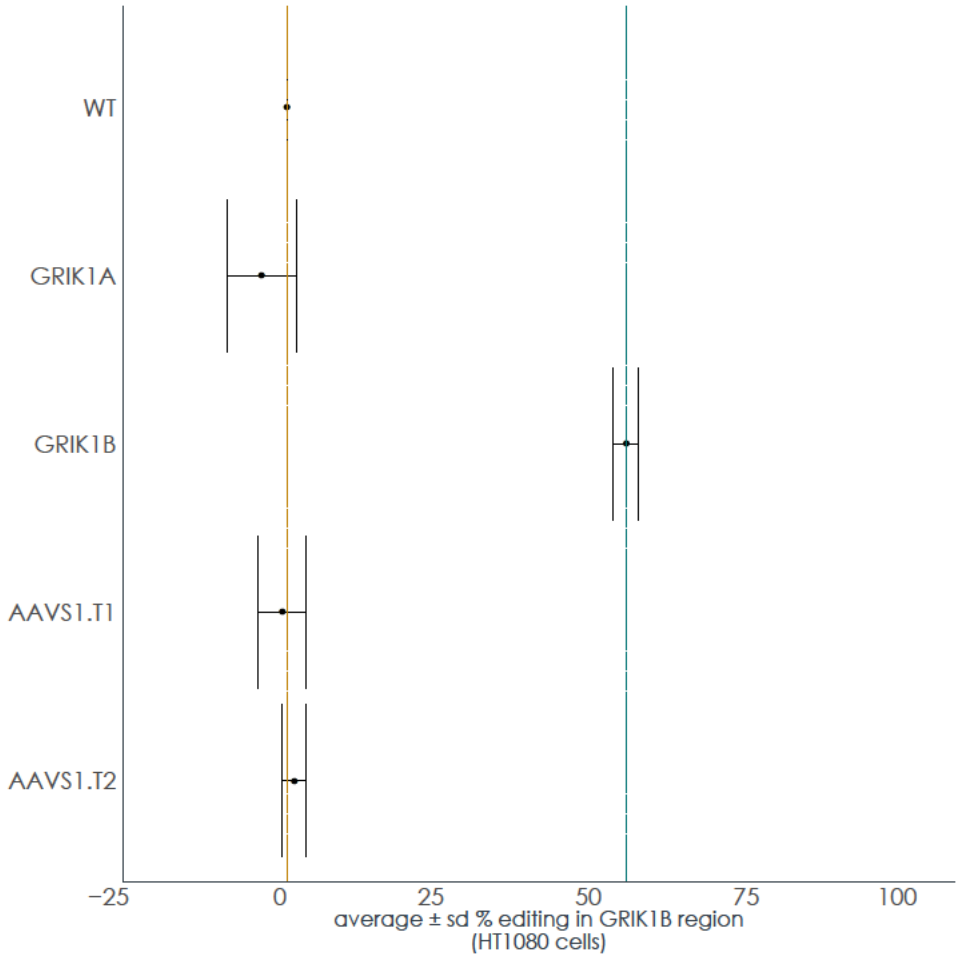
Figure 3-13: Estimated editing of *Cg.Grik1* at the Cg.Grik1.1 and Cg.Grik1.2 targets from CHO cell populations subjected to co-targeting and 6-tG selection based on dC_T from GEF-qPCR.



In comparison to the parental WT cells (CAT-S), the genome was edited at *Cg.Grik1* with the Cg.Grik1.1 gRNA while marginal changes were detected using the Cg.Grik1.2 gRNA.

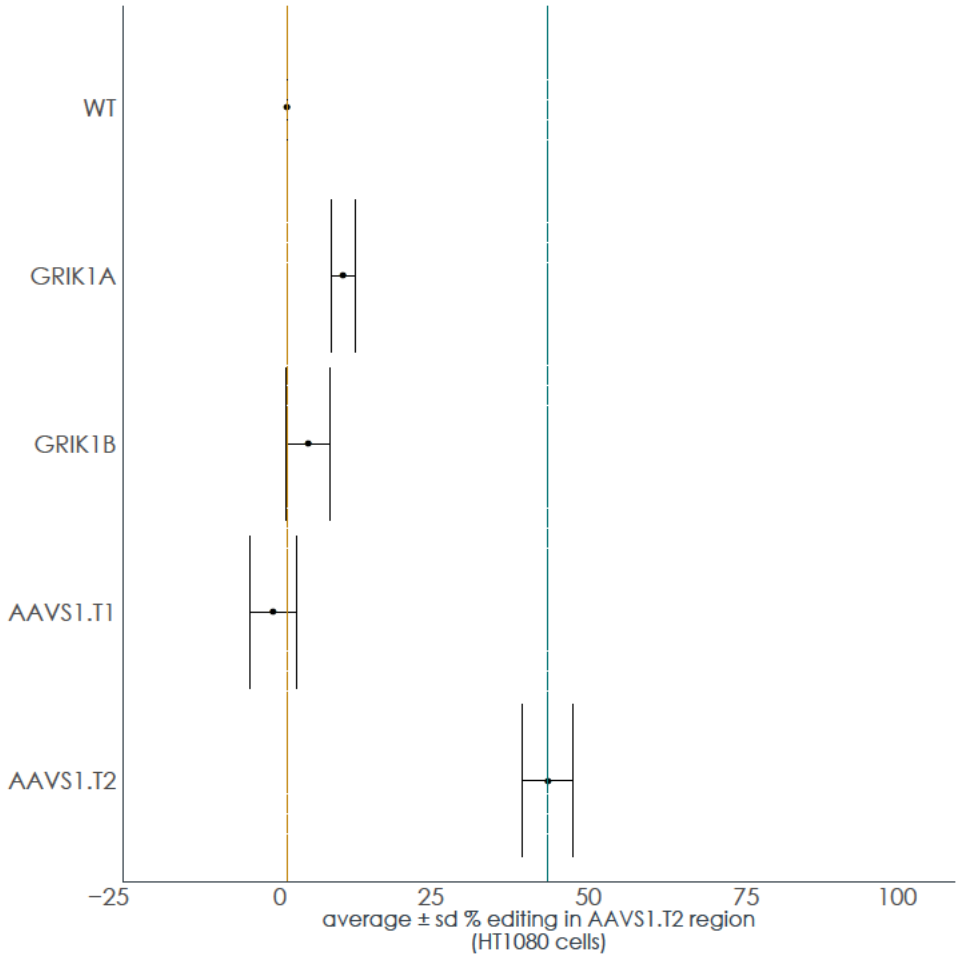
By establishing standard calibrations for the primers used in GEF-qPCR, we can also estimate the editing frequency from this data (**Figures 3-14** to **3-16**). Collectively, these results suggest that GEF-qPCR is sufficiently sensitive to single nucleotide resolution, enabling precise interrogation of the target integration site.

Figure 3-14: Estimated editing of *GRIK1* at the GRIK1B target from HT1080 cell populations subjected to co-targeting and 6-tG selection.



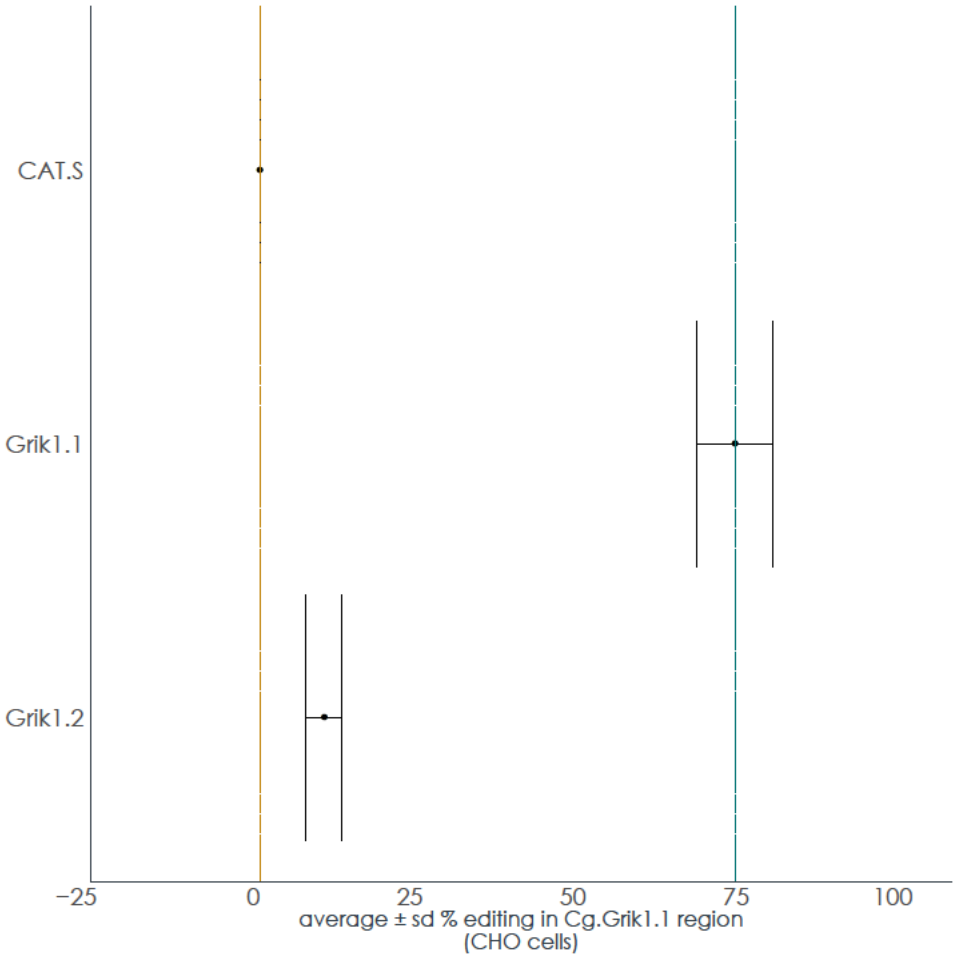
In comparison to the parental WT cells and other genomic DNA with edits at the AAVS1 locus, the genome was ~53% edited in *GRIK1* at the GRIK1B target site with the GRIK1B gRNA while no changes were detected using the GRIK1A gRNA.

Figure 3-15: Estimated editing of AAVS1 at the AAVS1.T2 target from HT1080 cell populations subjected to co-targeting and 6-tG selection.



In comparison to the parental WT cells and other genomic DNA with edits at the *GRIK1* locus, the genome was ~42% edited in AAVS1 at the T2 target site with the AAVS1.T2 gRNA while no changes were detected using the AAVS1.T1 gRNA.

Figure 3-16: Estimated editing of *Cg.Grik1* at the Cg.Grik1.1 and Cg.Grik1.2 targets from CHO cell populations subjected to co-targeting and 6-tG selection.



In comparison to the parental WT cells (CAT-S), the genome was ~75% edited at *Cg.Grik1* with the Cg.Grik1.1 gRNA while marginal changes (~10%) were detected using the Cg.Grik1.2 gRNA.

3.4. CONCLUSIONS

The work described in this chapter established the groundwork for a method to quantify the integration frequency at a target site mediated by CRIPSR/Cas9, estimating the selectivity of the transgene integration process. This work facilitates the evaluation of Cas9 variants with higher fidelity on transgene integration and an assessment of a coupled positive and negative selection strategy to improve targeted integration selectivity. It is

through an effective and precise generation of DSBs at the target site and subsequent repair with the transgene donor to that site that an integration process can be considered selective for cell line development. While we observed the expected behavior when evaluating Zeocin™ selection conditions in HT1080, HEK293, and CHO cells, we observed a significant detriment to cell viability with 6-tG selection such that resistant populations were generated with only a single treatment instead of media replacement with 6-tG. Even with the single treatment, these selected populations were still resistant to 6-tG at 5 μg/mL after recovery.

Based on this toxicity, coupling both Zeocin™ and 6-tG selection in a cell line development strategy would be extremely taxing on the cells and we concluded that the co-targeting approach should be reserved for evaluating gRNA designs. Furthermore, if cells are co-targeted for DSBs at *HPRT1*/*Cg.Hprt1* and another locus, there would be an increased chance of undesired chromosomal rearrangements and/or transgene integration into the *HPRT1*/*Cg.Hprt1* locus instead of the desired locus. Even though a selective cell line development strategy incorporating negative selection should eliminate cells with integrations at other loci, including *HPRT1*/*Cg.Hprt1*, the additional DSB and repair at *HPRT1*/*Cg.Hprt1* impose additional unnecessary strain on the cells since the target gRNA would already be validated for activity.

Through the *HPRT1*/*Cg.Hprt1* co-targeting approach[181], we identified gRNA designs (GRIK1B and Cg.Grik1.1) that are effective in HT1080, HEK293, and CHO cells. We can also use this method to conclude that other gRNA targets (GRIK1A, GRIK1C, and Cg.Grik1.2) are poor designs for these particular cell types based on the induced DSB activity and resulting NHEJ repair. Although we determined an effective selection concentration for 6-tG in HT1080, HEK293, and CHO cells, we observed much lower DSB activity for both *GRIK1* and AAVS1 loci in HEK293 cells based on measurements by our

PCR/TOPO TA recovery method (**Table 3-1**). In particular, the same GRIK1B gRNA showed lower activity in HEK293 than HT1080, suggesting that gRNA design may need to account for various cell types despite sharing the same genomic sequence. However, it is also possible that the observed differences are manifestations of different transfection efficiencies between the two cell types. Crucially, we did not measure any DSB activity at the AAVS1 locus in 6-tG resistant HEK293 cells, which conflicts with previous work describing the DSB activity at this locus[144]. The AAVS1.T1 and AAVS1.T2 gRNA targets are commonly used as positive controls, thus further investigation with our HEK293 cells are necessary to clarify the lack of DSB activity and DNA repair at these sites.

Chapter 4: Creating the message – engineering synthetic promoters for high transgene expression[3]

4.1. CHAPTER SUMMARY

To establish precise control of gene expression in mammalian hosts, we need to investigate the local genetic elements that govern transcription in addition to specifying the integration loci of these elements. Traditionally, promoter engineering enabled the transcriptional control required for bioproduction from mammalian hosts, focusing on achieving high expression levels. Despite recent advances in improving titers for therapeutic proteins such as antibodies to the 10 g/L scale, these high yields can only be achieved in select mammalian hosts. Regardless of the host or product, strong promoters are required to obtain these high levels of transgene expression. However, the promoters employed to drive this expression are rather limited in variety and are usually either viral-derived or screened empirically during vector design. To begin to move away from viral parts, we employed a more systematic approach to identify and design new synthetic promoters using endogenous elements. To do so, we established a workflow to design these elements by: (1) analyzing the transcriptomics profile of a specific cell line under a desired, representative cell culture condition, (2) identifying key genetic motifs using bioinformatics that can be used to rationally construct synthetic promoters, (3) building synthetic promoters using conventional DNA synthesis and molecular biology techniques, and (4) evaluating the performance of these synthetic promoters using model proteins. The resulting promoters perform comparably to the hCMV IE promoter variants tested, but with endogenous components. During this design-build-test cycle, we also investigated the

[3] The content in this chapter can be found in a previously authored publication. JKC conducted the experiments and analyses, and wrote the chapter.

underlying design rules for transcription factor binding site arrangement in synthetic promoters. Overall, this approach of using an 'omics-guided workflow for designing synthetic promoters facilitates the construction of high expression vectors for immediate use in current production hosts.

4.2. INTRODUCTION

In recent years, select mammalian hosts have become potent vehicles for the production of heterologous, therapeutic antibodies and proteins with industrial titers reaching and exceeding 10 g/L scale[211]. However, this capacity is not ubiquitous and is only possible in select hosts and for a subset of products, whereas difficult-to-express products still require intensive resources for their development[212, 213]. Optimal heterologous protein expression from mammalian hosts requires fine-tuning many parameters[213, 214] among which, the expression vector design plays a pivotal roles[42].

Inherently, one particular limitation in this field is the lack of genetic regulatory elements (namely, promoters) that can enable such high expression levels[42]. This underdeveloped toolkit poses a clear challenge for engineering complex biotechnology applications that involve multiple reactions or pathways (*e.g.* a complete heparin pathway[6]). Previously approaches for creating synthetic and hybrid promoters[107, 215, 216] employed both bottom-up and top-down approaches[217, 218]. Additionally, collections of regulatory promoters exist to enable complex functionality such as multi-cistronic control[219], epigenetic toggling[220], and a mammalian oscillator circuit[221]. Yet, the vast majority of promoter development has revolved around identifying, characterizing, and constructing hybrid promoters that are frequently viral-derived[5, 222-224], thus resulting in synthetic parts that are susceptible to silencing[225] with unreliable utility for long-term industrial processes. To bypass stability issues, the commonly used cytomegalovirus

immediate-early [CMV IE] promoter was modified to include a CG-rich region (CpG island)[226, 227] as well as alternative promoter variants[228]. However, these modifications do not remove the viral nature of these promoters since methylation was observed at both CpG and non-CpG sites[229] nor do they greatly expand the set of tools available. As a result, a set of synthetic promoters (ideally non-virally derived) will be of high utility for mammalian cell engineering applications.

This work seeks to establish a more systematic approach toward the rational design of synthetic promoter guided by high-throughput analysis such as microarray expression and RNA-seq. The underlying premise for these designs is to incorporate distinguishing features (*e.g.* putative transcription factor binding sites, [TFBSs]) over-represented in high expression promoters and ideally absent in low or moderate expression. A significant number of efforts (especially enabled by the ENCODE project[230-232]) have led to the cataloguing of transcription factors [TFs] and cognate TFBSs and their actions *in vivo*. As a result, a variety of TF databases are available including the JASPAR database[233] (primarily used in the work described in this chapter), TRANSFAC[234], MotifMap[235], UniProbe[236], and HOCOMOCO[237] (among others) for *H. sapiens*. While these databases contain a wealth of information that describes the endogenous expression program for a particular cell type, the use of these maps to prescribe transgene expression and design remains largely unexplored.

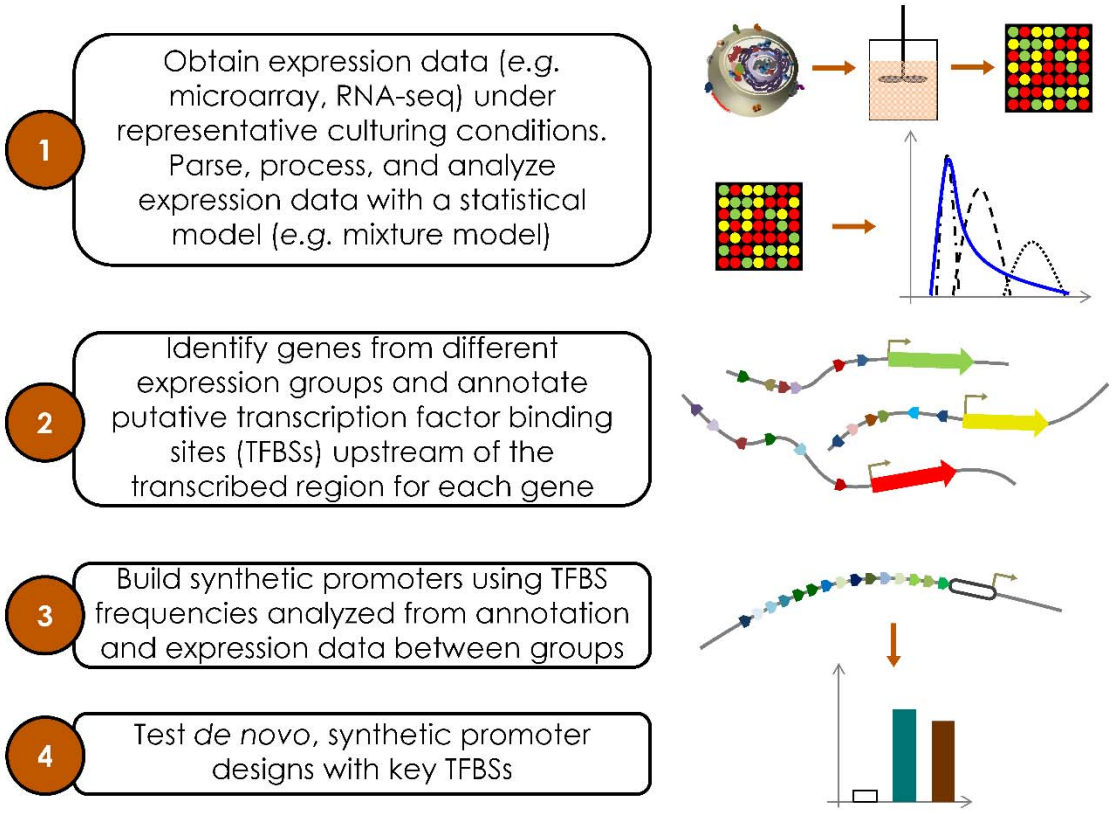
Here, we describe and demonstrate a generalizable workflow for designing synthetic promoters based on representative microarray expression data of a mammalian production host (in this case, for the HT1080 fibrosarcoma host cell line, which is used to produce three globally, commercially available therapeutics[135]). HT1080, in addition to other mammalian cell lines such as CHO and HEK293[238] are important industrial hosts commonly used to produce protein therapeutics, such as insulin, antibodies, cytokines,

enzyme replacement therapies, and growth factors[1]. Specifically, we use this expression data to identify TFBSs enriched in highly expressed promoters under desired conditions with the aid of a Gaussian Mixture Model [GMM] and bioinformatics. Next, we designed synthetic promoter scaffolds based on these TFBSs and evaluated their performance with two model proteins. The resulting promoters perform comparably to the hCMV IE promoter variants tested. Finally, we briefly investigated possible design rules for TFBS arrangement by creating multiple variants of our synthetic promoters and evaluating their ability to drive the expression.

4.3. RESULTS AND DISCUSSION

In this work, we established and executed a generalizable workflow (**Figure 4-1**) to analyze large expression data sets (such as from microarray) and generate designs for synthetic promoters for a given cell type. This workflow describes an expanded design-build-test cycle that is highly accessible: (1) Design: expression data derived from representative cell culture conditions best reflect synthetic promoter application; (2) Design: large expression data sets inform the design process for synthetic promoters; (3) Build: conventional DNA synthesis and standard molecular biology techniques are used to build the synthetic promoter designs for expression vectors; and (4) Test: these expression vectors are transfected into the cell line of interest for evaluation. Iterating upon this design-build-test cycle can further refine final designs and performance. For this project, we applied this expanded D-B-T workflow for creating synthetic promoters to the mammalian host HT1080.

Figure 4-1: Workflow to designing synthetic promoters from expression data.



A generalized workflow is established through this work to go from bioinformatics analysis to design and testing of synthetic promoters.

4.3.1. Processing and statistical modeling of gene expression data

A microarray expression data set (Illumina) of the HT1080 cell line collected at four distinct time-points throughout a bioreactor fermentation under representative industrial process conditions (spanning growth and production phase) was provided by Shire Human Genetic Therapies. Initially, this data comprising 48801 probes was pre-processed using a logarithm transformation to normalize expression values and mapped to genes in the human genome. Next, we modeled the expression data using a 3-component Gaussian Mixture Model (GMM) that assumed three populations of gene expression

existed in the data corresponding to high, moderate, and low expression. The parameters describing the Gaussian components of this model (μ_i, σ_i, and π_i) were determined using an expectation-maximization algorithm[239] with MATLAB software (Mathworks, R2014a). Qualitatively, this 3-component model adequately described the positively-skewed, log-transformed data at all time-points (**Figures 4-2A to 4-2D**), and this model provided a quantitative means to assess the probability of any given measured expression value belonging to each of the three expression profiles (high, moderate, and low).

Figure 4-2: Processing of Microarray data using a Gaussian Mixture Model.

A

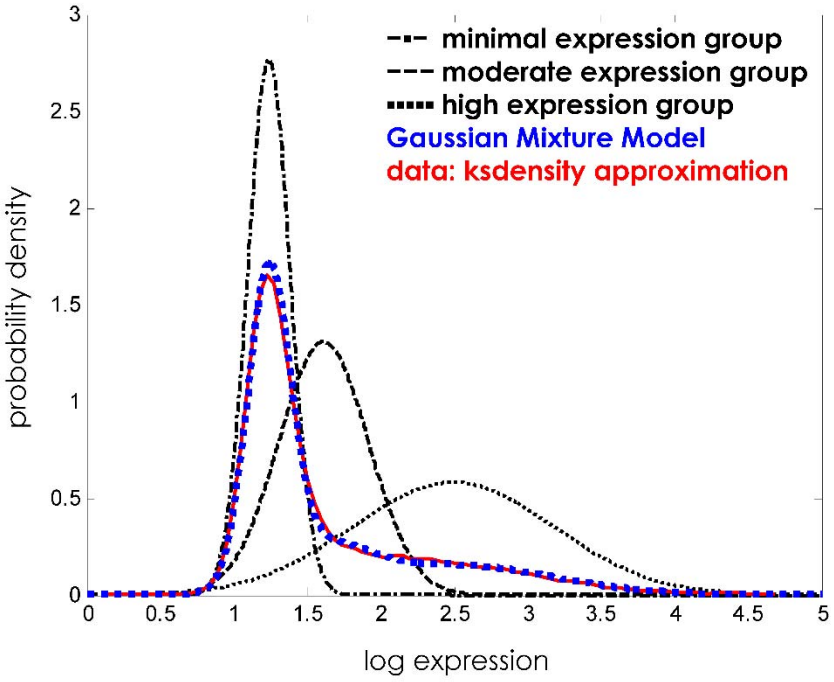
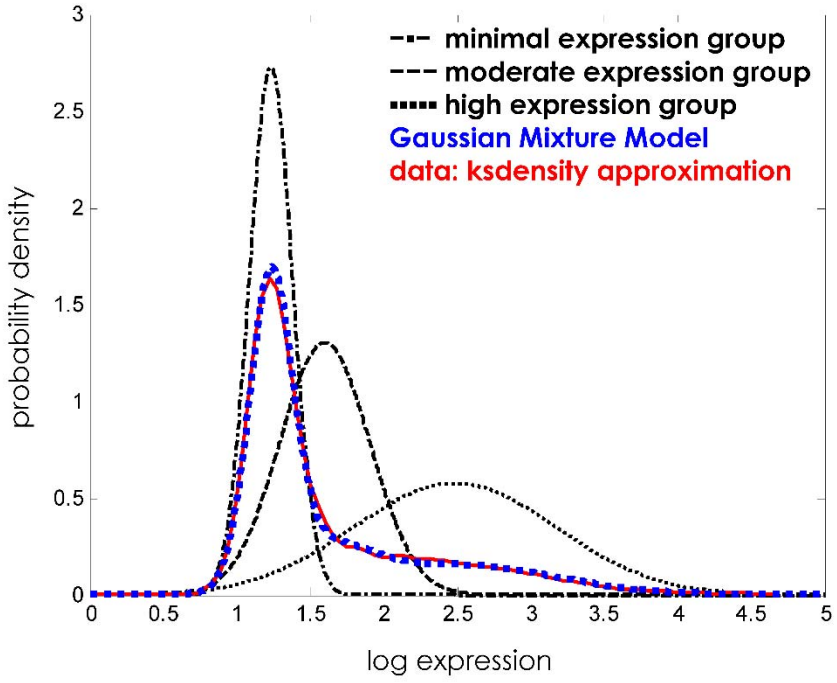


Figure 4-2, continued:
B



C

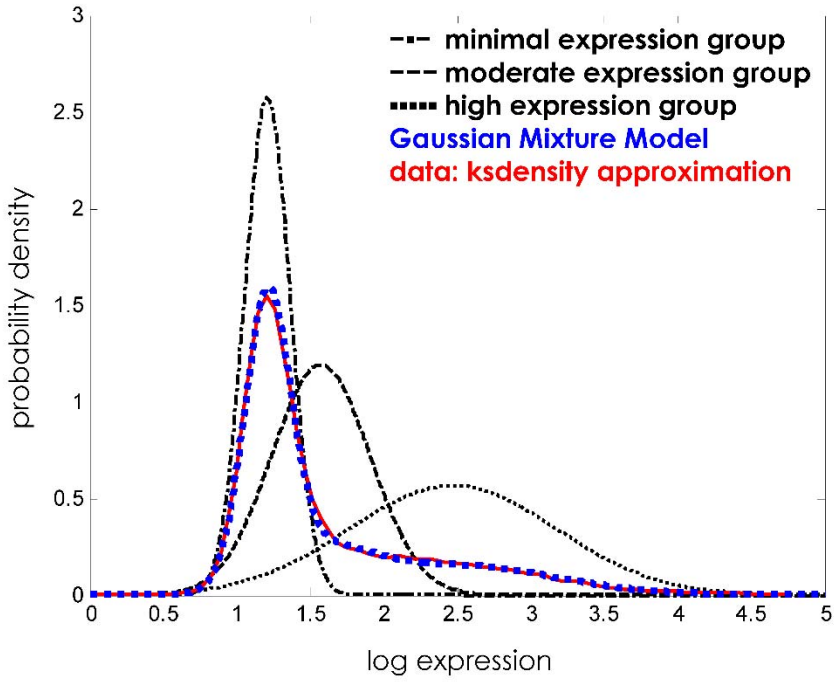
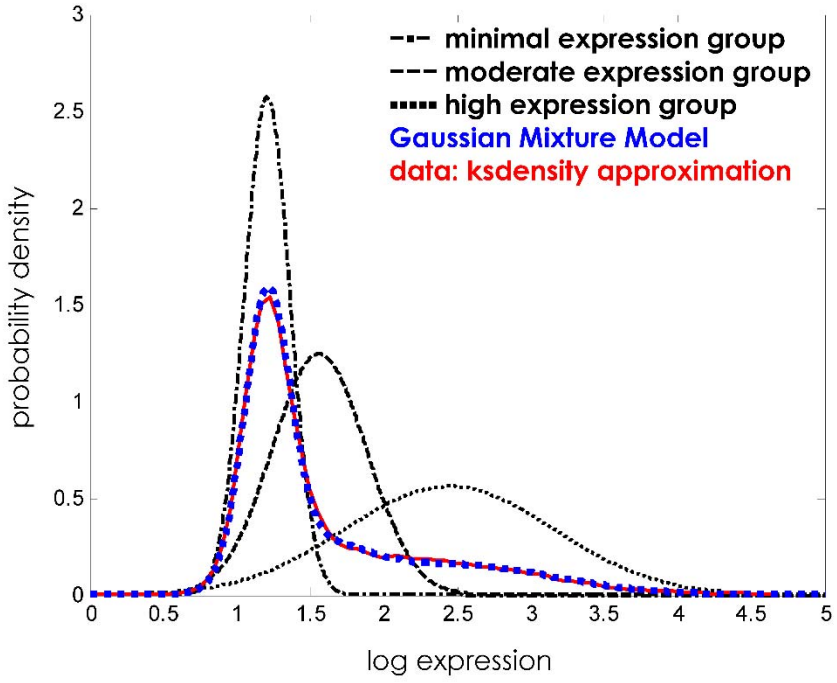


Figure 4-2, continued:
D



GMM (blue) with three components (black, various patterns) and ksdensity approximation (red) of log-transformed expression data at time-point **A)** t_1, **B)** t_2, **C)** t_3, and **D)** t_4.

Using this model, it was possible to evaluate expression profiles at each of the four timepoints and extract: (1) an expression threshold value (log-scale) that separated the "high" and "moderate" expression groups; (2) a false-positive probability associated with threshold (*i.e.* the probability that an expression value is incorrectly categorized as high expression); and (3) a false-negative probability associated with that threshold (*i.e.* the probability that an expression value is incorrectly categorized as low or moderate expression) (**Table 4-1**). Moreover, we chose a rather conservative threshold such that the probabilities $p(\textit{false-positive at threshold value}) = p(fp)|_{th} \approx 0.001$ was far less than $p(\textit{false-negative at threshold value}) = p(fn)|_{th} \approx 0.11\text{-}0.13$ across all time-points (**Table 4-1**). As a result, if anything, this threshold "cut-off" value would underestimate the number of

constituents in the high expression group. Similar sets of probabilities (**Table 4-1**) can be calculated across the other expression groups. These probabilities have great power over arbitrarily selecting a cut-off for "high" expression. For example, an arbitrary selection of the top 1% of available data (n = 488) as the "high expression group" would grossly underestimate membership for the group based the GMM with $p(fp)|_{th} \approx 0.001$ (n = 6,499). While the selection of discrete top m values would always be independent of using the statistical model, determining the members of the other expression groups (in this case, low and moderate), would not be as trivial especially for the skewed dataset observed (**Figure 4-2**). Ultimately, this model was used to map gene expression (and ultimately promoter sequences) with expression levels. While we chose the rational division of the microarray expression data into three groups representing discretized low, moderate, and high expression "groups", the number of groups (components) in the GMM is easily modifiable for alternative experimental data.

Table 4-1: Final Gaussian Mixture Model parameters.

time point	parameter	low group	moderate group	high group	moderate-high threshold	threshold *p(fp)|th*	threshold *p(fn)|th*	median *p(fn)|m*
t_1	mean, μ	1.2395	1.6116	2.4971	2.3704	1.00E-03	0.1141	0.0129
	std. deviation, σ	0.1443	0.3047	0.6856				
	probability, π	0.5759	0.1567	0.2674				
t_2	mean, μ	1.2325	1.5990	2.4879	2.3602	9.99E-04	0.1153	0.0135
	std. deviation, σ	0.1463	0.3056	0.6897				
	probability, π	0.5728	0.1569	0.2704				
t_3	mean, μ	1.2087	1.5721	2.4746	2.4095	1.00E-03	0.1257	0.0142
	std. deviation, σ	0.1551	0.3346	0.7023				
	probability, π	0.5662	0.1624	0.2714				
t_4	mean, μ	1.2070	1.5611	2.4496	2.3563	1.00E-03	0.1257	0.0162
	std. deviation, σ	0.1550	0.3198	0.7093				
	probability, π	0.5643	0.1550	0.2808				

Parameters from each time point of a representative HT1080 culture fermented in a bioreactor are presented. Final parameters are averaged from 10 independent optimizations of model parameters using an expectation-maximization algorithm. Values for the low, moderate, and high expression groups and the moderate-high threshold are in terms of the log(expression) from the microarray data set.

4.3.2. Elucidating key TFBSs for representative strong promoters

Using the GMM described above, we identified gene candidates based on their expression group and annotated their respective "promoter" regions with putative TFBSs using database-enabled bioinformatics. Specifically, we selected non-ribosomal, coding sequences and mapped their promoter sequences to the annotated human genome (GRCh38.p2 assembly, NCBI). For this test, we identified a subset of 20 of the most highly expressed genes across all time-points (**Supporting Information Table S1a** of the publication[240]), 20 genes with median level expression across all time-points (**Supporting Information Table S1b** of the publication[240]), and 20 randomly selected genes (**Supporting Information Table S1c** of the publication[240]) for further analysis. Mammalian promoters can vary greatly in size, ranging from 100-bp (base pair) scale to

over 1000-bp[222, 241, 242]. Therefore, we performed our analysis based on two sizes: a 2000-bp region and a 500-bp region preceding (in the 5' direction of) the transcription start site (based on GRCh38.p2 assembly). The goal of this analysis was to identify TFBSs enriched in the strong promoters found in the high expression group. All DNA sequences were annotated for these putative TFBSs based on consensus sequences from the JASPAR database[233] (vertebrates, *H. sapiens* only, **Supporting Information Table S2** of the publication[240]) using the "ApE – A plasmid Editor" software, v2.0.47 for visualization and tabulation. To account for native occurrences of putative TFBSs within a given 2000-bp region, we also analyzed up to 53 randomly selected 2000-bp regions from each human chromosome, which may include both intragenic and intergenic sequences. Finally, a similar annotation of several commonly used viral-derived and endogenous promoters[224, 243] was performed for comparison purposes (**Supporting Information Table S1d** of the publication[240]).

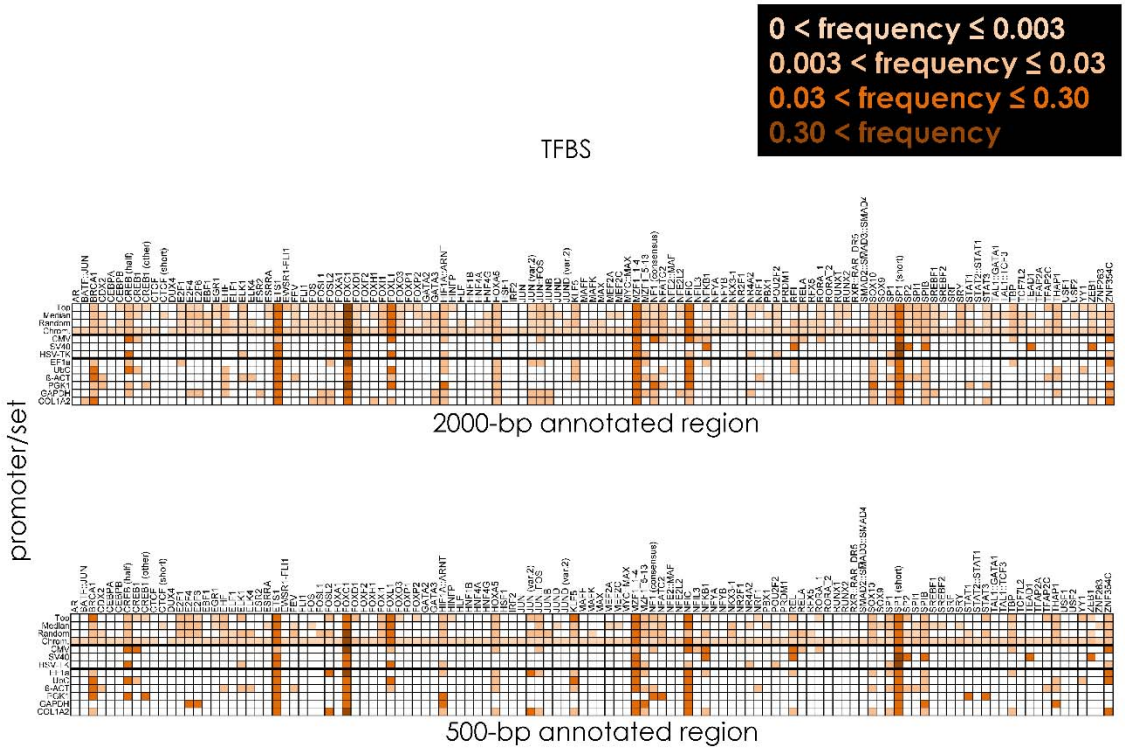
To identify enrichment of TFBSs amongst promoters belonging to specific expression groups, we chose to determine the frequencies of TFBSs as a ratio of the number of putative sites for each TFBS to the overall total number of TFBSs in each set. Immediate comparisons of these frequencies between the 2000-bp annotations and 500-bp annotations suggest minor differences due to promoter length (**Figure 4-3A**). More importantly, comparing these frequencies across promoters/promoter sets highlights differential enrichment for key TFBSs relevant to strong, measurable gene expression represented by the high expression group. For example, the data suggest an overrepresentation of activating TFs such as the Sp/KLF families[244] (Sp1, Sp2, and KLF5) and others also found in the hCMV IE gene enhancer[222]. These distinguishing features form the design basis of synthetic promoters. In addition to native promoters, comparisons were made with two conventional, viral-derived heterologous promoters used in research and large-scale

industrial applications (hCMV IE gene promoter and Simian Virus 40 [SV40] early promoter)[243] and two commonly used endogenous promoters (from *EEF1A1* and *UbC*)[224, 242, 243]. The putative TFBSs enriched in the high expression group encompass most of those annotated in the two viral-derived promoters (**Figures 4-3B** and **4-3C** and **Supporting Information Table S3a** and **S3b** of the publication[240]) and, as expected, all of the putative TFBSs annotated in the two endogenous promoters (**Figures 4-3D** and **4-3E** and **Supporting Information Table S3c** and **S3d** of the publication[240]).

Collectively, we utilized these comparisons to define a list of putative TFBSs for the first design cycle to construct (build) synthetic promoters for subsequent testing. We compared ~20% of the top frequencies (based on descending frequency as enriched in the high expression group) across the high, median, random, and chromosomal sets of genomic sequences. TFBS frequency is defined as the ratio of the number of times a given TFBS is found within the specific sequence window (2000-bp or 500-bp) to all TFBSs found in the same window for a given annotated set (top, median, random, chromosomal region). The resulting putative TFBSs (**Table 4-2**) considered to be enriched in high expression promoters (green, higher frequencies found in the high/top expression set compared to other sets) or enriched in background (red, higher frequencies found in the chromosomal set compared to other sets) were selected to create synthetic promoters.

Figure 4-3: Distribution of TFBS frequencies across promoter regions.

A



B

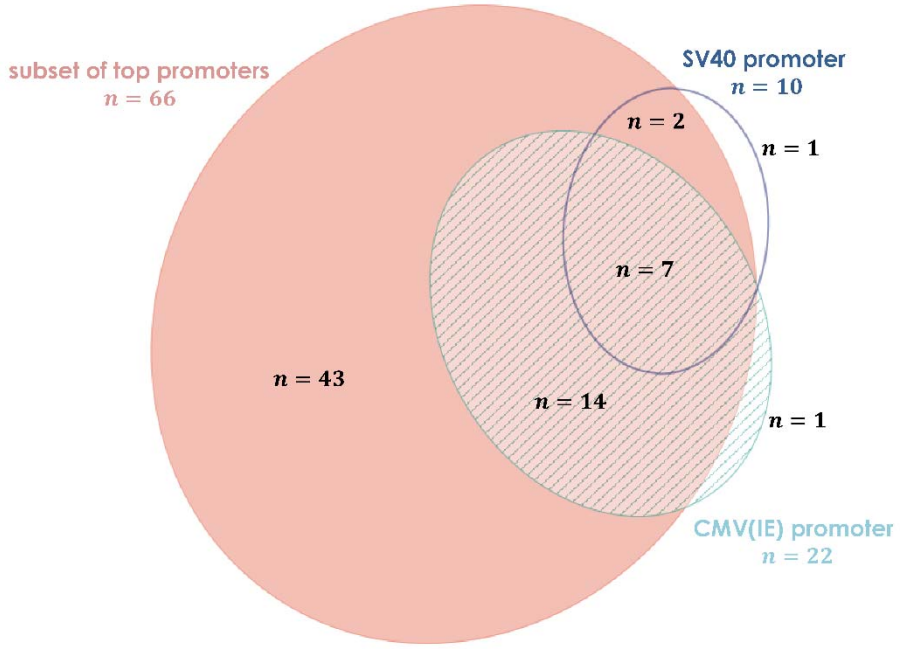
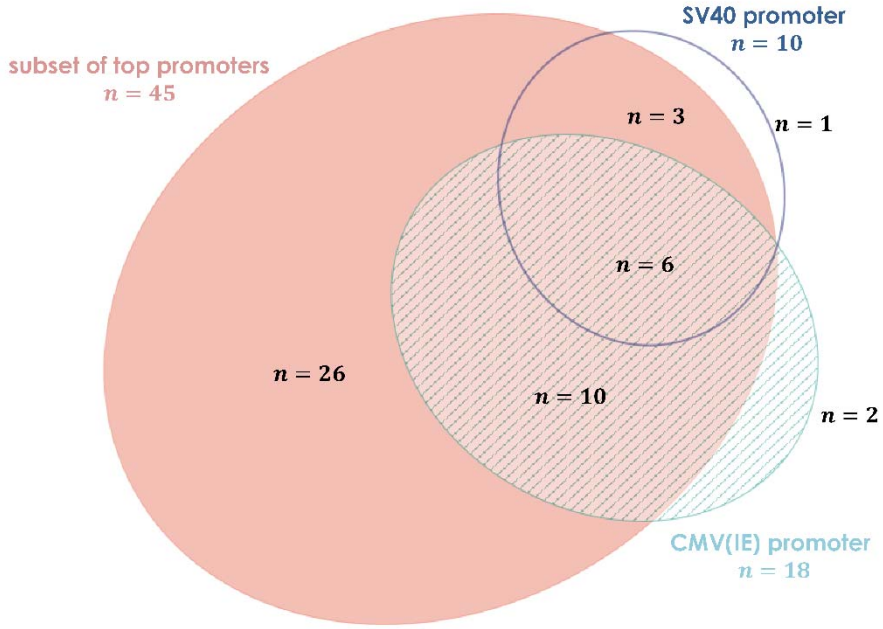


Figure 4-3, continued:
C



D

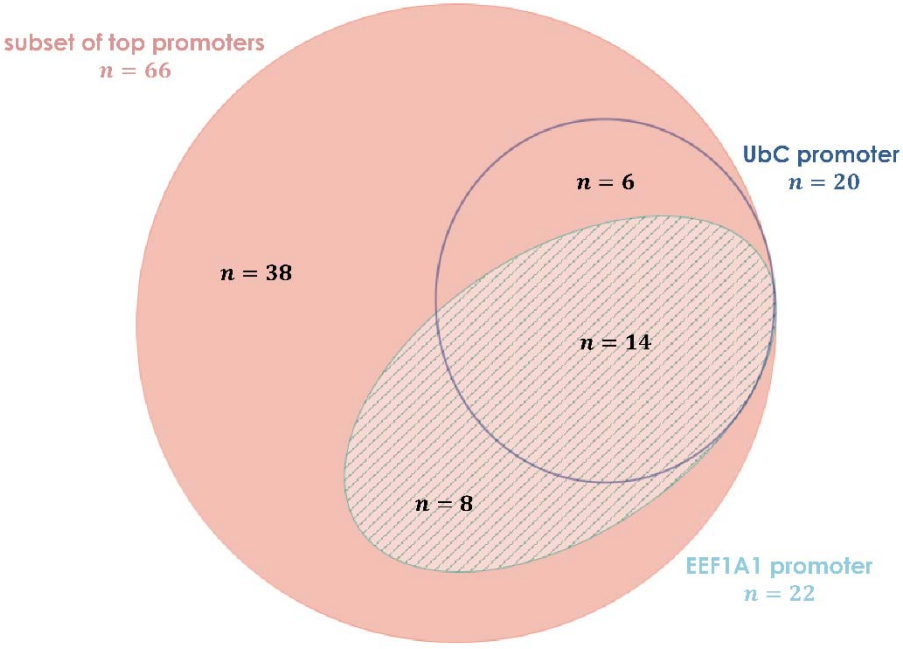
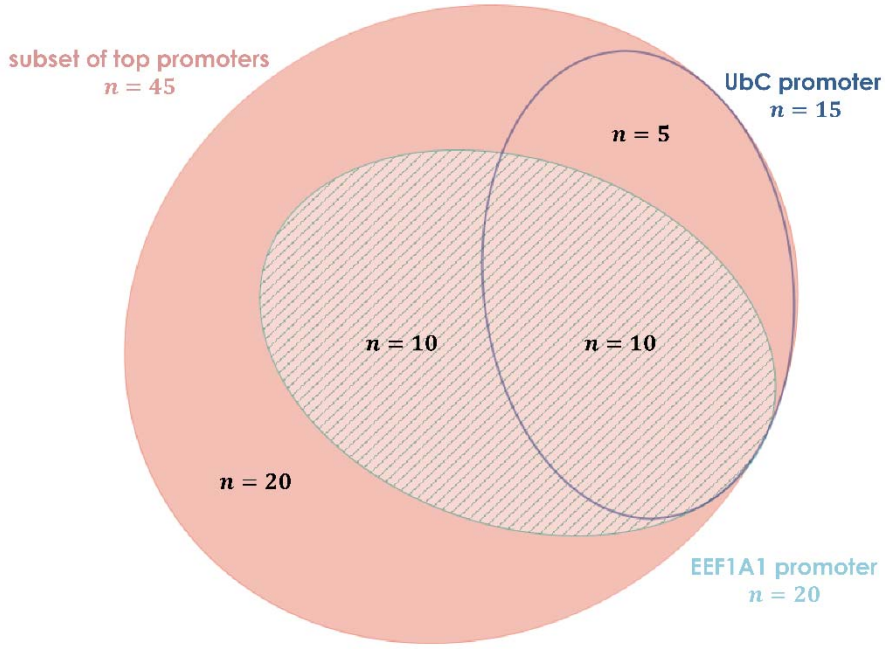


Figure 4-3, continued:
E



(**A**) Heat maps of the frequencies of putative TFBSs in a select DNA (promoter) region annotated with the JASPAR database. **top**: TFBS heat map based on 2000-bp annotation. **bottom**: TFBS heat map based on 500-bp annotation. (**B-E**) Euler diagrams of TFBSs found in viral promoters compared to the subset of top promoters based on **B)** 2000-bp annotation of the promoter region and **C)** 500-bp annotation of the promoter region, and found in commonly used endogenous promoters compared to the subset of top promoters based on **D)** 2000-bp annotation of the promoter region, and **E)** 500-bp annotation of the promoter region.

Table 4-2: Enriched TFBSs found across promoters.

2000-bp	500-bp
FOXC1	**FOXC1**
SP1 (short)	**SP1 (short)**
FOXL1	**MZF1_1-4**
MZF1_1-4	ETS1
ETS1	**FOXL1**
NFIC	NFIC
BRCA1	KLF5
ZNF354C	BRCA1
HOXA5	ZNF354C
SOX10	SPIB
SPIB	**CREB (half)**
HIF1A::ARNT	**HIF1A::ARNT**
KLF5	**HOXA5**
CREB (half)	**SP1**
FOXI1	**SOX10**
NFATC2	**SP2**
SP1	EHF
EHF	**TFAP2C**
SPI1	**THAP1**
SP2	**ELK1**
ELK1	**FOSL2**
SRY	**JUN (var.2)**
THAP1	JUN::FOS
	NF1 (consensus)
	STAT3
	YY1

List of TFBSs found to be enriched in high expression groups/subset of top promoters (green text), in background expression (red text), and no particular set association/enrichment (black text) for both sequence lengths (2000-bp and 500-bp) annotated. This enrichment is based on the TFBS frequency distribution found in **Supporting Information Figure S1a** and **S1b** of the publication[240].

4.3.3. Development and experimental testing of synthetic mammalian promoters using enriched TFBSs

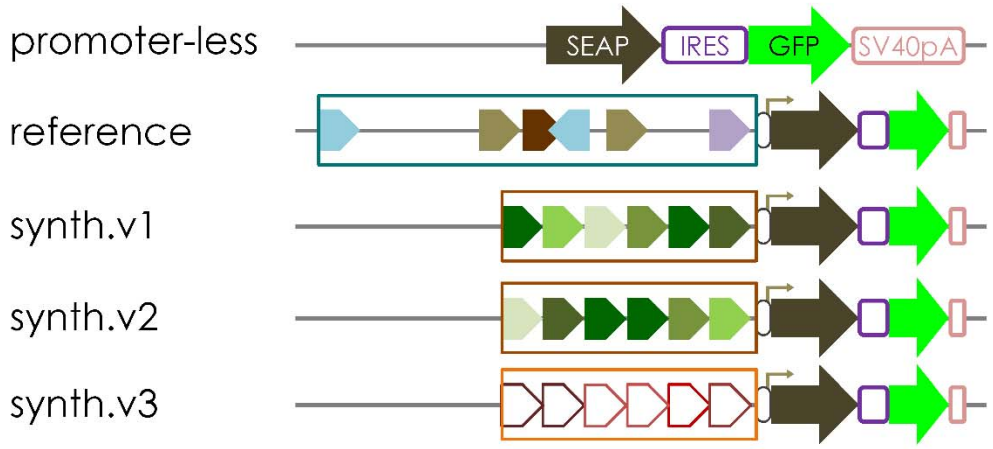
Leading to the final step of the design-build-test cycle, we created synthetic promoters based on the enriched TFBSs found in strong promoters and evaluated these promoters for their ability to drive reporter expression. The central premise of these designs is that functional, synthetic promoters can be created by concatenating TFBSs

represented in strong promoters. The corollary to this hypothesis is that weaker promoters will be created based on concatenations of TFBSs enriched in moderate or low expression and background datasets. For this D-B-T cycle, we designed and synthesized three synthetic promoter variants (**Figure 4-4**) based on the TFBSs listed in **Table 2** for subsequent testing. Two of these variants (v1 and v2) were comprised of only TFBSs enriched in the high expression group (**Table 4-2**, green text) and the third variant, v3, was comprised of only TFBSs enriched in background (**Table 4-2**, red text). Each of these designs involved randomly concatenating 19 TFBSs selected from the list (**Table 4-2**) and combining this element with the core region of hCMV IE promoter or the minimal region of the EF1α promoter from the *H. sapiens EEF1A1* gene. These TFBSs were selected based on their frequencies across the four frequency sets (top, median, random, and chromosomal, **Supporting Information Figure S1a** and **S1b** of the publication[240]): TFBSs with higher frequencies in the top promoter set (*e.g.* SP1) were included in variants v1 and v2 while TFBSs with higher frequencies in the chromosomal region or "background" (*e.g.* FOXC1) were included in variant v3. In each of these cases, we used frequency to guide the approximate ratio of the most abundant sites (*e.g.* SP1 to MZF1_1-4, FOXC1 to FOXL1).

Moreover, we did not incorporate unenriched TFBSs (*i.e.* those with similar frequencies across all four datasets (**Table 4-2**, black text)) in the synthetic variants due to the lack of association with any particular frequency set. Finally, we used the frequencies and ratios found in the 500-bp sequence window since the synthetic variants would be <500-bp. After synthesis, we tested these designs via transient transfection driving the expression of both hrGFP and SEAP reporter proteins (linked by the encephalomyocarditis virus IRES[245] as a bicistronic cassette) in the HT1080 cell line and we assayed for performance 48 hours post-transfection. In comparison to our negative controls (promoter-

less construct and HT1080 WT representing background expression levels), synthetic promoters v1 and v2 exhibited considerable functionality approaching reference (hCMV IE or full EF1α promoter) levels whereas variant v3 was barely functional (**Figures 4-5A** and **4-5B**). Furthermore, the synthetic promoter v1 seemed weaker than v2 when paired with hCMV IE promoter (**Figures 4-5A** and **4-5B**); however, we did not observe this difference when pairing these synthetic promoters with the minimal EF1α promoter despite seeding the same number of TFBSs for both synthetic designs. Annotation of the enhancer/proximal promoter region of these synthetic promoters using the same JASPAR database indicated that these 3 synthetic variants contained similar quantities of putative binding sites (53 for v1 and v2, 52 for v3; **Supporting Information Table S4a to S4d** of the publication[240]), thus the variable expression of both reporters reflecting the strength of these promoters were not due to a disparity in potential TFBSs. These expression patterns suggest some context dependency of the core promoter region and the importance of TFBS ordering in promoter function. Therefore, it is necessary to iterate through the workflow by incorporating TFBS arrangement considerations to further optimize synthetic promoter sequences. Nevertheless, these results highlight the premise that combining TFBSs enriched in sequences correlated with strong expression can act as a potent enhancer/proximal promoter region and result in functional promoters when coupled with a core promoter element. Furthermore, we have proven the corollary to the premise that coupling TFBSs enriched in background expression levels yielded a promoter sequence (v3) with minimal activity.

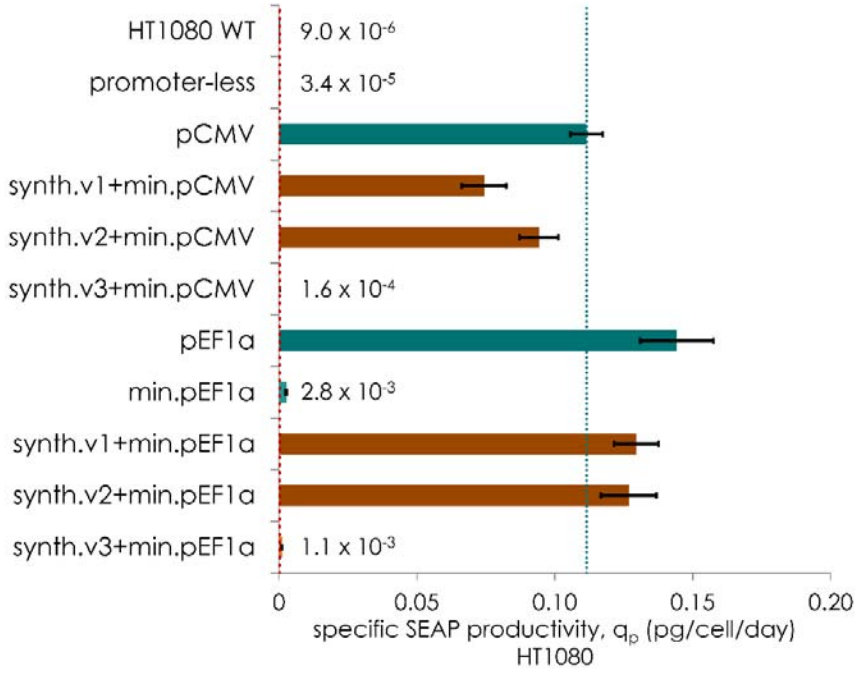
Figure 4-4: Schematic of synthetic promoter configurations used to drive dual-reporter expression.



Pentagons represent putative TFBSs based on consensus sequence annotation using **Supporting Information Table S2** of the publication[240]. The reference sequence of human CMV promoter (dark green) is equivalent to accession M60321.1, nucleotides 1-2105. In particular, putative TFBSs in the synth.v1 and synth.v2 promoters (in shades of green encompassed by dark orange/brown box) reflect those sites that are enriched in the annotated promoter regions of highly expressed genes, whereas the putative TFBSs in the synth.v3 promoter (outline in shades of red encompassed by light orange box) reflect those sites that are enriched in an annotated region of minimal/background expression levels.

Figure 4-5: Transient expression of two reporter proteins in HT1080 cells at 48-h and HEK293F cells at 16-h post-transfection using synthetic promoter variants.

A



B

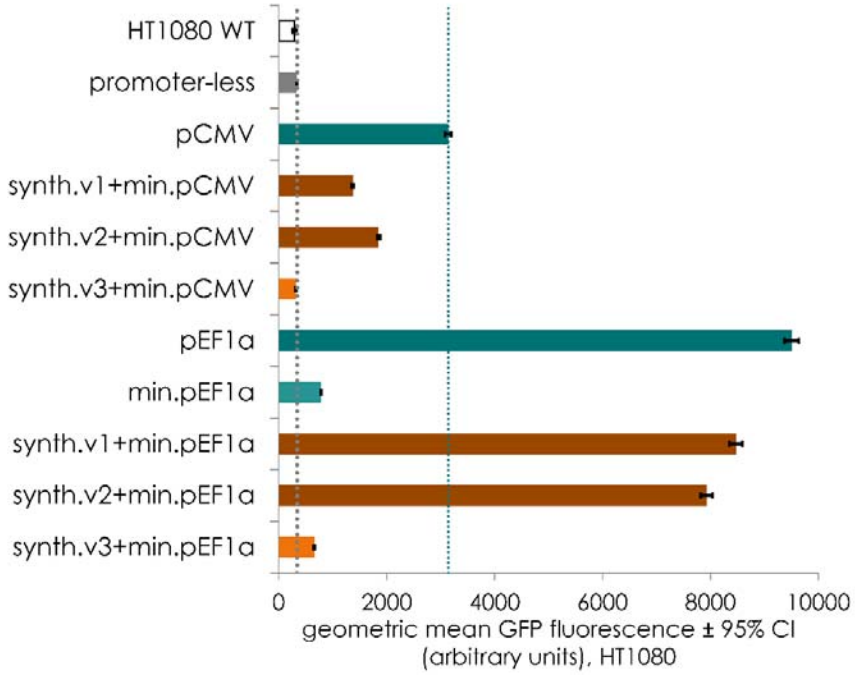
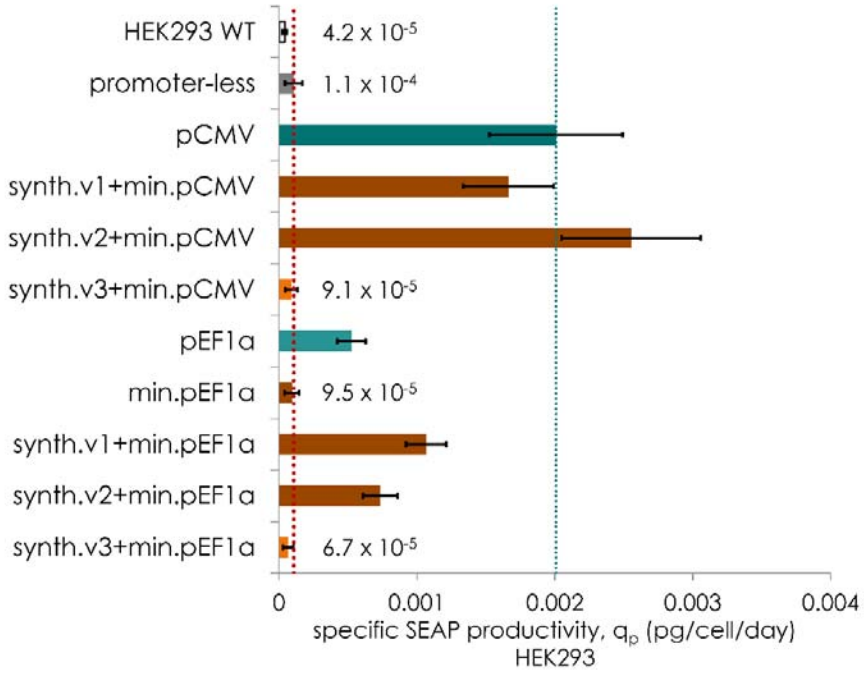


Figure 4-5, continued:
C



D

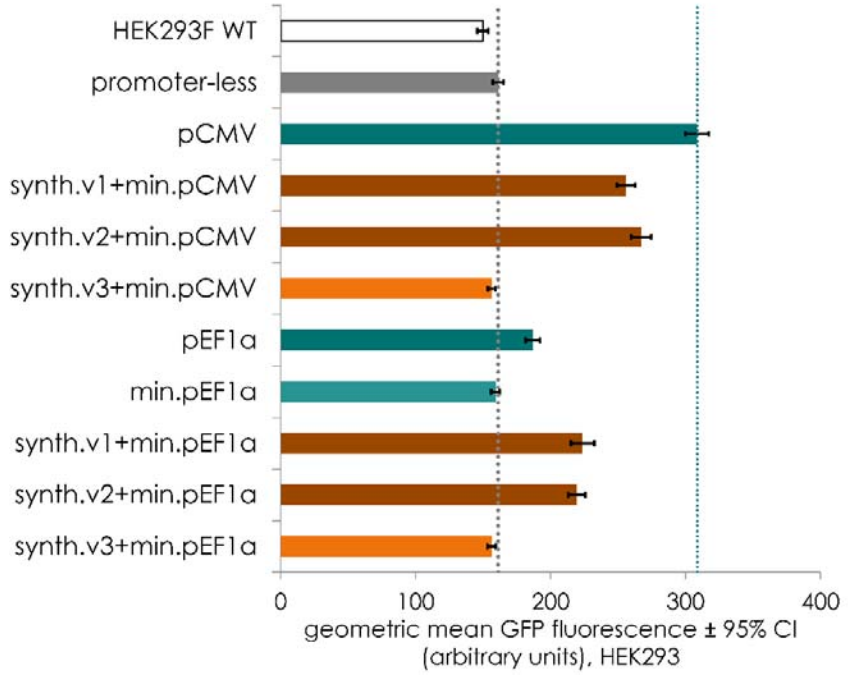


Figure 4-5, continued:
A) Expression of model secreted protein, secreted alkaline phosphatase (SEAP) from the first cistron of our dual-reporter construct driven by promoters described in **Figure 4-4**. Values represent average specific productivity over 48-h from 3 independent transfections and error bars represent the 95% confidence interval of the average specific productivity. **B)** Expression of model fluorescent protein (humanized Renilla green fluorescent protein, hrGFP) from the second cistron of our dual-reporter construct driven by promoters described in **Figure 4-4**. Values represent geometric mean fluorescence intensity from 3 independent transfections and error bars represent the 95% confidence interval of the geometric mean. **C)** SEAP expression and **D)** hrGFP expression from a single transfection via nucleofection in HEK293F cells quantified at 16-h post-transfection.

To briefly assess whether synthetic promoter designs for a particular host cell line can function in another, we measured transient SEAP and hrGFP expression using the same expression vectors at 16-h post-transfection via nucleofection in HEK293F cells. While the absolute expression levels were lower in HEK293F than HT1080 (compare **Figures 4-5A** with **4-5C**, **4-5B** with **4-5D**), the general trend of our synthetic promoters v1 and v2 having comparable expression to the reference promoters (hCMV IE/pCMV and pEF1α) is maintained even in HEK293F cells (**Figures 4-5C** and **4-5D**). Interestingly, while this workflow specifically analyzed the native expression profile of HT1080 cells and used this analysis to design promoter sequences for driving expression in HT1080 cells, they were able to function properly in an alternate host cell line. Both reference promoters showed comparable performance between the *H. sapiens* HT1080 and HEK293 cell lines[224], therefore we expected some cross-functionality and that similar transcription factors are recruited to the reference and synthetic promoters in these cell lines. Thus, this workflow can be readily applied to other host cell lines to generate functional *de novo* synthetic promoters using representative native expression data.

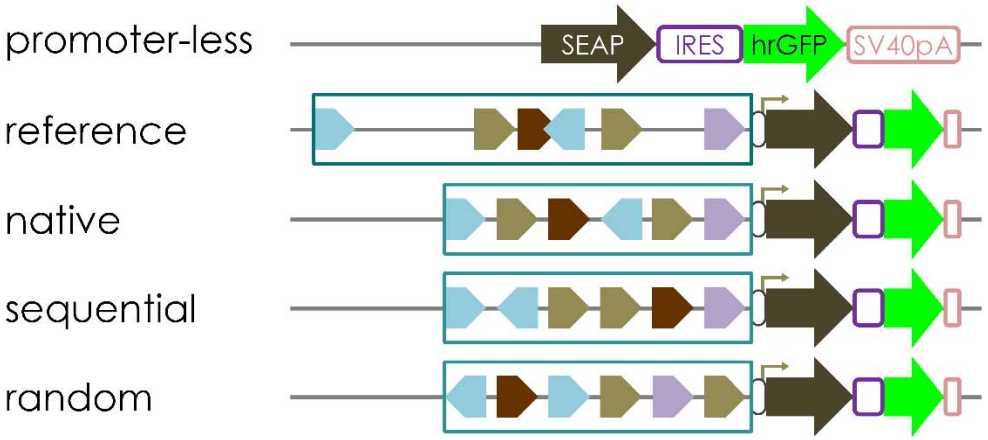
Simply gaining insight into which TFBSs may yield functional *de novo* synthetic promoters is insufficient for obtaining optimal designs in a single D-B-T cycle since the arrangement of these binding sites (such as their spacing, orientation, and order or adjacent

binding sites) is poorly understood and difficult to establish *a priori*. The large body of information found in transcription factor databases offers little insight into their overall arrangement (except potentially for known interactions, *e.g.* the AP-1 complex[246]). However, it is clear that the arrangement of the TFBSs used in a promoter can impact promoter activity as demonstrated in this work. Previous attempts at creating synthetic promoters relied on concatenations of TFBSs whether with intentional[5, 107] or random[247] arrangement of the TFBSs. Furthermore, the spacing between TFBSs is known to influence promoter strength[215]. Understanding the rules guiding TFBS arrangement will greatly optimize *de novo* promoter design and inform subsequent iterations of the D-B-T cycle.

To briefly investigate how TFBS arrangement can impact promoter functionality, we created three synthetic variants of the conventional viral-derived hCMV IE promoter (M60321.1) that have altered spacing and order of its annotated putative TFBSs (**Figure 4-6A**). These promoter variants were randomly generated using a computational algorithm[248] to contain 45% GC-content, approximately mimicking the overall GC-content found in the human genome and reference promoter sequence. The three synthetic hCMV IE promoter variants (native, sequential, and random order) all have annotated TFBSs spaced 10-bp apart. The reference hCMV IE promoter and its variants were used to drive the transient expression of the SEAP and hrGFP reporter proteins in HT1080 cells. Based on the SEAP reporter expression 48 hours post-transfection, both spacing and order of TFBSs impacts promoter strength (**Figure 4-6B**). Similarly, analyzing the geometric mean fluorescence intensity of the hrGFP reporter over the same 48-h period (**Figure 4-6C**) corroborates the trends observed with the SEAP reporter (**Figure 4-6B**).

Figure 4-6: Transient expression of two reporter proteins in HT1080 cells at 48-h post-transfection using synthetic hCMV-IE promoter variants.

A



B

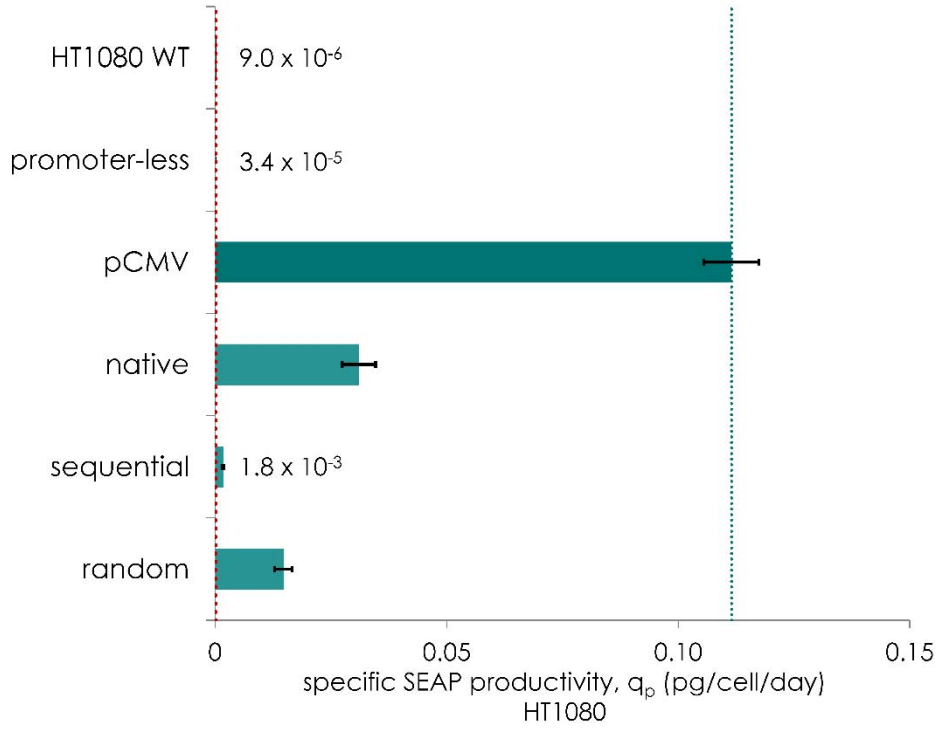
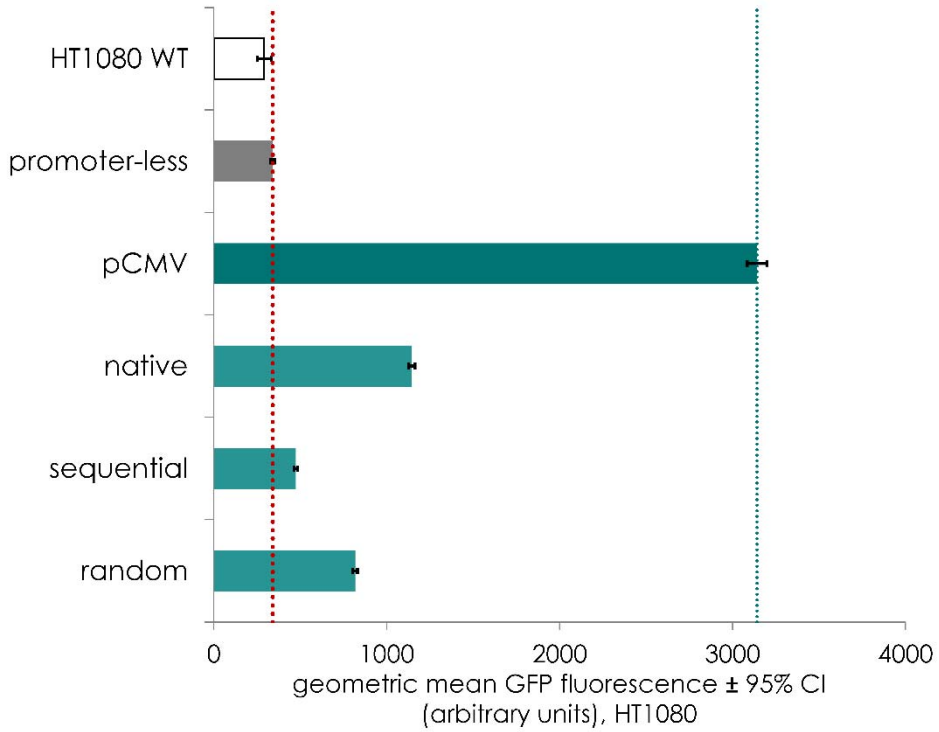


Figure 4-6, continued:
C



A) Schematic of synthetic hCMV-IE promoter variants used to drive our dual-reporter expression in this work. Pentagons represent putative TFBSs based on consensus sequence annotation using Supporting Information Table S2. The reference sequence of human CMV promoter (dark) is equivalent to accession M60321.1, nucleotides 1-2105. The 3 hCMV IE promoter variants with each TFBS spaced 10-bp apart exhibited reduced activity (light) and were constructed with the same identity and quantity of TFBSs as the reference sequence (dark). **B)** Expression of model secreted protein, secreted alkaline phosphatase (SEAP) from the first cistron of our dual-reporter construct driven by promoters described in **A**. Values represent average specific productivity over 48-h from 3 independent transfections and error bars represent the 95% confidence interval of the average specific productivity. **C)** Expression of model fluorescent protein (humanized Renilla green fluorescent protein, hrGFP) from the second cistron of our dual-reporter construct driven by promoters described in **A**. Values represent geometric mean fluorescence intensity from 3 independent transfections and error bars represent the 95% confidence interval of the geometric mean.

Interestingly, the data suggest that the spacing between annotated TFBSs exhibited greater influence on promoter strength (compare native in light bar with reference in dark bar) than the particular arrangement of the TFBSs themselves (compare native, sequential,

random in light colored bars) (**Figures 4-6B** and **4-6C**). We selected the 10-bp spacing to represent an average of the spacing between annotated TFBSs in the reference hCMV IE promoter and to approximate nearly a full turn of the DNA helix to facilitate interaction of transcription factors with their cognate binding sites, although periodic expression may result from using tandem repeats of certain TFBSs[249]. However, by creating a pre-defined static distance between putative TFBSs, it is likely that some synergistic interactions between TFs are disrupted[246, 250], ultimately reducing the promoter strength. The reduction in promoter length due to altered spacing between TFBSs does not fully account for its impact on promoter strength; the synthetic variants v1 and v2 (**Figures 4-5A** and **4-5B**) were stronger than the hCMV IE promoter variants (**Figures 4-6B** and **4-6C**) with an even shorter promoter sequence footprint overall (341-bp vs. 1301-bp for reference hCMV IE and 702-bp for hCMV IE variants). Thus, the spacing, order, and composition of TFBSs are important for generating functional synthetic promoters and these parameters can be used to further tune the strength of synthetic promoters in subsequent cycles of the workflow described in this work. For *de novo* synthetic sequences, it is straightforward to generate random designs as the first cycle and subsequent cycles of this workflow can investigate the spacing between TFBSs in these designs. With the prevalence targeted genome editing tools and techniques[251], the initial synthetic promoter designs can be refined with subsequent workflow cycles to obtain optimal functionality in the desired genomic context.

4.4. CONCLUSIONS

This work demonstrated the utility of an 'omics guided workflow to create novel, synthetic promoters for a mammalian cell host. These diverse, synthetic sequences comprise a suite of non viral-derived sequences, potentially reducing their susceptibility of

epigenetic silencing and increasing long-term stability. By adjusting the TFBS composition and arrangement, a set of *de novo* promoters can be developed to properly tune the protein intermediates in a metabolic pathway of mammalian hosts as one of the tools governing transgene expression[252]. Moreover, these synthetic promoters had an overall reduced sequence footprint compared to reference promoters, thus increasing the overall utility for applications such as (heterologous) metabolic pathway designs[6] and gene delivery vector designs[253], akin to efforts in reducing the regulatory sequence footprint in a conventional metabolic engineering host[254]. It is possible to expand the combinatorial space in this work to enable many permutations of synthetic randomized concatenations of a particular set of TFBSs to be constructed and evaluated[255]. Ultimately, this workflow incorporates contemporary high-throughput methodologies to construct functional promoter elements that can facilitate applications ranging from the study of fundamental processes to immediate use in large-scale industrial processes.

Chapter 5: Creating the message – alternative approaches to promoter engineering

5.1. CHAPTER SUMMARY

Chapter 4 established an approach to rationally create *de novo* promoters using endogenous expression information, but other promoter engineering approaches can exploit the endogenous sequences and their activity. Traditionally, promoter engineering involved isolation and testing of segments of endogenous DNA sequences, which are often viral-derived. These segments would be further dissected to identify essential and/or superfluous DNA regions for promoter activity through the expression of a reporter gene. This breakdown of promoter segments is typically comprised of an enhancer region, a proximal promoter region, and a core promoter region[256]. The work described in this chapter investigated approaches to engineer these 3 regions to create hybrid promoters. In doing so, we successfully generated hybrid promoter variants based on endogenous sequences that are comparable to a strong viral-derived promoter in expression and concomitantly revealed that the 5' UTR, in particular the intron preceding the coding sequence, can drastically affect promoter strength in mammalian cells.

5.2. INTRODUCTION

As described in Chapter 4, a promoter typically includes the core promoter, proximal promoter, and distal enhancer elements. Many previous efforts have explored the promoter activity of isolated putative promoter sequences that encompass all three elements in order to characterize their ability to drive heterologous gene expression[222, 224, 257-262]. To this end, only a handful of sequences are predominantly used based on the desire to drive high levels of expression[224] and these sequences are often viral-derived. However, the foreign nature of these viral-derived sequences can elicit cellular responses (*e.g.* silencing) that reduce or remove their transcriptional activity[226, 229], leading to unstable

gene expression over time. Nonetheless, the initial characterization of these promoters facilitated the subsequent derivatization to incorporate novel functionality such as response to heavy metals[247], and paved the foundation for rational library screening approaches to identify stronger promoter variants[5, 107, 215, 263]. Similar efforts utilizing both rational library screens and synthetic approaches created sets of sequences suitable for fine-tuned regulation of gene expression to enable metabolic engineering of microbial systems[110, 111, 248, 254, 264], and this work adapts some of these methodologies to create hybrid mammalian promoters.

In particular, the core promoter region interacts with RNA polymerase II to facilitate downstream transcription and this region can be dissected into many critical components[265, 266]. These components (TATA box, Initiator, general transcription factor binding sites such as TFIIB, TFIID, etc.) can be rationally combined to create synthetic core promoters with a substantial increase in activity compared to endogenous sequences[267]. Therefore, we investigated the premise that functional core promoters can be synthetically created and that multiple core promoters could increase gene expression in the HT1080 mammalian host cell.

To further expand the limited set of non-viral derived promoters that are functional in mammalian hosts, we adopted an alternative to the bioinformatics approach described in Chapter 4. Again, by harnessing the native expression program, we can pinpoint precisely which genes are highly expressed and therefore explore their corresponding 5' regulatory sequence responsible for the high expression levels. The use of this guided approach is based on the same hypothesis in Chapter 4: highly expressed endogenous genes are controlled by strong promoters.

The work described here is akin to previous work characterizing the functionality of commonly used promoters[222, 224, 257-262] and require a fully sequenced genome. However,

these regulatory regions are often poorly annotated at best in most host cell genomes, and the work described here explores some of these sequences and their ability to drive heterologous gene expression. Another critical aspect of gene regulation in higher eukaryotes is the abundance of introns in their coding DNA sequences and in their 5' UTR, and these introns can critically impact promoter functionality[268-273]. By incorporating both native and synthetic intron sequences into promoter designs, we explored their impact on heterologous gene expression as an additional layer of control embedded within a promoter.

5.3. RESULTS AND DISCUSSION

5.3.1. Evaluating core promoter designs for increasing heterologous gene expression

The previous rational design of a super core promoter was comprised of a variety of commonly used core promoter elements from the CMV-IE viral gene, adenovirus major-late viral gene, and *D. melanogaster* sequences[267]. Subsequent characterizations of additional core promoter elements suggest that novel core promoter designs can be tailored to specific applications to elicit the desired expression[274-276]. While RNA polymerase II machinery is highly conserved across eukaryotic species, the elements of the core promoter and mode of transcription initiation by a specific core promoter can vary, especially in vertebrates[266, 277].

We revisited the design of the super core promoter and created a variant that included two additional elements that interact with TFIIB[278] and three additional elements from the major histocompatibility complex (MHC) class I genes core promoter[279] dubbed UCP shown below (**Figure 5-1**). To assess the impact of coupling these two core promoters with an enhancer region, we generated 3 enhancer variants derived from the hCMV-IE enhancer (**Figure 5-2**). We compared the ability of this novel core promoter

against the core promoter from the hCMV-IE gene (cpCMV) to drive gene expression in HT1080 cells by measuring the fluorescence intensity of the hrGFP reporter protein. We quantified transient expression at 24h, 48h, and 72h post-transfection, and found insignificant differences between the UCP or cpCMV (**Figure 5-3**). Furthermore, there were insignificant differences when these core promoter elements were coupled to the CMV enhancer variants (**Figure 5-3**). As expected, since these two core promoters are statistically indistinguishable, the coupling of the hCMV-IE enhancer to either core promoter recapitulated the functionality of the full hCMV-IE promoter (pCMV, **Figure 5-3**). While this particular UCP core promoter design did not demonstrate an improvement over cpCMV in its ability to drive gene expression, the comparable expression suggested that some of these key elements are dispensible. Therefore, future designs that would fully replace the viral-derived sequences can serve as alternatives if a particular application, such as for gene therapy or immunotherapy, requires an avoidance of viral-derived sequences.

Figure 5-1: Conserved elements of the core promoter derived from hCMV-IE (cpCMV) and the rationally designed UCP.

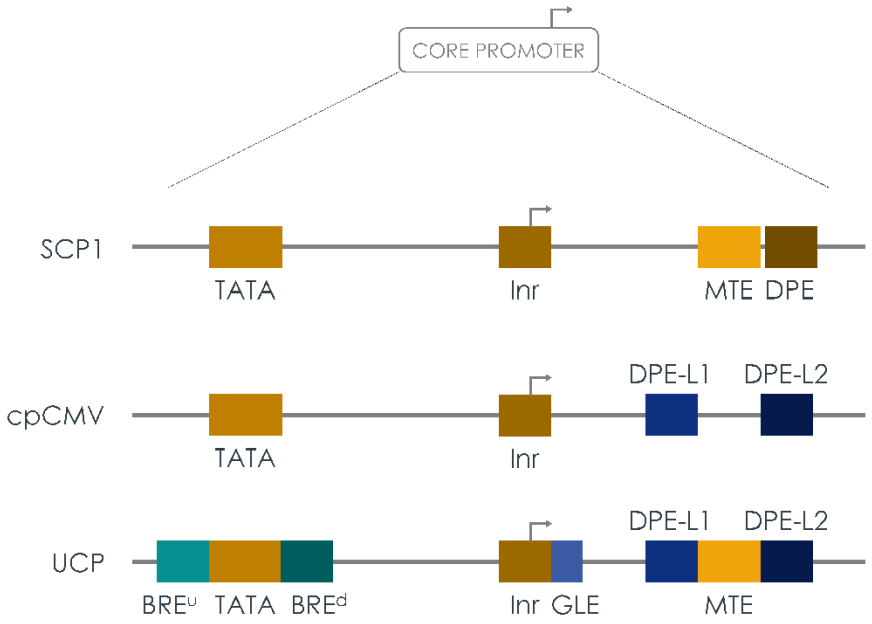


Figure 5-2: Enhancer variants of the hCMV-IE gene used in conjunction to evaluate two core promoter designs.

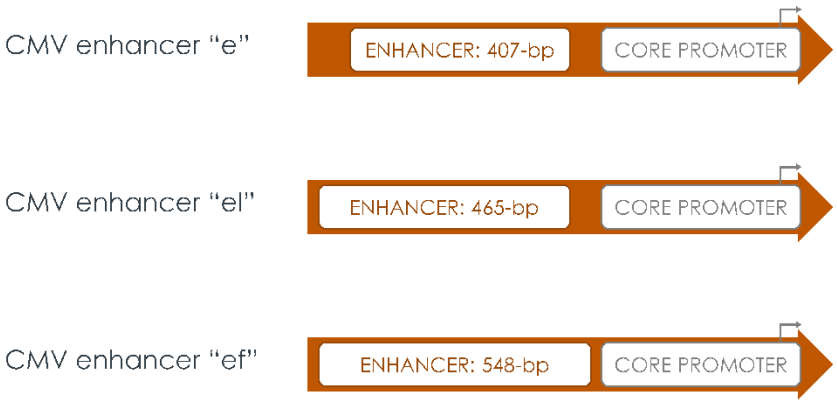
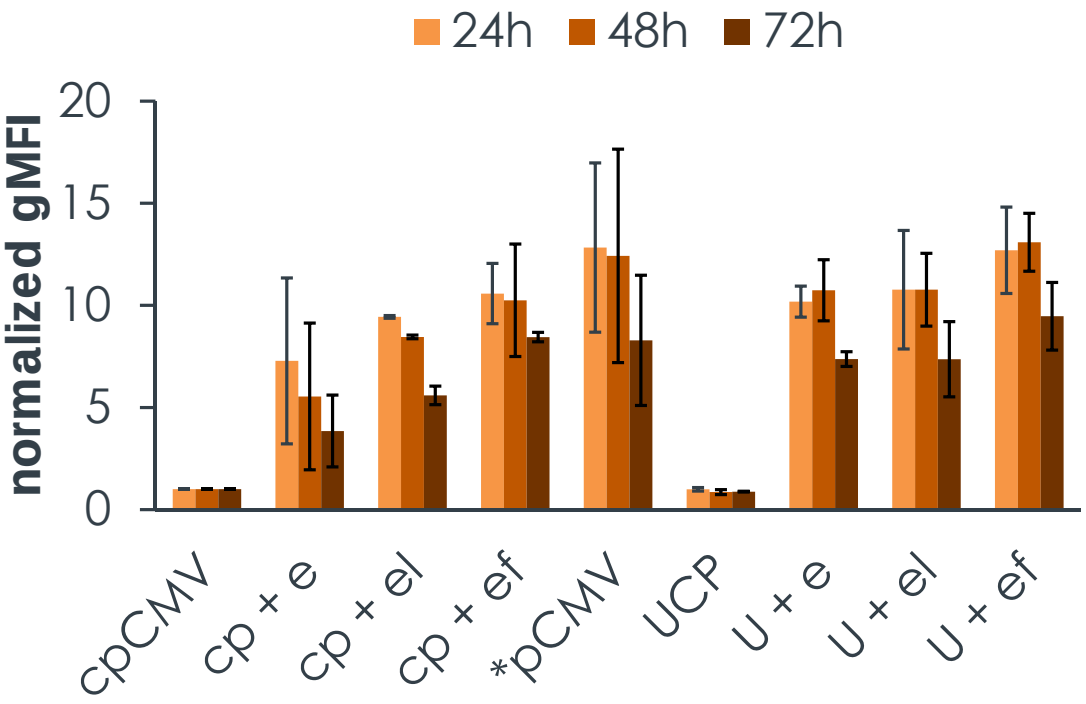


Figure 5-3: Normalized geometric mean fluorescence intensity (gMFI) of hrGFP driven by two core promoter designs

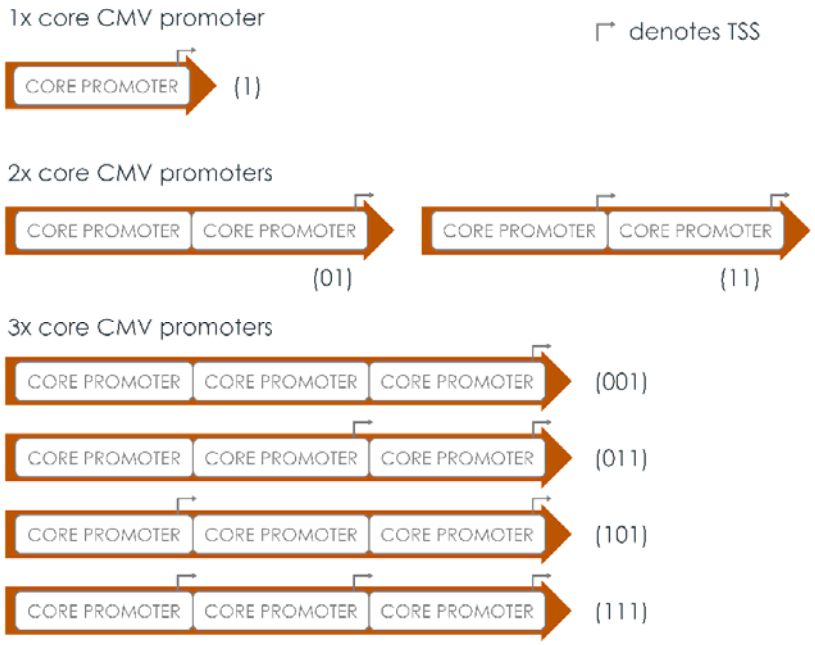


Core promoters along with their coupling with enhancer variants derived from hCMV-IE gene at 24h (light orange), 48h (orange), and 72h (brown) post-transfection.

Inspired by the work in *E. coli* to boost gene expression through using tandem core promoters[264], we explored the possibility of using multiple core promoters with defined TATA boxes and with/without initiator sequences to improve heterologous gene expression in mammalian production hosts. It is expected that with the multiple initiator sequences in these core promoter variants that there would be multiple focused transcription start sites. Despite this 5' UTR heterogeneity, there would be a net increase of the transcript leading to increased expression levels when compared to a single core promoter. Based on the extensive work that characterized the core promoter from the viral hCMV-IE gene, we created 6 additional multiple-core promoter variants from this well-

studied sequence to incorporate permutations of including/excluding the initiator sequence (*i.e.* a transcription start site) for each additional core promoter sequence (**Figure 5-4**).

Figure 5-4: Multiple core promoter variants evaluated for their transcription capacity.

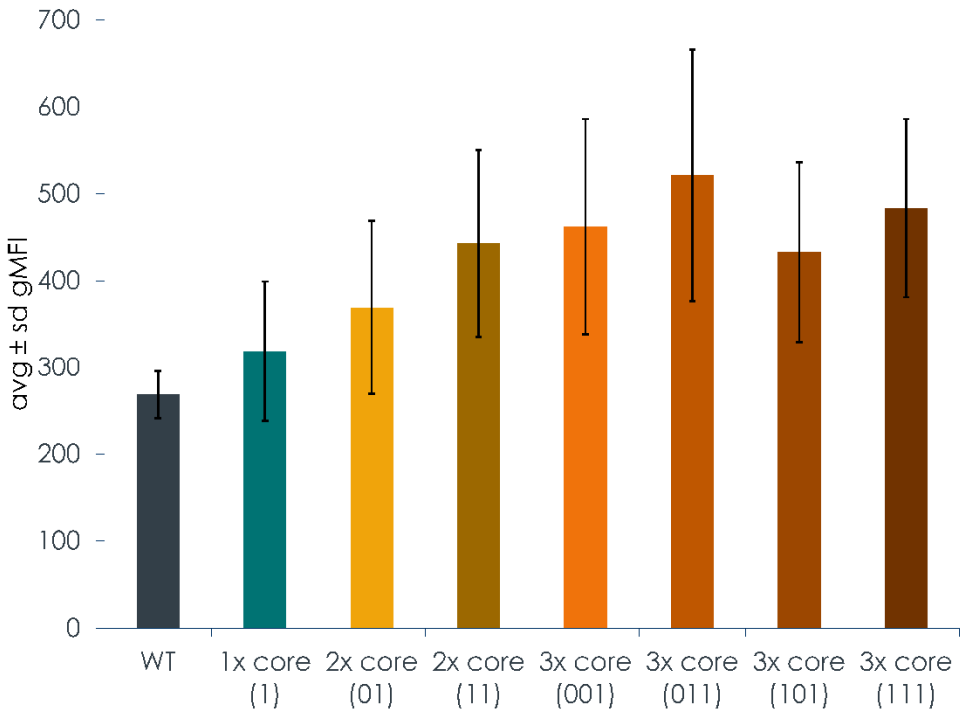


The extra gray right angle arrow corresponds to an expected transcription start site, denoted as "1" in the coded representation of the promoter.

We transfected our model host cells, HT1080, with these core promoter variants driving the expression of our fluorescent model protein hrGFP. Interestingly, our data did not suggest that the core promoter configurations had any impact on transient hrGFP expression as quantified by flow cytometry analysis 48h post-transfection (**Figure 5-5**). However, we analyzed the proportion of hrGFP expressing (positive) and non-expressing (negative) cells post-transfection, the data suggested that there was an increase in the proportion of cells that are expressing hrGFP (**Figure 5-6**). By combining this approach with novel core promoter designs, this particular component of a promoter can be easily tailored to a particular application, and these core promoters can be independently coupled

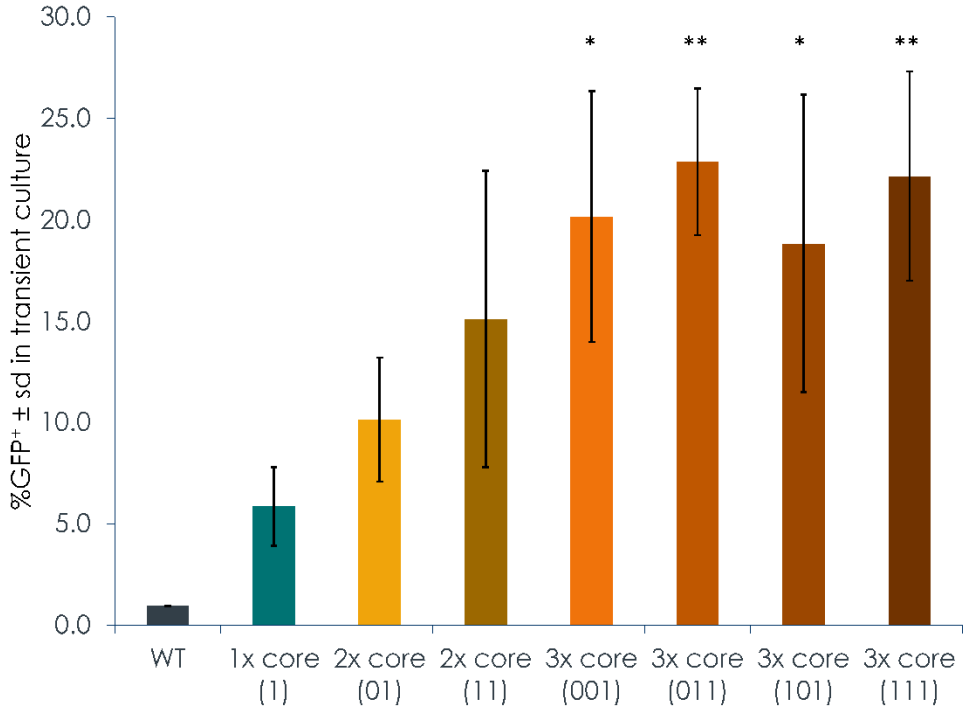
to the enhancer regions produced by the Design-Build-Test workflow described in Chapter 4.

Figure 5-5: hrGFP transient expression measured from HT1080 cells 48h post-transfection



The average geometric mean fluorescence intensity (gMFI) is reported and accounting for the standard deviation in this measurement (represented by the error bars), there is no statistically significant difference when multiple core promoters are used to drive gene expression. WT corresponds to the autofluorescence level from HT1080 cells.

Figure 5-6: The same samples from **Figure 5-5** represented by the percentage of the analyzed cell population that is expressing hrGFP.



By incorporating additional core promoter elements, a higher proportion of cells are expressing the reporter protein, albeit at the same expression level (see **Figure 5-5**). WT corresponds to the autofluorescence level from HT1080 cells. *denotes $p < 0.1$ and **denotes $p < 0.05$ relative to the single core promoter (1x core).

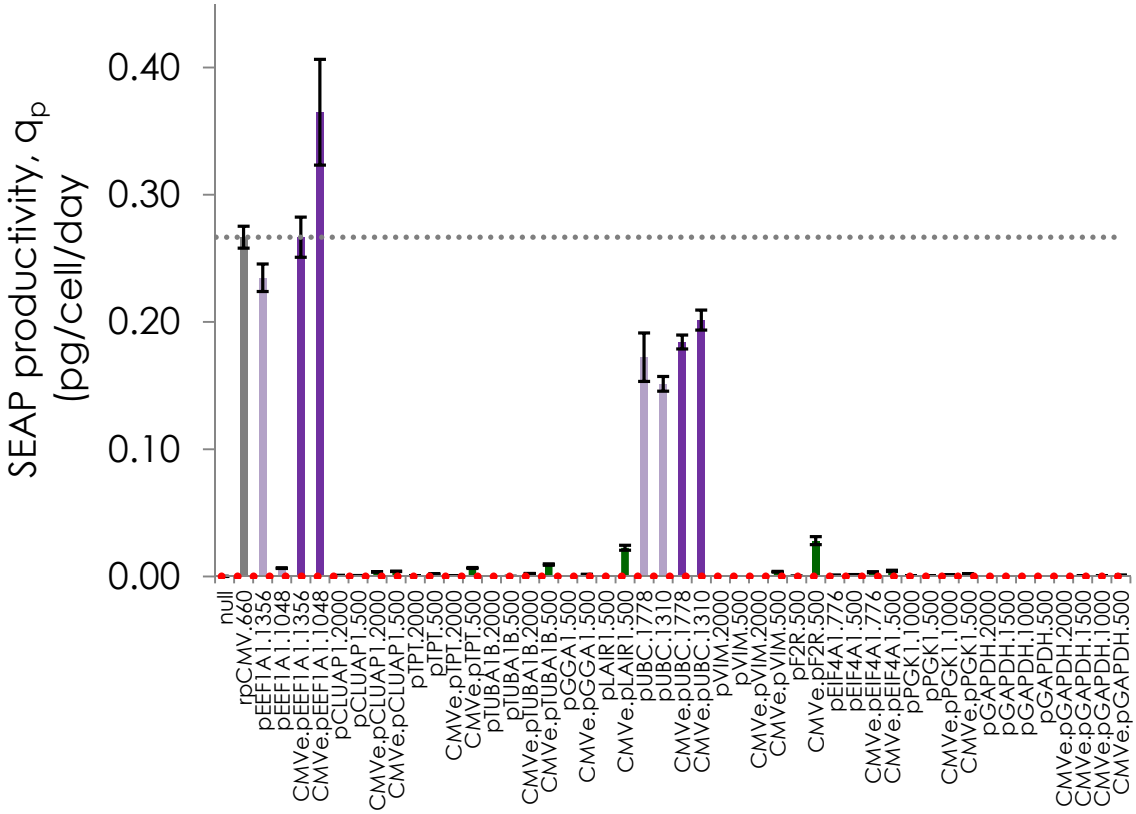
5.3.2. Rational design of endogenous hybrid promoters

Although the core promoter region is essential for interacting with RNA polymerase II and basal transcription factors to facilitate transcription[277], the proximal promoter/enhancer and distal enhancer regions harbor additional transcription factor binding sites that interact with the core region to boost transcriptional activity[256]. By leveraging the expression data of the HT1080 cells analyzed in Chapter 4, we can screen endogenous 5' UTR sequences from the same highly expressed genes for promoter activity. This is motivated by the same hypothesis that highly expressed genes have strong promoters responsible for driving that expression level. Since conventionally used

promoter sequences vary from 500-bp to over 2000-bp[224], we extracted versions of the 5' UTR from 12 endogenous genes of the human genome (GRCh38.p2 build) that include approximately 500-bp or 2000-bp 5' of the annotated TSS and 40-bp 3' of the annotated TSS to include a portion of the first transcribed exon (see **Supporting Information** from publication[240] described in Chapter 4), retaining the endogenous core promoter. These putative promoters correspond to the regions annotated for bioinformatics analysis in Chapter 4.

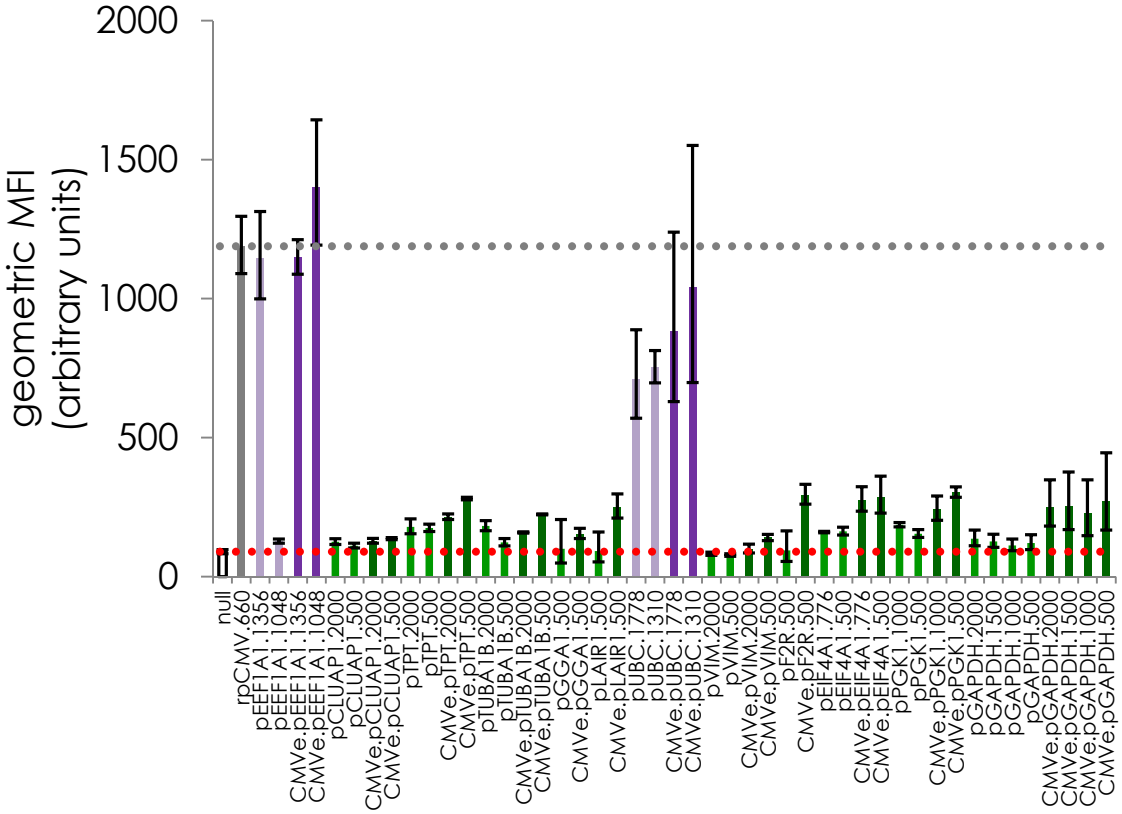
These putative sequences were cloned into the same vector used in Chapter 4 that would drive the expression of a secreted reporter protein (SEAP) and a fluorescent protein reporter (hrGFP) simultaneously. We quantified the ability of these putative promoter sequences to drive gene expression transiently 48h post-transfection using an *in-vitro* assay for detecting SEAP production (NovaBright™ SEAP Enzyme Reporter Gene Chemiluminescent Detection System 2.0, Invitrogen) and flow cytometry to measure fluorescence intensity (BD LSRII Fortessa, 200V). To further evaluate the compatibility of these putative promoters with a well-characterized enhancer element, these endogenous promoter sequences were also coupled with the enhancer region from hCMV-IE (576-bp enhancer corresponding to 534-1109 from accession number M60321.1). Despite the high expression of these endogenous genes in HT1080 as measured by microarray analysis (Shire Genetic Therapies), only a few of these 5' UTR contained sequences capable of driving strong expression (**Figures 5-7** and **5-8**). Only the promoter region from the previously characterized *EEF1A1* gene[261] and *UBC* gene[269] demonstrated activity that is comparable to the viral hCMV-IE promoter (rpCMV.660).

Figure 5-7: Transient SEAP productivity (expression) 48h post-transfection in HT1080 driven by putative promoters derived from highly expressed genes.



Error bars represent 95% CI of SEAP productivity from 3 or more biological replicates. The purple fill color denotes promoters with intron sequences included in the promoter region. The numbers following the promoter name denotes the promoter length. CMVe denotes promoter sequences with the hCMV-IE enhancer coupled, represented by the darker fill color. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

Figure 5-8: Transient hrGFP expression 48h post-transfection in HT1080 driven by putative promoters derived from highly expressed genes.

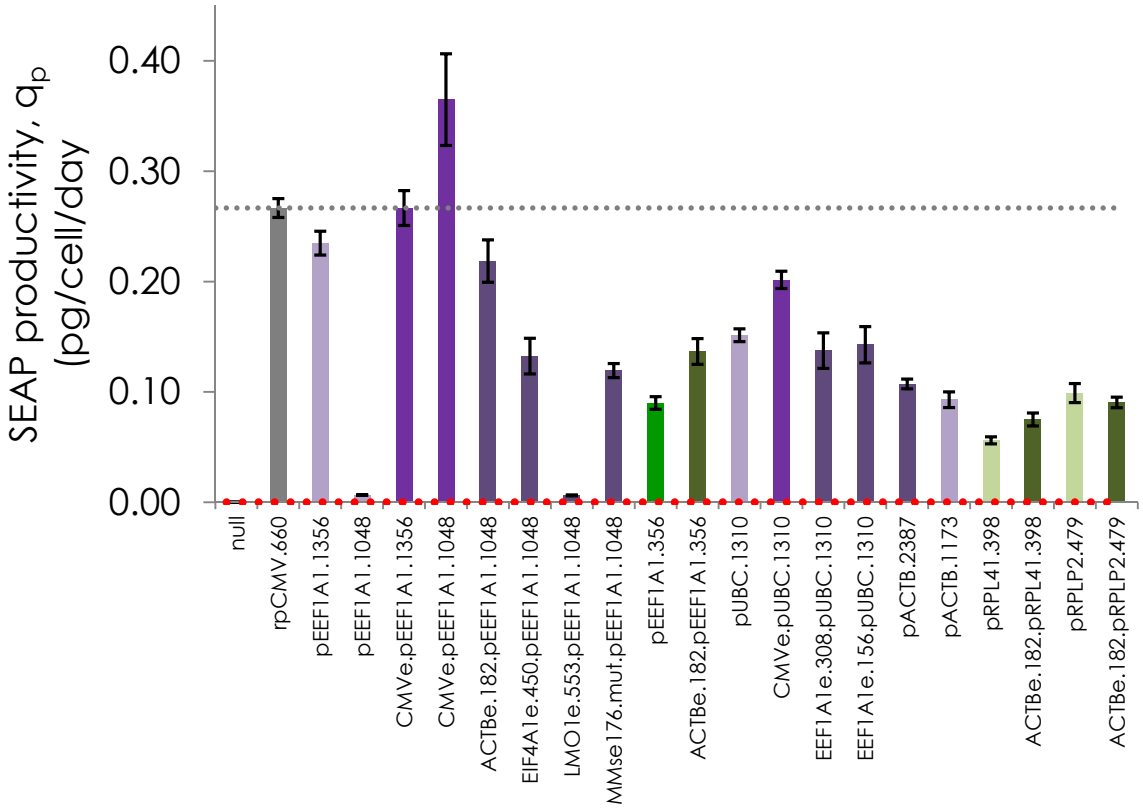


Error bars represent standard deviation of geometric mean fluorescence intensity from 3 or more biological replicates. The purple fill color denotes promoters with intron sequences included in the promoter region. The numbers following the promoter name denotes the promoter length. CMVe denotes promoter sequences with the hCMV-IE enhancer coupled, represented by the darker fill color. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

The results in **Figures 5-7** to **5-8** suggest that most of the endogenous promoter sequences are activated by the CMV enhancer. Therefore, despite already containing an endogenous enhancer region, these promoters can be further activated with an additional enhancer element. We evaluated additional putative endogenous promoter sequences derived from the beta-actin promoter[270, 280] and two ribosomal proteins identified from our microarray expression data set (*RPL41* and *RPLP2*). We also investigated pairings of a

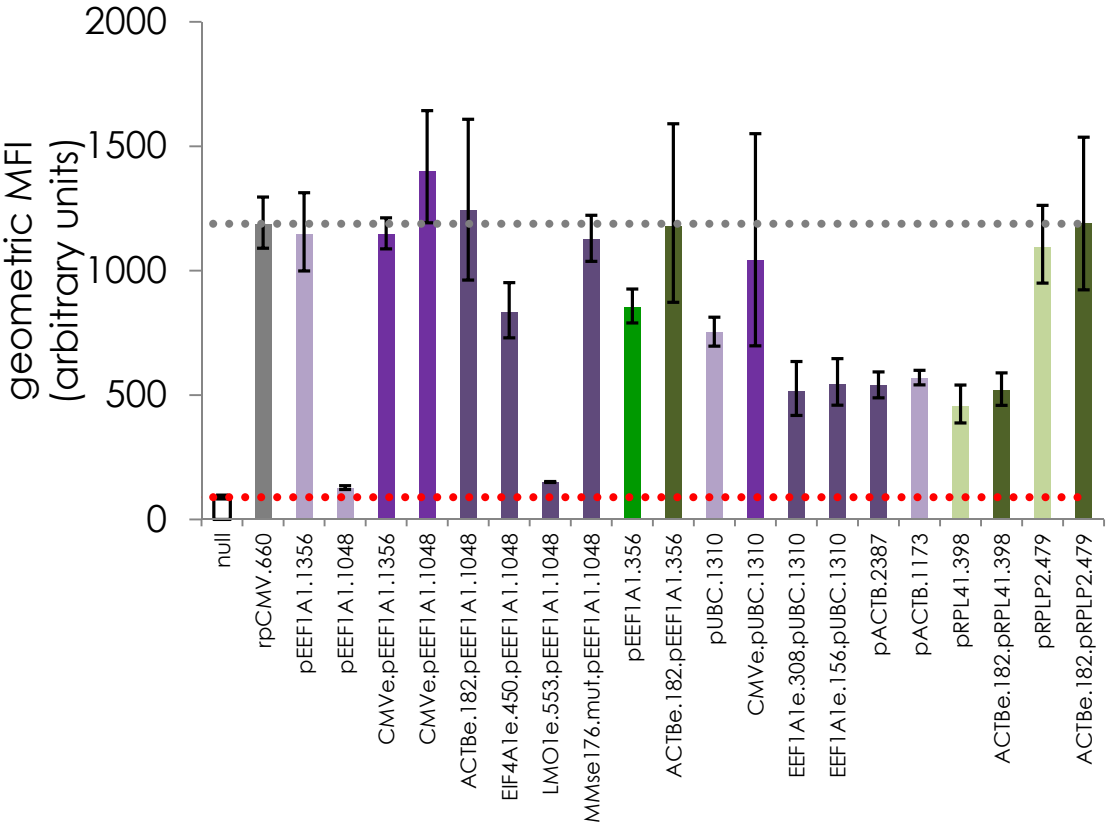
subset of these endogenous promoters (from *EEF1A1*, *UBC*, *RPL41*, and *RPLP2*) with putative enhancer regions derived from *EEF1A1*, *EIF4A1*, or *ACTB* to create hybrid promoters using purely endogenous sequences. Since the pEEF1A1.1048 promoter behaves like a minimal promoter that can be activated with a large dynamic range, we screened two additional endogenous elements that have been characterized as super-enhancer elements from *LMO1*[281] and a mutant variant of enhancer ID 176 (sequence found in **Appendix B**) from Lovén, *et al* that showed a strong signature for enhancer activity in multiple myeloma cells[282]. The 182-bp enhancer region derived from *ACTB* coupled with 2 *EEF1A1* promoter variants (1048-bp "minimal" and 356-bp) resulted in hybrid promoters that exhibited comparable strength to the CMV promoter (**Figures 5-9** and **5-10**) using purely endogenous sequences. The same enhancer region showed minimal influence when coupled with the ribosomal promoters; however, these endogenous promoters were able to drive gene expression at two distinct levels (**Figures 5-9** and **5-10**), suggesting that regulatory regions of ribosomal proteins can also be exploited. Unfortunately, when two variants of the enhancer region from the 356-bp *EEF1A1* promoter (308-bp and 156-bp) were coupled to the 1310-bp *UBC* promoter, the resulting hybrid promoters yielded no additional benefit to gene expression (**Figures 5-9** and **5-10**). This data corroborates previous efforts in constructing hybrid promoters by coupling additional enhancer regions, but the high variation in functionality of these hybrid promoters maintains that this approach is extremely empirical.

Figure 5-9: Transient SEAP productivity (expression) 48h post-transfection in HT1080 driven by endogenous promoters and hybrid promoters with modified enhancers.



Error bars represent 95% CI of SEAP productivity from 3 or more biological replicates. The purple fill color denotes promoters with intron sequences included in the promoter region. The numbers following the promoter name denotes the promoter length. Hybrid promoters with additional enhancer regions are represented by the darker fill color. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

Figure 5-10: Transient hrGFP expression 48h post-transfection in HT1080 driven by other endogenous promoters and hybrid promoters with modified enhancers.



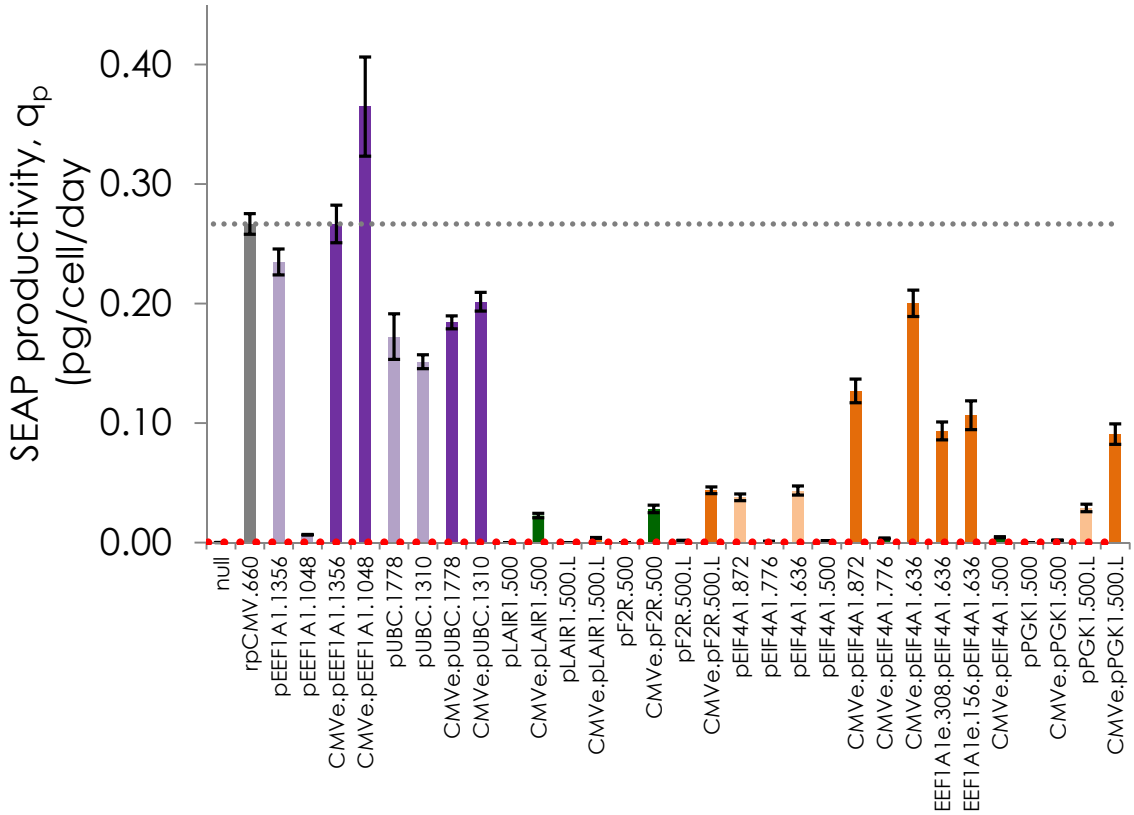
Error bars represent standard deviation of geometric mean fluorescence intensity from 3 or more biological replicates. The purple fill color denotes promoters with intron sequences included in the promoter region. The numbers following the promoter name denotes the promoter length. Hybrid promoters with additional enhancer regions are represented by the darker fill color. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

Based on the expression of our reporters with promoters derived from *EIF4A1* (**Figures 5-7** and **5-8**), critical analysis of the endogenous sequence revealed an extended 5' UTR before the start codon. Thus, we wanted to explore the contribution of the 5' UTR on promoter activity and designed additional hybrid promoters that contained extended 5' UTR regions. Specifically, we assessed permutations of promoters derived from *LAIR1*, *F2R*, *PGK1*, and *EIF4A1* that included the full length of their endogenous 5' UTR up to

the corresponding start codon in these genes based on their relatively poor ability to drive reporter expression (**Figures 5-7** and **5-8**). Given the benefit of coupling the CMV enhancer to these promoters, we wanted to verify that the 5' UTR extension is complementary to the activation by the enhancer region 5' of the transcription start site.

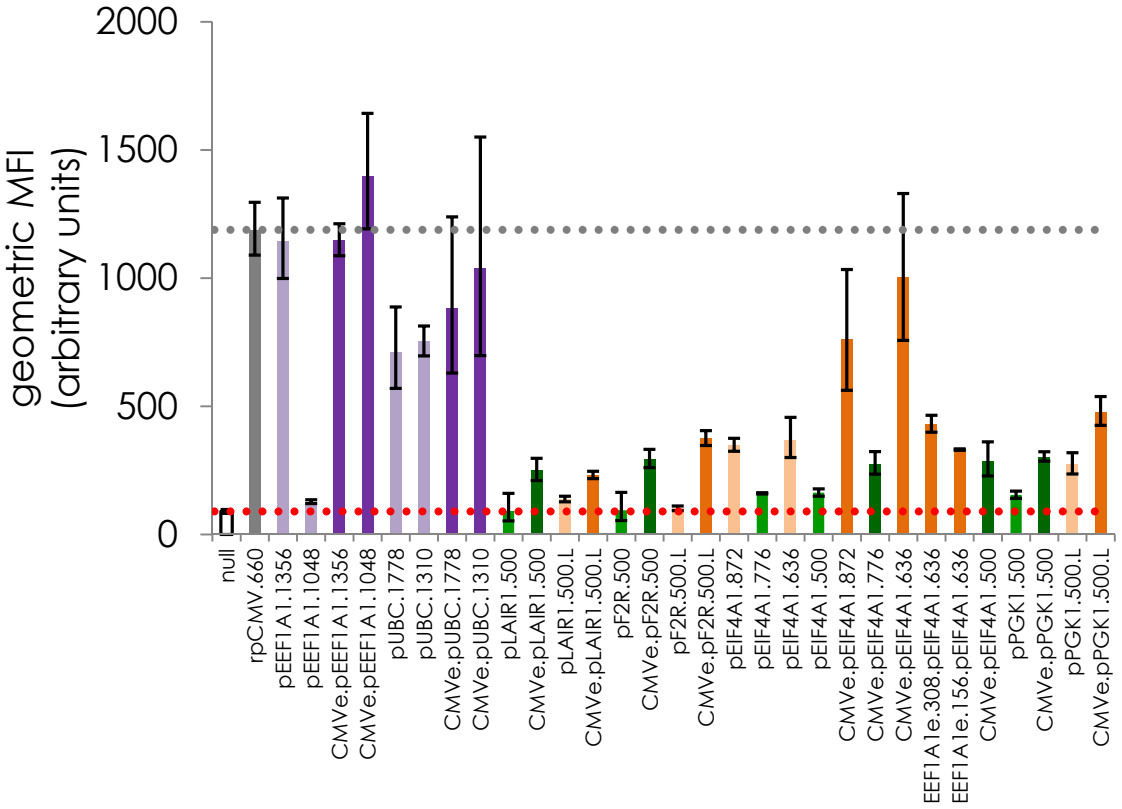
The transient hrGFP and SEAP expression 48h post-transfection from these hybrid promoters with 5' UTR modifications suggest that the promoter strength can be improved by incorporating a suitable 5' UTR, but there is also context dependency with regards to the enhancer region (**Figures 5-11** and **5-12**). For example, the enhancer and 5' UTR can be cooperative (*e.g.* CMV enhancer with 636-bp *EIF4A1* promoter compared with the 500-bp variant), yet the same hybrid promoter with the extended 5' UTR is not influenced by the addition of the 308-bp or 156-bp enhancer variants derived from *EEF1A1*. Increasing the 5' UTR by itself is insufficient to generate hybrid promoters that are comparable to the CMV promoter without incorporating an additional enhancer region. The improved activity from incorporating the native 5' UTR (**Figures 5-11** and **5-12**) suggests that the putative promoters showed minimal activity (**Figures 5-7** and **5-8**) because they were no longer in the appropriate genomic context when employed in a transient expression vector. This is further corroborated by the *EEF1A1*- and *UBC*-derived promoters that include their first intron as a 5' UTR showing significant activity (**Figures 5-7** and **5-8**). Thus, the genomic context is indeed important when hijacking endogenous sequences for gene regulatory function.

Figure 5-11: Transient SEAP productivity (expression) 48h post-transfection in HT1080 driven by hybrid promoters with modified 5' UTR.



Error bars represent 95% CI of SEAP productivity from 3 or more biological replicates. The purple fill color denotes promoters with intron sequences included in the promoter region. The orange fill color denotes promoters with extended 5' UTR. Numbers following the promoter name denotes the promoter length. CMVe denotes promoter sequences with the hCMV-IE enhancer coupled, represented by the darker fill color. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

Figure 5-12: Transient hrGFP expression 48h post-transfection in HT1080 driven by hybrid promoters with modified 5' UTR.



Error bars represent standard deviation of geometric mean fluorescence intensity from 3 or more biological replicates. The purple fill color denotes promoters with intron sequences included in the promoter region. The orange fill color denotes promoters with extended 5' UTR. Numbers following the promoter name denotes the promoter length. CMVe denotes promoter sequences with the hCMV-IE enhancer coupled, represented by the darker fill color. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

5.3.3. Promoter redesign with introns

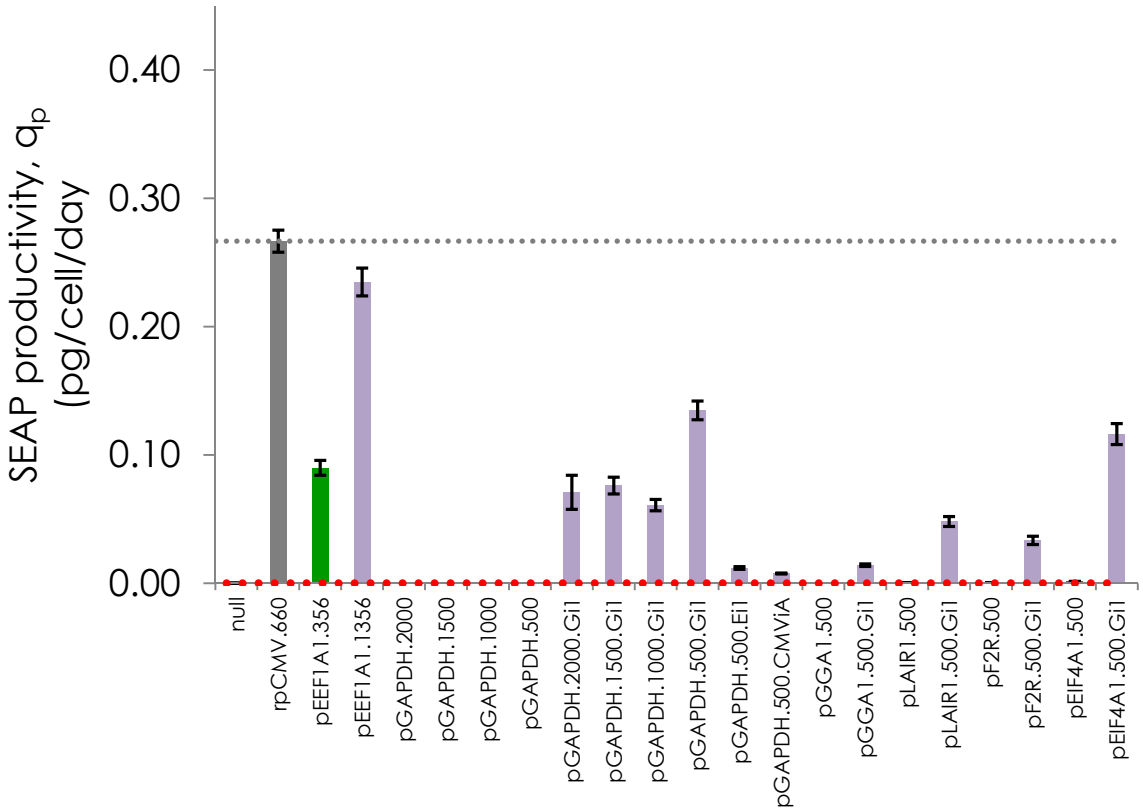
Along the same vein as incorporating a native 5' UTR to modulate promoter activity, we showed that the endogenous promoters with introns (*EEF1A1*- and *UBC*-derived) and their hybrid derivatives are the predominant ones that exhibited comparable functionality to the strong CMV promoter (**Figures 5-7** to **5-10**). Therefore, we suspected

that the presence of an intron was responsible for the strong promoter activity since using the *EEF1A1* enhancer regions were insufficient to elicit this effect (**Figures 5-9** to **5-12**). Given the significant DNA footprint that introns typically occupy, it is expected that these sequences would contain additional transcription factor binding sites that can activate or repress gene expression[283]. It was previously shown that the intron A and its variants from the CMV promoter can drastically affect its strength[273, 284]. In the case of the *GAPDH* promoter variants, which we expected should have strong promoter activities, sequence analysis revealed that the endogenous gene locus contained a 240-bp intron prior to the start codon found in the second exon. Therefore, we created additional *GAPDH* promoter variants with the same 5' enhancer regions and included this first intron and flanking exons. We also rationally combined this *GAPDH* endogenous intron and flanking exons with some of the putative promoters derived from *GGA1*, *LAIR1*, and *F2R* that showed negligible activities originally (**Figures 5-7** and **5-8**) and subsequently detectable activities when their 5' UTR were extended (**Figures 5-11** and **5-12**). Furthermore, we created an additional *EIF4A1* promoter variant with the same *GAPDH* intron 1 and flanking exons.

When the previous *GAPDH* promoter variants were coupled to its cognate intron 1, we observed increases in gene expression as expected (**Figures 5-13** and **5-14**). However, pairing the *GAPDH* intron 1 with other endogenous promoters resulted only in a modest increase in promoter strength (**Figures 5-13** and **5-14**), suggesting that the improvement observed with the *GAPDH* promoters with the intron is unlikely to be dictated by additional activating transcription factor binding sites within the intron and flanking exons. It is possible that these other promoters contain similar binding sites to those found in the *GAPDH* intron 1 and flanking exons, resulting in insignificant contributions. Due to the improvements observed when the *GAPDH* promoters were coupled with its cognate intron, we also evaluated two additional variants by combining the 500-bp *GAPDH* promoter with

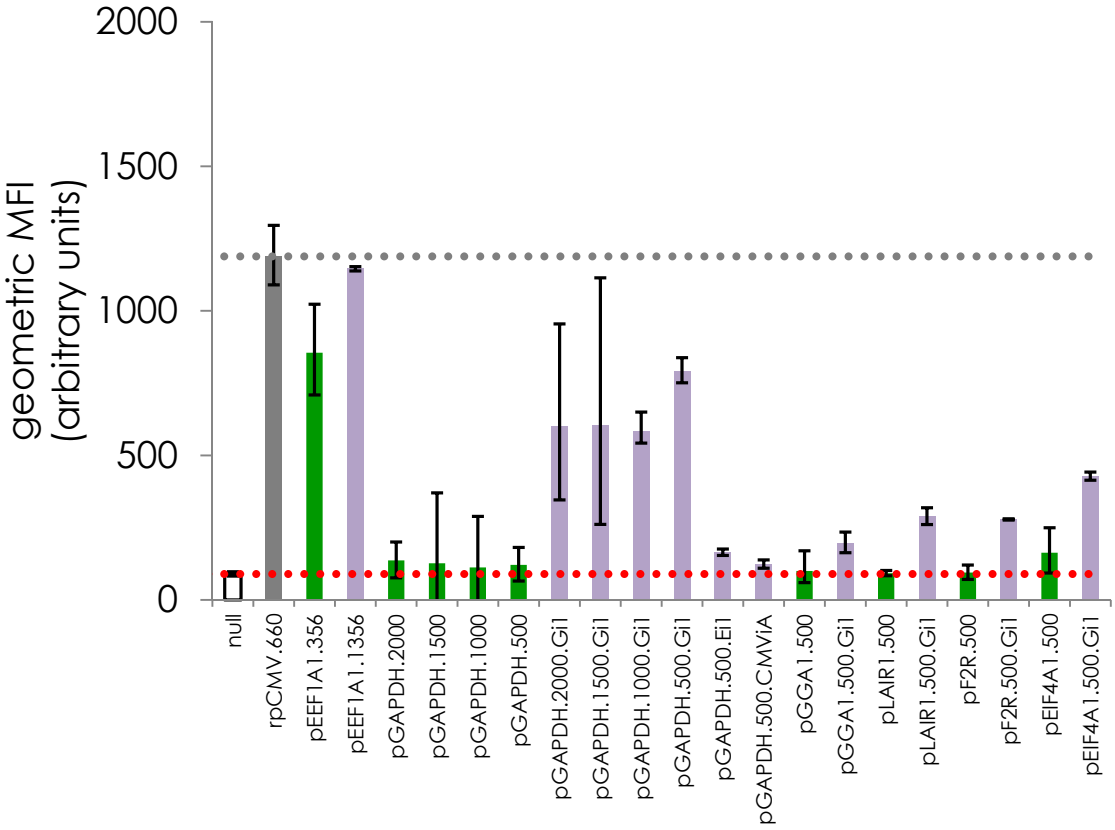
the *EEF1A1* intron 1 with its flanking exons or the hCMV-IE intron A with its flanking exons. Interestingly, despite both introns demonstrating an increase in promoter activity when paired with their cognate promoters, they negatively impacted the *GAPDH* promoter (**Figures 5-13** and **5-14**).

Figure 5-13: Transient SEAP productivity (expression) 48h post-transfection in HT1080 driven by hybrid promoters with modified introns.



Error bars represent 95% CI of SEAP productivity from 3 or more biological replicates. The purple fill color denotes promoters with intron sequences included in the promoter region. The numbers following the promoter name denotes the promoter length. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

Figure 5-14: Transient hrGFP expression 48h post-transfection in HT1080 driven by hybrid promoters with modified introns.

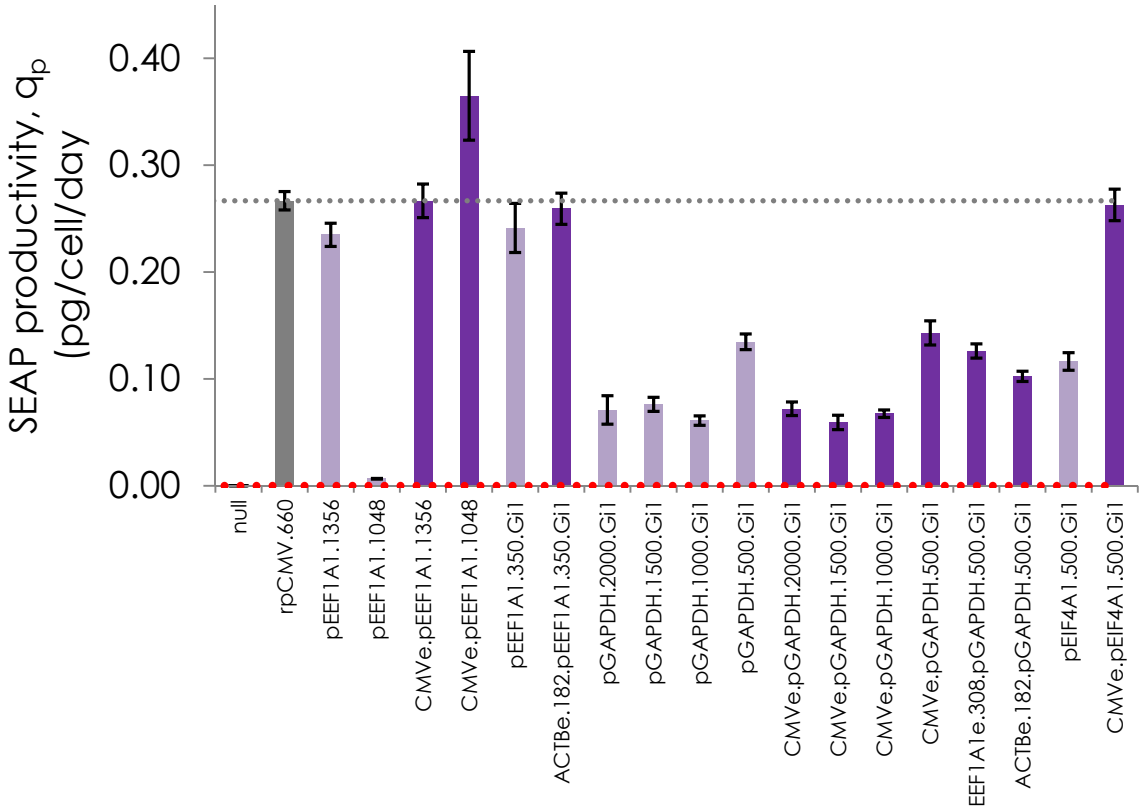


Error bars represent standard deviation of geometric mean fluorescence intensity from 3 or more biological replicates. The purple fill color denotes promoters with intron sequences included in the promoter region. The numbers following the promoter name denotes the promoter length. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

Lastly, we investigated a few additional hybrid promoter variants that utilized the *GAPDH* intron 1 with other enhancer regions (based on data found in **Figure 5-13** and **5-14**) to drive gene expression as quantified by the hrGFP and SEAP reporters (**Figures 5-15** and **5-16**). Interestingly, the combination of these two major regulatory regions showed an inconsistent interplay: the addition of the enhancer regions did not always increase gene expression (*e.g. GAPDH* variants with CMV enhancer). In fact, the 500-bp *GAPDH* hybrid

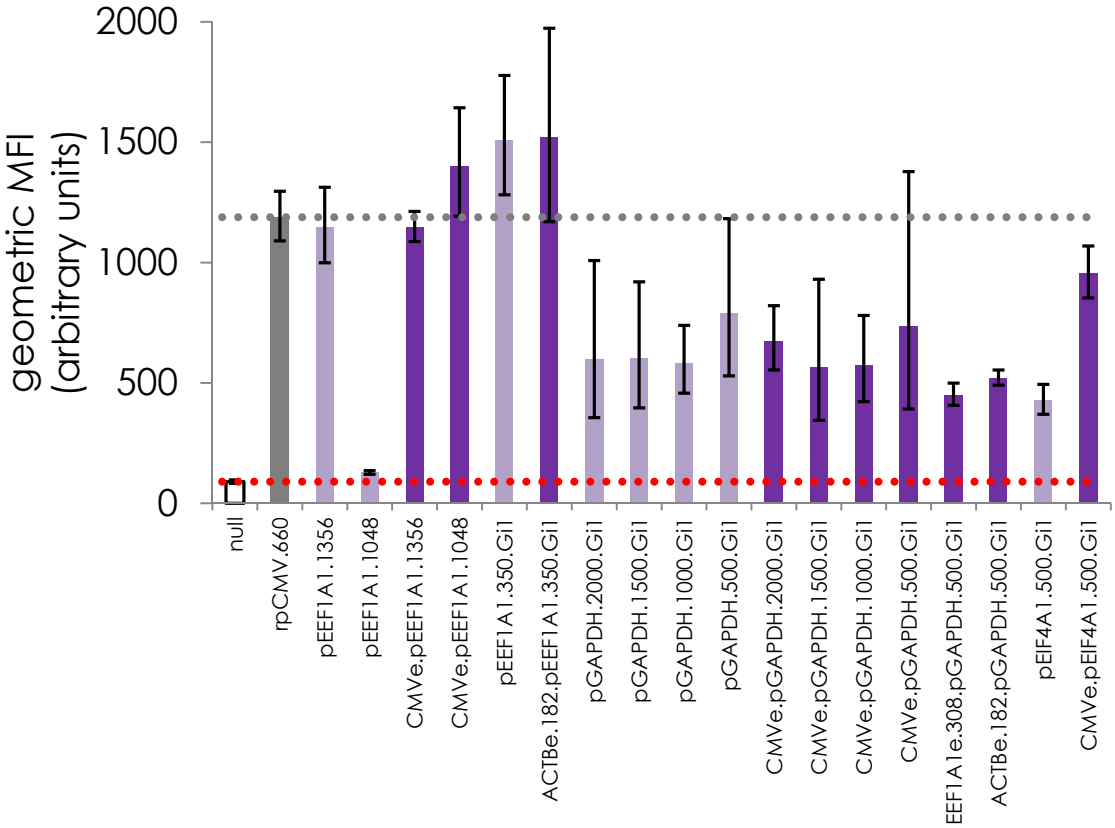
promoter with *GAPDH* intron 1 was not affected by the 3 distinct enhancers coupled to this particular promoter (**Figures 5-15** and **5-16**). However, a hybrid promoter based on the 500-bp *EIF4A1* promoter with *GAPDH* intron 1 showed a marked increase when it was coupled to the CMV enhancer (**Figures 5-15** and **5-16**). Based on these results, it is difficult to identify which element is dominant when they are used in conjunction. Instead, permutations of these elements derived from endogenous sequences must be evaluated empirically at this time as clear design rules for hybrid promoters cannot be elucidated from this data.

Figure 5-15: Transient SEAP productivity (expression) 48h post-transfection in HT1080 driven by hybrid promoters with modified enhancers and introns.



Error bars represent 95% CI of SEAP productivity from 3 or more biological replicates. The purple fill color denotes promoters with intron sequences included in the promoter region. The numbers following the promoter name denotes the promoter length. Hybrid promoters with additional enhancer regions are represented by the darker fill color. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

Figure 5-16: Transient hrGFP expression 48h post-transfection in HT1080 driven by hybrid promoters with modified enhancers and introns.

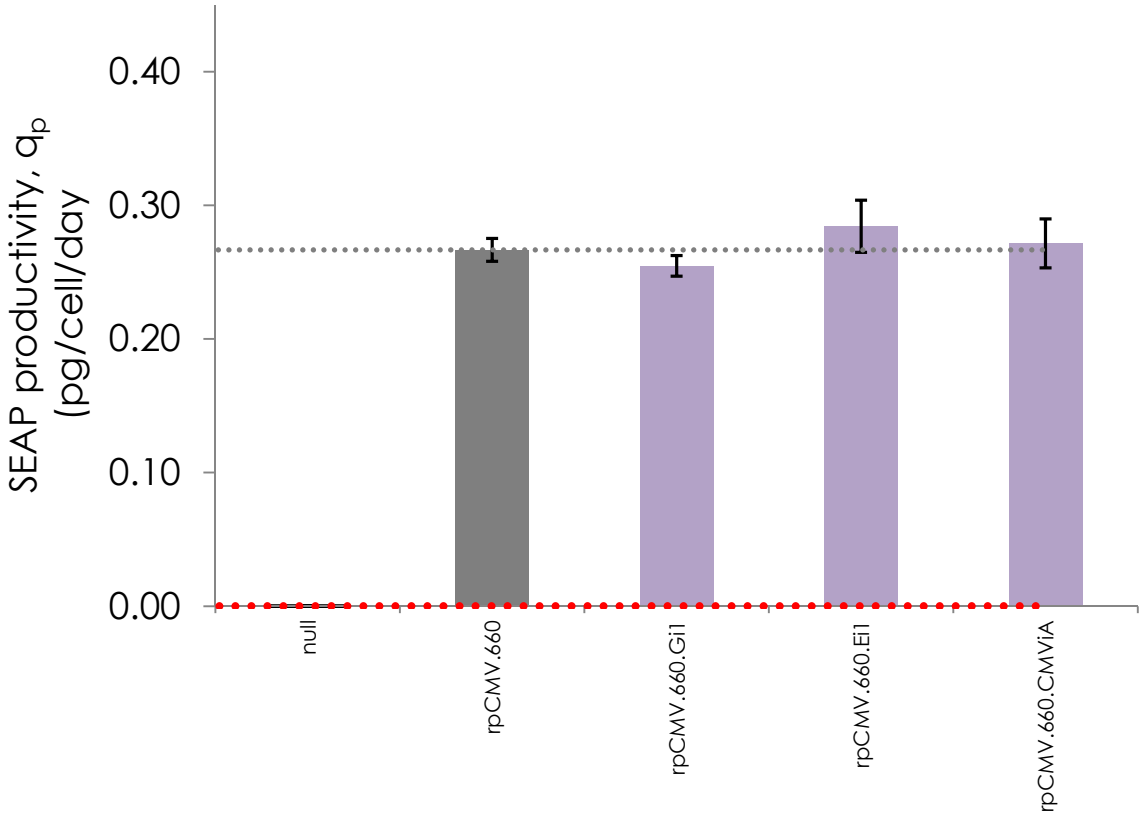


Error bars represent standard deviation of geometric mean fluorescence intensity from 3 or more biological replicates. The purple fill color denotes promoters with intron sequences included in the promoter region. The numbers following the promoter name denotes the promoter length. Hybrid promoters with additional enhancer regions are represented by the darker fill color. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

Based on these inconsistent benefits to gene expression, we wanted to verify the impact of using an intron with the CMV promoter as previously reported[273, 284]. We designed hybrid promoter variants of the CMV promoter that contained the *EEF1A1* or *GAPDH* intron 1 and their flanking exons in addition to the endogenous variant with the hCMV-IE intron A. We measured transient reporter expression of SEAP and hrGFP driven by this particular set of CMV hybrid promoters and found that the introns had no impact

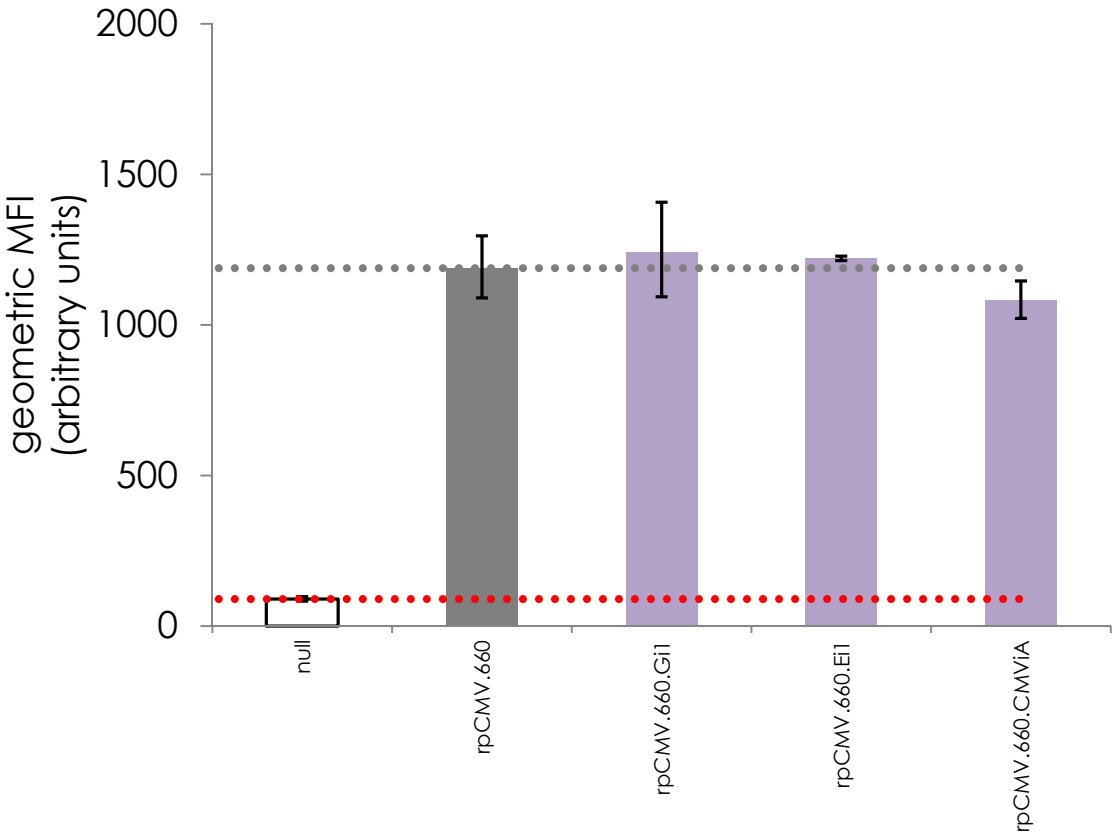
on promoter activity (**Figure 5-17** and **5-18**). Thus, it is unlikely that any observed benefits are due to these regions hosting additional transcription factor binding sites that activate transcription.

Figure 5-17: Transient SEAP productivity (expression) 48h post-transfection in HT1080 driven by CMV hybrid promoters with intron variants.



Error bars represent 95% CI of SEAP productivity from 3 or more biological replicates. The numbers following the promoter name denotes the promoter length. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

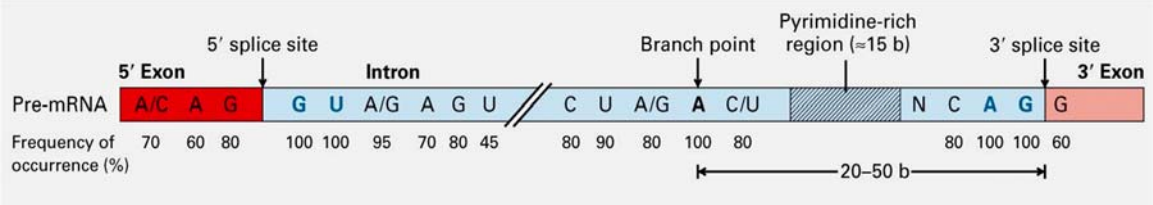
Figure 5-18: Transient hrGFP expression 48h post-transfection in HT1080 driven by CMV hybrid promoters with intron variants.



Error bars represent standard deviation of geometric mean fluorescence intensity from 3 or more biological replicates. The numbers following the promoter name denotes the promoter length. Gray dotted line represents the expression level driven by the strong CMV promoter and the red dotted line represents background expression/autofluorescence from a promoter-less construct.

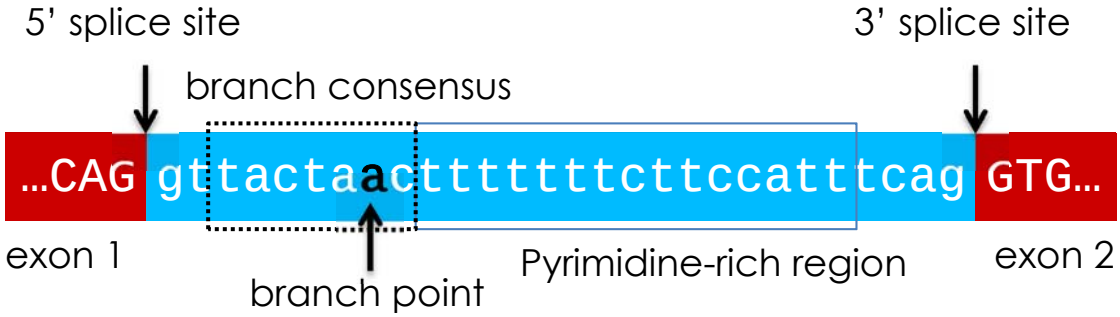
Another suggested mechanism is that the splicing machinery can stimulate transcription machinery, leading to successful gene expression[285]. To further evaluate this finding, we designed a synthetic minimal intron (iS1) based on conserved sequence elements (**Figure 5-19**) and the *EEF1A1* pyrimidine-rich region and flanking exons (**Figure 5-20**). This intron is devoid of any putative transcription factor binding sites based on the sequences from JASPAR database[233, 286], allowing us to interrogate the impact of intron processing on gene expression directly.

Figure 5-19: Typical intron structure with approximate consensus sequence based on nucleotide frequencies.



Reproduced from Stevens lecture, BIO 395J Fall 2013, The University of Texas at Austin.

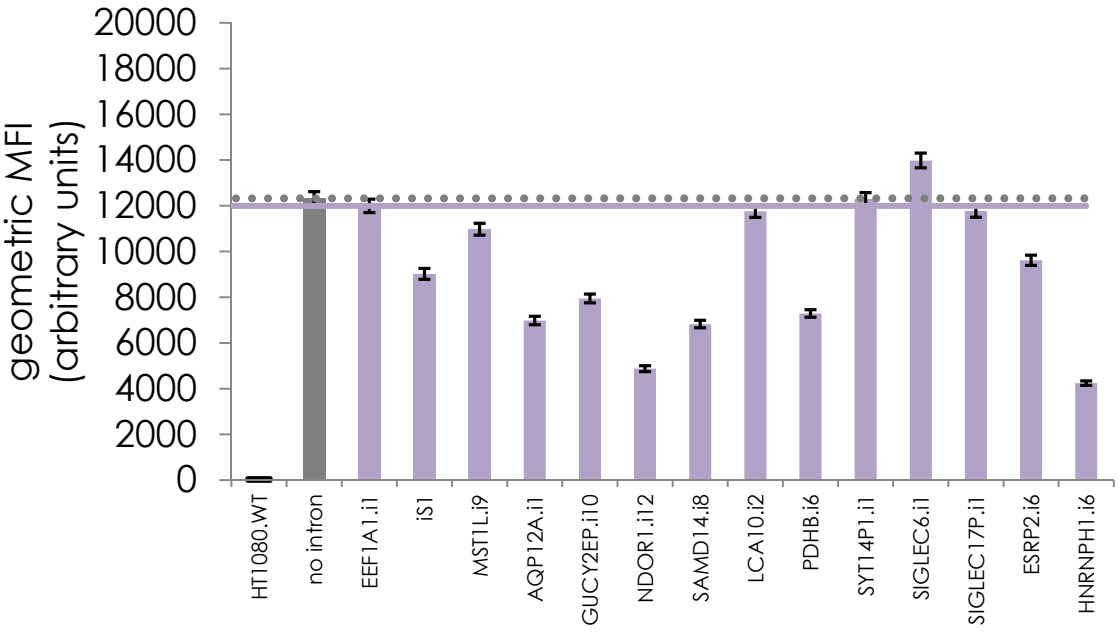
Figure 5-20: Sequence of synthetic intron iS1 based on conserved sequences and a 19-bp pyrimidine-rich region from *EEF1A1* flanked by the *EEF1A1* exons 1 and 2.



With such a minimal sequence, it is unclear whether the splicing machinery would be able to properly process this mRNA given the size of the spliceosome[287]. Therefore, we screened additional "ultra-short" introns found in the human genome to exploit the minimal sequence space and minimize the potential of these sequences to harbor transcription factor binding sites. The splicing of the 56-bp intron from *HNRNPH1* and the 43-bp intron from *ESRP2* were confirmed[287], and we also built test constructs based on the 10 shortest introns that were identified by Piovesan, *et al*[288]. The processing of the 49-bp intron of *NDOR1* found in this data set[288] was also experimentally verified previously[287]. These 12 minimal endogenous introns were also flanked by the *EEF1A1* exons 1 and 2 to maintain genomic context consistency and consistency in the final 5' UTR of the reporter mature mRNA.

We compared the ability of the CMV promoter with and without each of these minimal introns to drive gene expression as quantified by our hrGFP reporter. As a positive control, we included the *EEF1A1* intron 1 flanked by its endogenous exons. The control promoter without an intron had the native exons replaced by the *EEF1A1* exons 1 and 2 for context consistency. Interestingly, these minimal introns exhibited significantly different effect on gene expression (**Figure 5-21**). These hybrid promoters with minimal introns cover almost a 3-fold dynamic range, further highlighting the need for empirical screening of endogenous sequences to identify desirable elements for a particular application.

Figure 5-21: Transient hrGFP expression 48h post-transfection in HT1080 driven by CMV hybrid promoters with intron variants in a single replicate.



Error bars represent 95% CI of geometric mean fluorescence intensity. The numbers following the promoter name denotes the promoter length. Gray dotted line represents the expression level driven by the strong CMV promoter and the purple solid line represents the expression from the CMV promoter coupled to the native intron 1 from *EEF1A1*.

5.4. CONCLUSIONS

Rational and library-based promoter engineering approaches have been demonstrated to yield strong components[107, 110, 111, 215, 247, 248, 254] and characterization of endogenous genes led to the exploitation of these regulatory sequences to drive heterologous gene expression, especially in mammalian cells[222, 224, 258, 261, 269, 289]. Through integrating an additional layer of classification based on endogenous expression levels, we can narrow down the desired sequence space screened in order to generate strong hybrid promoters. Overall, by evaluating several of the elements that regulate transcription at the promoter level, we constructed hybrid promoters solely relying on endogenous sequences that drove a wide range of expression levels, relieving the dependence on viral-derived elements. Most importantly, to address the goal of replacing strong viral-derived elements, we were able to construct a hybrid promoter (pEEF1A1.350.Gi1) that exhibited comparable activity (1.1-fold) to the strong CMV promoter. Although this hybrid promoter required slightly larger sequence footprint relative to the CMV promoter (770-bp vs. 660-bp), it required substantially less sequence footprint than the strongest endogenous promoter evaluated from *EEF1A1* (770-bp vs. 1356-bp).

The use of multiple core promoters in mammalian cells did not result in the same level of improvement in gene expression previously observed in *E. coli*[264], but we were able to increase the percentage of transfected cells that expressed our transgene. This suggested that the multiple core promoter elements were in fact functional based on the increase in percentage of transgene expressing cells. Furthermore, this repetitive core promoter architecture can be extrapolated for other applications such as creating a hybrid core promoter that would be functional across multiple cell types by combining multiple core promoters unique to these cell types.

In evaluating endogenous promoters and designing hybrid promoters in the more traditional fashion of exploring enhancer regions, we observed stark differences between many of our designs. Many of the endogenous promoters could not drive gene expression, suggesting a high dependence on genomic context for their activity. It is unclear whether this context is a result of chromatin organization to govern expression[290] or if a more comprehensive inclusion of the local regulatory sequence is required for promoter activity. However, incorporating additional genomic context through an extended 5' UTR yielded increased promoter activity, suggesting that this is a viable approach to generate hybrid promoters for mammalian cell types. Also, endogenous enhancers did not exhibit a consistent activation when paired with other endogenous promoters, indicating that more fundamental design rules are required. This was further exemplified when these enhancer regions were coupled to promoters with extended 5' UTR or introns.

With regards to the minimal introns, the variable gene expression observed (**Figure 5-21**) cannot be solely attributed to whether the intron processing was experimentally confirmed since the hybrid promoters with introns derived from *NDOR1*, *ESRP2*, and *HNRNPH1* are distinctly different. Nonetheless, the resulting hybrid promoters were only comparable to the promoter without an intron, confirming the lack of benefit observed when the CMV promoter was coupled with its cognate intron A in our data (**Figures 5-17** and **5-18**). This data suggests that the intron could be a superfluous element for the CMV promoter to drive gene expression based on the data with an underlying assumption. The impact on transcription could be masked if the translation machinery is saturated, thus any increase in mRNA abundance would not increase protein levels. However, there were instances where protein levels were significantly lower than our control promoter without an intron, but it is unclear from protein expression data alone whether this is due to lower mRNA abundance or ineffective processing of mRNA due to these introns.

Ultimately, through our investigation of the putative promoter structure in the mammalian genome, we revealed that the core promoter, enhancer, and 5' UTR can drastically influence gene expression in a combinatorial, yet currently unpredictable, manner. The inconsistent interplay between these elements recapitulated the need for empirical and systematic evaluation of each element. Through these efforts, we gained some insight into the design rules and impact of each element on gene expression that collectively could lead to *de novo* rational promoter design for target expression levels.

Chapter 6: Prolonging the message – engineering terminators for high transgene expression

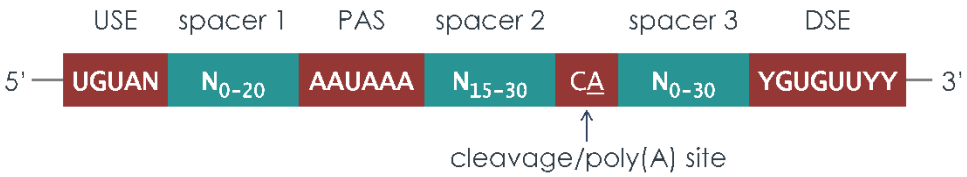
6.1. CHAPTER SUMMARY

Many of the previous efforts to boost transgene expression in mammalian hosts focus on the promoter element, as described in Chapters 4 and 5. Understandably, the promoter is the element responsible for generating mRNA and therefore the primary driver for gene expression. However, the terminator (3' UTR) of the transgene can also modulate gene expression by controlling the stability of the transgene mRNA and exerting influence on post-transcriptional processing such as intron splicing[291-293]. Akin to the work described in Chapters 4 and 5 to replace viral promoters with endogenous and synthetic components, we generated rational endogenous and synthetic terminators that are comparable in terms of tuning gene expression and require less sequence footprint relative to a standard viral counterpart. Ultimately, this novel set of terminators expanded our genetic toolkit for engineering mammalian host cells.

6.2. INTRODUCTION

The terminator of any coding sequence is often overlooked with minimal emphasis during transgene design, especially in mammalian cells, as exemplified by the common usage of viral derived terminators[294]. In order to generate rational designs of terminators, we require an understanding of the essential elements that comprise a functional terminator. Significant efforts have characterized various viral and endogenous terminators[295-301], resulting in a suggested generic structure for this particular element[302]. This generic structure is comprised of an upstream sequence element (USE), the highly conserved hexameric polyadenylation signal (PAS), a cleavage/polyadenylation site, and the downstream sequence element (DSE) (**Figure 6-1**).

Figure 6-1: Generic structure of mammalian terminators as described by Proudfoot[302].



USE represents the upstream sequence element based on a consensus sequence[303], PAS represents the highly conserved hexameric polyadenylation signal[300], and DSE represents the downstream sequence element[304].

These key regions in the terminator have similar structure to yeast terminators that were previously exploited to design synthetic terminators for *S. cerevisiae* and showed functionality in another yeast *Y. lipolytica*[305], suggesting that synthetic designs based these conserved regions of the terminator are broadly functional. Along the same vein, combining the endogenous element from the human gastrin gene with the viral SV40 terminator improved reporter gene expression at the transcript level in 3 model mammalian cell lines, affirming some modularity to these elements[306]. While it was unclear from this work what mechanism was directly responsible for the increased mRNA levels, this work suggested that terminators can be rationally designed to influence gene expression in mammalian cells, much like the previous efforts in yeast[305]. More importantly, transcription termination can improve gene expression regardless of the promoter in yeast[305, 307] and mammalian cells[292], indicating that the terminator design can be orthogonal to the promoter engineering efforts described in Chapters 4 and 5.

Previous studies characterized some of these key elements of the terminator and identified their impact on post-transcriptional regulation[308]. The most highly conserved element is the hexameric PAS, with minimal deviation from the AAUAAA consensus[300, 309]. The USE was reported to contain sequences that interact with splicing factors, and these interactions promote 3' end formation[303]. Likewise, the consensus sequence for the DSE was essential for 3' end formation[301, 304, 310, 311]. Despite the generic structure of

mammalian terminators, it was suggested that some of these elements are dispensable[312]. In the absence of the hexameric PAS, other USEs can direct the recognition of the appropriate polyadenyation site and poly(A) addition[313, 314]. In fact, a synthetic sequence comprised of only the PAS and DSE spaced 22- to 23-bp apart was sufficient to terminate transcription[315]. Therefore, we adopted a similar approach in exploiting the key elements found in the conserved terminator structure for rational designs[305] to create and subsequently evaluate a set of novel endogenous and synthetic mammalian terminators. The expanded availability of characterized mammalian terminators facilitates the replacement of viral-derived sequences and offer additional tools to fine-tune gene expression in these cells.

6.3. RESULTS AND DISCUSSION

6.3.1. Rational design and evaluation of endogenous and synthetic terminators

It was previously reported that many housekeeping genes have mRNA with relatively long half-lives[316], suggesting that the terminators of these genes have sequences in their 3' UTR that confer this attribute. In addition, some of these genes (*e.g. EEF1A1*, *GAPDH*, and *ACTB*) are also among the most highly expressed in the genome based on our microarray expression data analyzed in Chapter 4. Therefore, we first explored variants of the 3' UTR from these three genes as terminators and evaluated their impact on the expression of the fluorescent reporter hrGFP. Terminator sequences from *EEF1A1*, *GAPDH*, and *ACTB* were selected based on approximations of the key terminator elements previously described, most notably the USEs and DSEs (**Table 6-1**). These sequences are not fully annotated in the genome, thus variants were selected and evaluated similar to previous characterization work with promoter identification and verification. We also evaluated a terminator derived from the 3' UTR of the non-coding *RPL41* gene as an

alternative to key housekeeping coding genes (**Table 6-1**). As a control, we also conducted a truncation analysis of the commonly used SV40 terminator to estimate the effects of the USEs and DSEs and terminator impact on hrGFP expression (**Table 6-1**).

Table 6-1: Table of native/endogenous terminator sequences with putative spacer regions in lower case.

terminator	sequence 5' > 3'	length	Notes
f.SV40pA	CAGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAAccattataagctgcAATAAAcaagttaacaacAAcaattgcattcatttTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTA	222	SV40 late terminator
m.SV40pA.1	TGTAAccattataagctgcAATAAAcaagttaacaacAAcaattgcattcatttTATGTTTC	62	"minimal" SV40 late terminator, minimal USE, minimal DSE
m.SV40pA.2	TGTAAccattataagctgcAATAAAcaagttaacaacAAcaattgcattcatttTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTT	92	"minimal" SV40 late terminator, minimal USE, short DSE
m.SV40pA.3	CAGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAAccattataagctgcAATAAAcaagttaacaacAAcaattgcattcatttTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTT	192	"minimal" SV40 late terminator, full USE, short DSE
m.SV40pA.4	TGTGATGCTATTGCTTTATTTGTAAccattataagctgcAATAAAcaagttaacaacAAcaattgcattcatttTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTA	142	"minimal" SV40 late terminator, short USE, full DSE
T.EEF1A1.1	TGGTATtcattacaaacttgctcactacAATAAAtgaattttaagctttAAgatgaagtggcaTTTCTTTT	71	EEF1A1 terminator, short USE, short DSE
T.EEF1A1.2	TGGTATtcattacaaacttgctcactacAATAAAtgaattttaagctttAAgatgaagtggcaTTTCTTTTAACAGTTACTATGTTGGAATTGGTTACAAATTTTGGAGTGGATTTCAAAAGTGAGAGCTAACTTCAGTTGATTTCAAGGTAGTGCTTGGCTTTTTTTGTTTA	173	EEF1A1 terminator, short USE, long DSE
T.EEF1A1.3	TGTGAAACCCAGTGTCTTAGACAACTGTGGCTTGAGCACCACCTGCTGGTATtcattacaaacttgctcactacAATAAAtgaattttaagctttAAgatgaagtggcaTTTCTTTT	117	EEF1A1 terminator, long USE, short DSE
T.EEF1A1.4	TGTGAAACCCAGTGTCTTAGACAACTGTGGCTTGAGCACCACCTGCTGGTATtcattacaaacttgctcactacAATAAAtgaattttaagctttAAgatgaagtggcaTTTCTTTTAACAGTTACTATGTTGGAATTGGTTACAAATTTTGGAGTGGATTTCAAAAGTGAGAGCTAACTTCAGTTGATTTCAAGGTAGTGCTTGGCTTTTTTTGTTTA	219	EEF1A1 terminator, long USE, long DSE
T.GAPDH.1	TGTCATGTACcatcAATAAAgtaccctgtgctcaaccagtTActtgtcctgtcttattctagGGTCTGGGGCAGAGGGGAGGGAAGCTGGGCTTGTGTC	99	GAPDH terminator, short USE
T.GAPDH.2	TGTAGACCCCTTGAAGAGGGGAGGGGCCTAGGGAGCCGCACCTTGTCATGTACcatcAATAAAgtaccctgtgctcaaccagtTActtgtcctgtcttattctagGGTCTGGGGCAGAGGGGAGGGAAGCTGGGCTTGTGTC	142	GAPDH terminator, long USE

Table 6-1, continued:

terminator	sequence 5' > 3'	length	Notes
T.ACTB.f	GCGGACTATGACTTAGTTGCGTTACACCCTTTCTTGACAAAACCTAACTTGCGCAGA AAACAAGATGAGATTGGCATGGCTTTATTTGTTTTTTTTGTTTTGTTTTGGTTTTTTTT TTTTTTTTGGCTTGACTCAGGATTTAAAAACTGGAACGGTGAAGGTGACAGCAGTCG GTTGGAGCGAGCATCCCCCAAAGTTCACAATGTGGCCGAGGACTTTGATTGCACAT TGTTGTTTTTTTAATAGTCATTCCAAATATGAGATGCGTTGTTACAGGAAGTCCCTTG CCATCCTAAAAGCCACCCCACTTCTCTCTAAGGAGAATGGCCCAGTCCTCTCCCAAG TCCACACAGGGGAGGTGATAGCATTGCTTTCGTGTAAATTATGTAATGCAAAATTTT TTTAATCTTCGCCTTAATACTTTTTTATTTTGTTTTATTTTGAATGATGAGCCTTCGTG CCCCCCCTTCCCCCTTTTTTGTCCCCCAACTTGAGATGTATGAAGGCTTTTGGTCTCC CTGGGAGTGGGTGGAGGCAGCCAGGGCTTACCTGTACactgacttgagaccagttgAATAAAag tgcacaccttaaaaatGAggccaagtgtgacTTTGTGGTGTGGCTGGGTTGGGGGCAGCAGAGGGTG AACCCTGCAGGAGGGTGAACCCTGCAAAAGGGTGGGGCAGTGGGGGCCAACTTGT CCTTACCCAGAGTGCAGGTGTGTGGAGATCCCTCCTGCCTTGACATTGAGCAGCCTT AGAGGGTGGGGGAGGCTCAGGGGTCAGGTCTCTGTTCCTGCTTATTGGGGAGTTCC TGGCCTGGCCCTTCTATGTCTCCCCAGGTACCCCAGTTTTTCTGGGTTCACCCAGAGT GCAGATGCTTGAGGAGGTGGGAAGGGACTATTTGGGGGTGTCTGGCTCAGGTGCCA TGCCTCACTGGGGCTGGTTGGCACCTGCATTTCCTGGGAGTGGGGCTGTCTCAGGGT AGCTGGGCACGGTGTTCCCTTGAGTGGGGGTGTAGTGGGTGTTCCTAGCTGCCACGC CTTTGCCTTCACCTATGGGA	1065	ACTB reference sequence (same for both mRNA variants), full USE sequence prior to PAS, long DSE
T.ACTB.1	GCGGACTATGACTTAGTTGCGTTACACCCTTTCTTGACAAAACCTAACTTGCGCAGA AAACAAGATGAGATTGGCATGGCTTTATTTGTTTTTTTTGTTTTGTTTTGGTTTTTTTT TTTTTTTTGGCTTGACTCAGGATTTAAAAACTGGAACGGTGAAGGTGACAGCAGTCG GTTGGAGCGAGCATCCCCCAAAGTTCACAATGTGGCCGAGGACTTTGATTGCACAT TGTTGTTTTTTTAATAGTCATTCCAAATATGAGATGCGTTGTTACAGGAAGTCCCTTG CCATCCTAAAAGCCACCCCACTTCTCTCTAAGGAGAATGGCCCAGTCCTCTCCCAAG TCCACACAGGGGAGGTGATAGCATTGCTTTCGTGTAAATTATGTAATGCAAAATTTT TTTAATCTTCGCCTTAATACTTTTTTATTTTGTTTTATTTTGAATGATGAGCCTTCGTG CCCCCCCTTCCCCCTTTTTTGTCCCCCAACTTGAGATGTATGAAGGCTTTTGGTCTCC CTGGGAGTGGGTGGAGGCAGCCAGGGCTTACCTGTACactgacttgagaccagttgAATAAAag tgcacaccttaaaaatGAggccaagtgtgacTTTGTGGTGTGGCTGGGTTGGGGGCAGCAGAGGGTG AACCCTGCAGGAGGGTGAACCCTGCAAAAGGGTGGGGCAGTGGGGGCCAAC	700	ACTB reference sequence (same for both mRNA variants), full USE sequence prior to PAS, 100-bp DSE after poly(A) site

Table 6-1, continued:

terminator	sequence 5' > 3'	length	Notes
T.ACTB.2	GCGGACTATGACTTAGTTGCGTTACACCCTTTCTTGACAAAACCTAACTTGCGCAGA AAACAAGATGAGATTGGCATGGCTTTATTTGTTTTTTTTGTTTTGTTTTGGTTTTTTTT TTTTTTTTGGCTTGACTCAGGATTTAAAAACTGGAACGGTGAAGGTGACAGCAGTCG GTTGGAGCGAGCATCCCCCAAAGTTCACAATGTGGCCGAGGACTTTGATTGCACAT TGTTGTTTTTTTAATAGTCATTCCAAATATGAGATGCGTTGTTACAGGAAGTCCCTTG CCATCCTAAAAGCCACCCCACTTCTCTCTAAGGAGAATGGCCCAGTCCTCTCCCAAG TCCACACAGGGGAGGTGATAGCATTGCTTTCGTGTAAATTATGTAATGCAAAATTTT TTTAATCTTCGCCTTAATACTTTTTTATTTTGTTTTATTTTGAATGATGAGCCTTCGTG CCCCCCCTTCCCCCTTTTTTGTCCCCCAACTTGAGATGTATGAAGGCTTTTGGTCTCC CTGGGAGTGGGTGGAGGCAGCCAGGGCTTACCTGTACactgacttgagaccagttgAATAAAag tgcacaccttaaaaatGAggccaagtgtgacTTTGTGGTGTGGCTGGGTTGGGGG	637	ACTB reference sequence (same for both mRNA variants), full USE sequence prior to PAS, 37-bp DSE after poly(A) site
T.ACTB.3	TTGCTTTCGTGTAAATTATGTAATGCAAAATTTTTTTAATCTTCGCCTTAATACTTTT TTATTTTGTTTTATTTTGAATGATGAGCCTTCGTGCCCCCCCTTCCCCCTTTTTTGTCC CCCAACTTGAGATGTATGAAGGCTTTTGGTCTCCCTGGGAGTGGGTGGAGGCAGCC AGGGCTTACCTGTACactgacttgagaccagttgAATAAAagtgcacaccttaaaaatGAggccaagtgtgacTTT GTGGTGTGGCTGGGTTGGGGGCAGCAGAGGGTGAACCCTGCAGGAGGGTGAACCCT GCAAAAGGGTGGGGCAGTGGGGGCCAAC	333	ACTB reference sequence, short USE sequence prior to PAS for both mRNA variants, 100-bp after poly(A) site
T.ACTB.4	TTGCTTTCGTGTAAATTATGTAATGCAAAATTTTTTTAATCTTCGCCTTAATACTTTT TTATTTTGTTTTATTTTGAATGATGAGCCTTCGTGCCCCCCCTTCCCCCTTTTTTGTCC CCCAACTTGAGATGTATGAAGGCTTTTGGTCTCCCTGGGAGTGGGTGGAGGCAGCC AGGGCTTACCTGTACactgacttgagaccagttgAATAAAagtgcacaccttaaaaatGAggccaagtgtgacTTT GTGGTGTGGCTGGGTTGGGGG	270	ACTB reference sequence, short USE sequence prior to PAS for both mRNA variants, 37-bp after poly(A) site
T.RPL41.f	ACCGCTAGCTTGTTGCACCGTGGAGGCCACAGGAGCAGAAACATGGAATGCCAGAC GCTGGGGATGCTGGTACAAGTTGTGGGACTGCATGCTACTGTCTAGAGCTTGTCTCA ATGGATCTAGAACTTCATCGCCCTCTGATCGCCGATCACCTCTGAGACCCACCTTGC TCATAAACAAAATGCCCATGTTGGTCCTCTGCCCTGGACCTGTGACATTCTGGACTA TTTCTGTGTTTATTTGTGGCCGAGTGTAACAACCATATAATAAAtcacctcttccgctgttttagctg aagaattaaatCAtcttgtctattaTGTTTTTTATGGTTCCATCGGGTGGGGGTTTTCTGTCATTAGA GTTTGCCCTGTCACTACCTGTGCTATGGAGGGTATCAAAGCTATA	408	RPL41 reference sequence, full USE sequence prior to PAS, 100-bp after poly(A) site

Table 6-2: Table of synthetic terminator sequences

terminator	sequence 5' > 3'	length	Notes
Tm.synth.1	TGTAGACCCCTTGAAGAGGGGAGGGGCCTAGGGAGCCGCACCTTGTCATGTACcatcAATAAAgtaccctgtgctcaaccagtTActtgtcctgtcttattctagTGTGTTTT	113	GAPDH long USE, GAPDH spacer 1, GAPDH spacer 2, GAPDH poly(A) site, GAPDH spacer 3, DSE consensus
Tm.synth.2	TGTAGACCCCTTGAAGAGGGGAGGGGCCTAGGGAGCCGCACCTTGTCATGTACcatcAATAAAgtaccctgtgctcaaccagtTActtgtcctgtcttattctagTCTGTGTGTTGGTTTTTTGTGTG	128	GAPDH long USE, GAPDH spacer 1, GAPDH spacer 2, GAPDH poly(A) site, GAPDH spacer 3, Levitt consensus
Tm.synth.3	TGTAGACCCCTTGAAGAGGGGAGGGGCCTAGGGAGCCGCACCTTGTCATGTACcatcAATAAAgtaccctgtgctcaaccagtTActtgtcctgtcttattctagTGTGTTTTTCTGTGTGTTGGTTTTTTGTGTG	136	GAPDH long USE, GAPDH spacer 1, GAPDH spacer 2, GAPDH poly(A) site, GAPDH spacer 3, DSE+Levitt consensus
Tm.synth.4	TGTAATGTAATGTAATGTAAcatcAATAAAgtaccctgtgctcaaccagtTActtgtcctgtcttattctagGGTCTGGGGCAGAGGGGAGGGAAGCTGGGCTTGTGTC	109	4x USE consensus, GAPDH spacer 1, GAPDH spacer 2, GAPDH poly(A) site, GAPDH spacer 3, GAPDH DSE
Tm.synth.5	TGTAATGTAATGTAATGTAAcatcAATAAAgtaccctgtgctcaaccagtTActtgtcctgtcttattctagTGTGTTTT	80	4x USE consensus, GAPDH spacer 1, GAPDH spacer 2, GAPDH poly(A) site, GAPDH spacer 3, DSE consensus
Tm.synth.6	TGTAATGTAATGTAATGTAAcatcAATAAAgtaccctgtgctcaaccagtTActtgtcctgtcttattctagTCTGTGTGTTGGTTTTTTGTGTG	95	4x USE consensus, GAPDH spacer 1, GAPDH spacer 2, GAPDH poly(A) site, GAPDH spacer 3, Levitt consensus
Tm.synth.7	TGTAATGTAATGTAATGTAAcatcAATAAAgtaccctgtgctcaaccagtTActtgtcctgtcttattctagTGTGTTTTTCTGTGTGTTGGTTTTTTGTGTG	103	4x USE consensus, GAPDH spacer 1, GAPDH spacer 2, GAPDH poly(A) site, GAPDH spacer 3, DSE+Levitt consensus
Tm.synth.8	TGTAATGTAATGTAATGTAAAATAAAgtaccctgtgctcaaccagtTATGTGTTTT	56	4x USE consensus, GAPDH spacer 2, GAPDH poly(A) site, DSE consensus
Tm.synth.9	TGTAATGTAATGTAATGTAAAATAAAgtaccctgtgctcaaccagtTATCTGTGTGTTGGTTTTTTGTGTG	71	4x USE consensus, GAPDH spacer 2, GAPDH poly(A) site, Levitt consensus

Table 6-2, continued:

terminator	sequence 5' > 3'	length	Notes
Tm.synth.10	TGTAATGTAATGTAATGTAAAATAAAgtaccctgtgctcaaccagtTATGTGTTTTTCTGTGTGTTGGTTTTTTGTGTG	79	4x USE consensus, GAPDH spacer 2, GAPDH poly(A) site, DSE+Levitt consensus
Tm.synth.11	TGGTATtcattacaaacttgctcactacAATAAAtgaattttaagctttAAgatgaagtggcaTGTGTTTT	71	EEF1A1 short USE, EEF1A1 spacer 1, EEF1A1 spacer 2, EEF1A1 poly(A) site, EEF1A1 spacer 3, DSE consensus
Tm.synth.12	TGGTATtcattacaaacttgctcactacAATAAAtgaattttaagctttAAgatgaagtggcaTCTGTGTGTTGGTTTTTTGTGTG	86	EEF1A1 short USE, EEF1A1 spacer 1, EEF1A1 spacer 2, EEF1A1 poly(A) site, EEF1A1 spacer 3, Levitt consensus
Tm.synth.13	TGGTATtcattacaaacttgctcactacAATAAAtgaattttaagctttAAgatgaagtggcaTGTGTTTTTCTGTGTGTTGGTTTTTTGTGTG	94	EEF1A1 short USE, EEF1A1 spacer 1, EEF1A1 spacer 2, EEF1A1 poly(A) site, EEF1A1 spacer 3, DSE+Levitt consensus
Tm.synth.14	TGTAATGTAATGTAATGTAAtcattacaaacttgctcactacAATAAAtgaattttaagctttAAgatgaagtggcaTTTCTTTTAACAGTTACTATGTTGGAATTGGTTACAAATTTTGGAGTGGATTTCAAAAGTGAGAGCTAACTTCAGTTGATTTCAAGGTAGTGCTTGGCTTTTTTTGTTTA	187	4x USE consensus, EEF1A1 spacer 1, EEF1A1 spacer 2, EEF1A1 poly(A) site, EEF1A1 spacer 3, EEF1A1 long DSE
Tm.synth.15	TGTAATGTAATGTAATGTAAtcattacaaacttgctcactacAATAAAtgaattttaagctttAAgatgaagtggcaTGTGTTTT	85	4x USE consensus, EEF1A1 spacer 1, EEF1A1 spacer 2, EEF1A1 poly(A) site, EEF1A1 spacer 3, DSE consensus
Tm.synth.16	TGTAATGTAATGTAATGTAAtcattacaaacttgctcactacAATAAAtgaattttaagctttAAgatgaagtggcaTCTGTGTGTTGGTTTTTTGTGTG	100	4x USE consensus, EEF1A1 spacer 1, EEF1A1 spacer 2, EEF1A1 poly(A) site, EEF1A1 spacer 3, Levitt consensus
Tm.synth.17	TGTAATGTAATGTAATGTAAtcattacaaacttgctcactacAATAAAtgaattttaagctttAAgatgaagtggcaTGTGTTTTTCTGTGTGTTGGTTTTTTGTGTG	108	4x USE consensus, EEF1A1 spacer 1, EEF1A1 spacer 2, EEF1A1 poly(A) site, EEF1A1 spacer 3, DSE+Levitt consensus
Tm.synth.18	TGTAATGTAATGTAATGTAAAATAAAtgaattttaagctttAATGTGTTTT	51	4x USE consensus, EEF1A1 spacer 2, EEF1A1 poly(A) site, DSE consensus
Tm.synth.19	TGTAATGTAATGTAATGTAAAATAAAtgaattttaagctttAATCTGTGTGTTGGTTTTTTGTGTG	66	4x USE consensus, EEF1A1 spacer 2, EEF1A1 poly(A) site, Levitt consensus

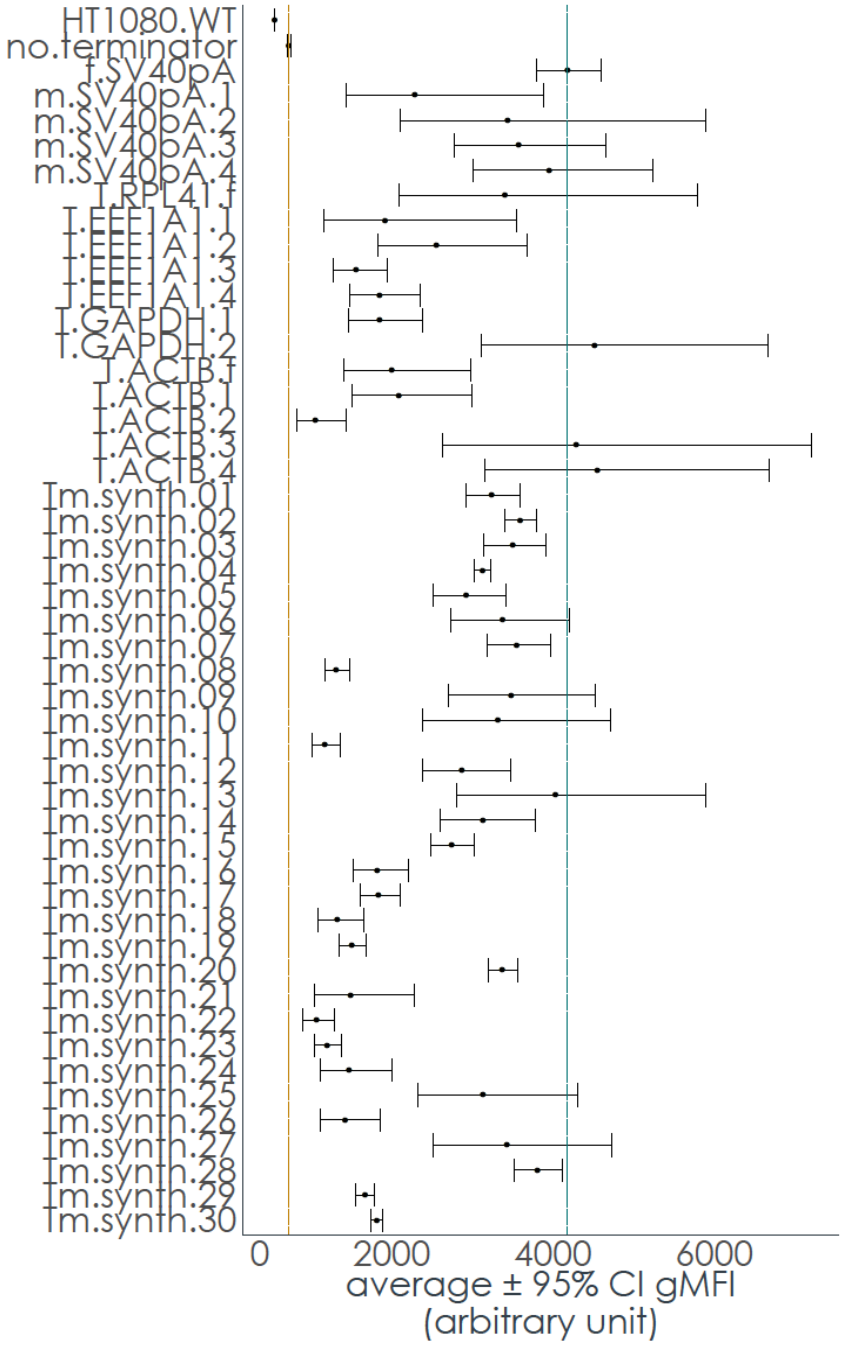
Table 6-2, continued:

terminator	sequence 5' > 3'	length	Notes
Tm.synth.20	TGTAATGTAATGTAATGTAAAATAAAtgaattttaagctttAATGTGTTTTTCTGTGTGTTGGTTTTTTGTGTG	74	4x USE consensus, EEF1A1 spacer 2, EEF1A1 poly(A) site, DSE+Levitt consensus
Tm.synth.21	TGTAATGTAATGTAATGTAAAATAAAagtgcacaccttaaaaatGATGTGTTTTTCTGTGTGTTGGTTTTTTGTGTG	77	4x USE consensus, ACTB spacer 2, ACTB poly(A) site, DSE+Levitt consensus
Tm.synth.22	TGTAATGTAATGTAATGTAAAATAAAagtgcacaccttaaaaatCATGTGTTTTTCTGTGTGTTGGTTTTTTGTGTG	77	4x USE consensus, ACTB spacer 2, poly(A) site consensus, DSE+Levitt consensus
Tm.synth.23	TGTAATGTAATGTAATGTAAAATAAAgtaccctgtgctcaaccagtCATGTGTTTT	56	4x USE consensus, GAPDH spacer 2, poly(A) site consensus, DSE consensus
Tm.synth.24	TGTAATGTAATGTAATGTAAAATAAAgtaccctgtgctcaaccagtCATCTGTGTGTTGGTTTTTTGTGTG	71	4x USE consensus, GAPDH spacer 2, poly(A) site consensus, Levitt consensus
Tm.synth.25	TGTAATGTAATGTAATGTAAAATAAAgtaccctgtgctcaaccagtCATGTGTTTTTCTGTGTGTTGGTTTTTTGTGTG	79	4x USE consensus, GAPDH spacer 2, poly(A) site consensus, DSE+Levitt consensus
Tm.synth.26	TGTAATGTAATGTAATGTAAAATAAAtgaattttaagctttCATGTGTTTT	51	4x USE consensus, EEF1A1 spacer 2, poly(A) site consensus, DSE consensus
Tm.synth.27	TGTAATGTAATGTAATGTAAAATAAAtgaattttaagctttCATCTGTGTGTTGGTTTTTTGTGTG	66	4x USE consensus, EEF1A1 spacer 2, poly(A) site consensus, Levitt consensus
Tm.synth.28	TGTAATGTAATGTAATGTAAAATAAAtgaattttaagctttCATGTGTTTTTCTGTGTGTTGGTTTTTTGTGTG	74	4x USE consensus, EEF1A1 spacer 2, poly(A) site consensus, DSE+Levitt consensus
Tm.synth.29	TGTAATGTAATGTAATGTAAAATAAAgtaccctgtgctcaaccagtCACGTGTTATTCATAAGCATT	67	4x USE consensus, GAPDH spacer 2, poly(A) site consensus, MC4R DSE[312]
Tm.synth.30	TGTAATGTAATGTAATGTAAAATAAAgtaccctgtgctcaaccagtCAGTTGTGTGTGTTG	61	4x USE consensus, GAPDH spacer 2, poly(A) site consensus, #7 DSE from Pérez Cañadillas, *et al* (CstF-64 RRM)[317]

Since previous work demonstrated that rationally designed synthetic terminators are functional in eukaryotes[305, 315], we used the generic mammalian terminator structure (**Figure 6-1**) as a scaffold and created 30 permutations (**Table 6-2**) by filling the spacer sequences with endogenous sequences from *EEF1A1*, *GAPDH*, and *ACTB*. We also explored putative USEs and DSEs from these endogenous genes in addition to their consensus sequences[302], as well as two DSEs identified in other reports[312, 317]. Lastly, variants were created using the polyadenylation site consensus dinucleotide 'CA' instead of the native site from *EEF1A1*, *GAPDH*, and *ACTB* to evaluate the impact of this key terminator element.

These terminators were paired with a moderate strength promoter characterized in Chapter 5 and derived from *EIF4A1* (636-bp promoter pEIF4A1.636) to drive hrGFP reporter expression. We quantified the impact of the various terminators on gene expression by measuring the transient hrGFP expression 48h post-transfection in HT1080 cells. Of the 4 truncated variants of the SV40 terminator, only m.SV40pA.1 affected gene expression negatively relative to the full length terminator (f.SV40pA) (**Figure 6-2**). Based on these results (**Figure 6-2**), we identified endogenous terminator sequences that are functional and influence gene expression to different extents, with 3 (T.GAPDH.2, T.ACTB.3, and T.ACTB.4) that were obviously comparable to the full length 222-bp SV40 terminator (f.SV40pA). From our 30 synthetic variants, the Tm.synth.13 terminator was clearly indistinguishable from the full length SV40 terminator (f.SV40pA) in terms of its impact on hrGFP expression (**Figure 6-2**). Overall, 14 of our 30 synthetic terminators comprised solely of endogenous and consensus sequences behave comparably to the SV40 terminator based on this expression data (**Figure 6-2**).

Figure 6-2: Transient hrGFP expression 48h post-transfection with various terminators in HT1080 cells



Expression reported as average geometric mean fluorescence intensity (gMFI) in arbitrary units. Error bars represent the 95% CI of the average gMFI from at least 3 independent transfections.

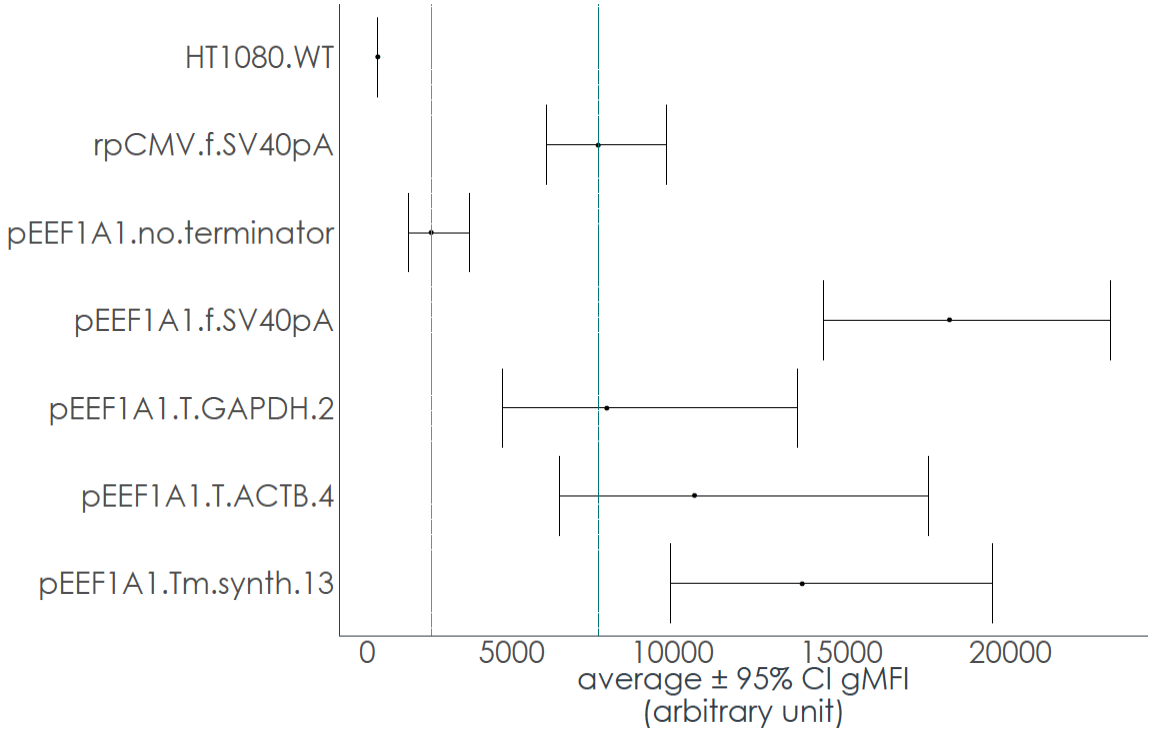
6.3.2. Full replacement of viral components with endogenous/synthetic elements

Based on the comparability of gene expression levels from several endogenous and synthetic terminators to the SV40 terminator (**Figure 6-2**), we established that these novel terminators are suitable replacements for the viral-derived sequence. We compared hrGFP expression between several transgene designs to further evaluate the possibility of completely replacing strong viral-derived regulatory elements. Our transgene containing viral-derived elements use the common strong CMV promoter coupled to the SV40 terminator while our endogenous and synthetic variants use the strong endogenous promoter (pEEF1A1.1356) with our 3 terminators T.GAPDH.2, T.ACTB.4, and Tm.synth.13. When we measured transient hrGFP expression 48h post-transfection, the 3 fully endogenous/synthetic constructs were comparable to, and 1.9x stronger (Tm.synth.13) than, the viral-derived construct, indicating that we can obtain high transient transgene expression in HT1080 without relying on viral-derived components (**Figure 6-3** and **Table 6-3**).

Most importantly, we identified endogenous and synthetic terminator variants that resulted in comparable gene expression relative to the common SV40 terminator (T.GAPDH.2, T.ACTB.3, T.ACTB.4, and Tm.synth.13, **Figures 6-2** and **6-3**). Interestingly, these novel variants were either 1.5x larger (333-bp) or nearly 2.4x smaller (94-bp) in sequence length (**Tables 6-1** and **6-2**) relative to the viral terminator (222-bp), indicating that a particular sequence length was not necessary for strong terminator activity. Furthermore, the 63-bp reduction in the DSE between T.ACTB.3 and T.ACTB.4 resulted in negligible impact on gene expression (**Figure 6-2**), suggesting that the 24-bp G/T-rich DSE from *ACTB* was sufficient. However, when comparing the other 3 *ACTB*-derived terminators with an extended 555-bp USE (T.ACTB.f, T.ACTB.1, and T.ACTB.2, **Figure 6-2**), the reduction of the corresponding DSE from the 452-bp or 87-bp version to the 24-

bp version resulted in much lower gene expression (compare T.ACTB.2 with other *ACTB*-derived terminators, **Figure 6-2**). Even within this small group of terminator sequences derived from the same endogenous sequence, the variable expression highlights that these key elements of a terminator have significant interplay.

Figure 6-3: Comparison of strong viral-derived elements with fully endogenous/synthetic elements for regulating gene expression.



Transient hrGFP expression 48h post-transfection in HT1080 cells reported as average geometric mean fluorescence intensity (gMFI) in arbitrary units. Error bars represent the 95% CI of the average gMFI from at least 3 independent transfections. rpCMV corresponds to reference CMV promoter, pEEF1A1 corresponds to full length 1356-bp promoter derived from *EEF1A1*.

Table 6-3: Fully endogenous/synthetic genetic elements compared with viral-derived elements based on hrGFP expression.

terminator pair	expression difference	$p_{adjusted}$
rpCMV.f.SV40pA-HT1080.WT	2.032	0.000
rpCMV.f.SV40pA-pEEF1A1.no.terminator	0.605	0.000
rpCMV.f.SV40pA-pEEF1A1.f.SV40pA	-0.411	0.000
rpCMV.f.SV40pA-pEEF1A1.T.GAPDH.2	-0.016	1.000
rpCMV.f.SV40pA-pEEF1A1.T.ACTB.4	-0.156	0.275
rpCMV.f.SV40pA-pEEF1A1.Tm.synth.13	-0.282	0.003

rpCMV corresponds to reference CMV promoter, pEEF1A1 corresponds to full length 1356-bp promoter derived from *EEF1A1*. Statistical significance determined by ANOVA and Tukey's HSD post-hoc testing. Expression difference corresponds to the difference of log mean hrGFP expression of the particular terminator relative to SV40.

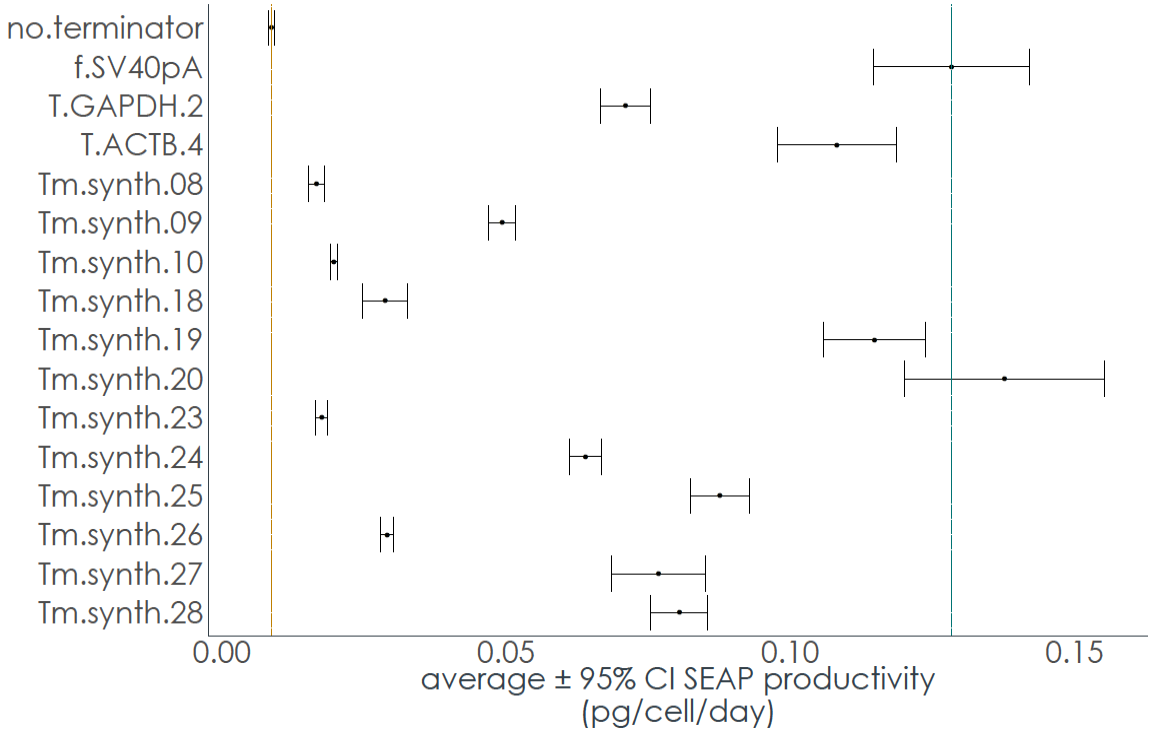
6.3.3. Confirming terminator functionality under different genomic context

While we designed these synthetic variants based on the generic mammalian terminator structure and therefore they should be independent of the gene/transgene they regulate, we wanted to confirm that their functionality remains comparable under two additional genomic contexts. We examined a subset of our endogenous and synthetic terminator designs for their ability to modulate the expression of secreted alkaline phosphatase (SEAP), a secreted reporter protein, and also their interplay with a stronger promoter (1356-bp promoter pEEF1A1.1356 derived from *EEF1A1*, commonly regarded as EF1α) to regulate hrGFP expression.

We observed similar trends in SEAP expression to our previous hrGFP data (**Figure 6-4**). With this subset of terminators, we obtained 11-fold dynamic range of expression, which is greater than the 5.7-fold range measured with our hrGFP reporter (**Figure 6-2**) for this same subset. Interestingly, the *ACTB*-derived terminator (T.ACTB.4) was only minimally influenced by this change in preceding coding sequence and remained comparable to the SV40 terminator, but the *GAPDH*-derived terminator (T.GAPDH.2) was only half as functional (**Figure 6-4**). Also, the synthetic terminator Tm.synth.19 behaved comparably to the SV40 terminator based on SEAP expression but not hrGFP expression

(compare **Figure 6-2** and **6-4**), suggesting some interaction between terminators and preceding coding sequence. Based on the SEAP reporter, only Tm.synth.19 and Tm.synth.20 were comparable to SV40 in terms of gene expression (**Figure 6-4**).

Figure 6-4: Transient SEAP expression 48h post-transfection with terminator subset in HT1080 cells

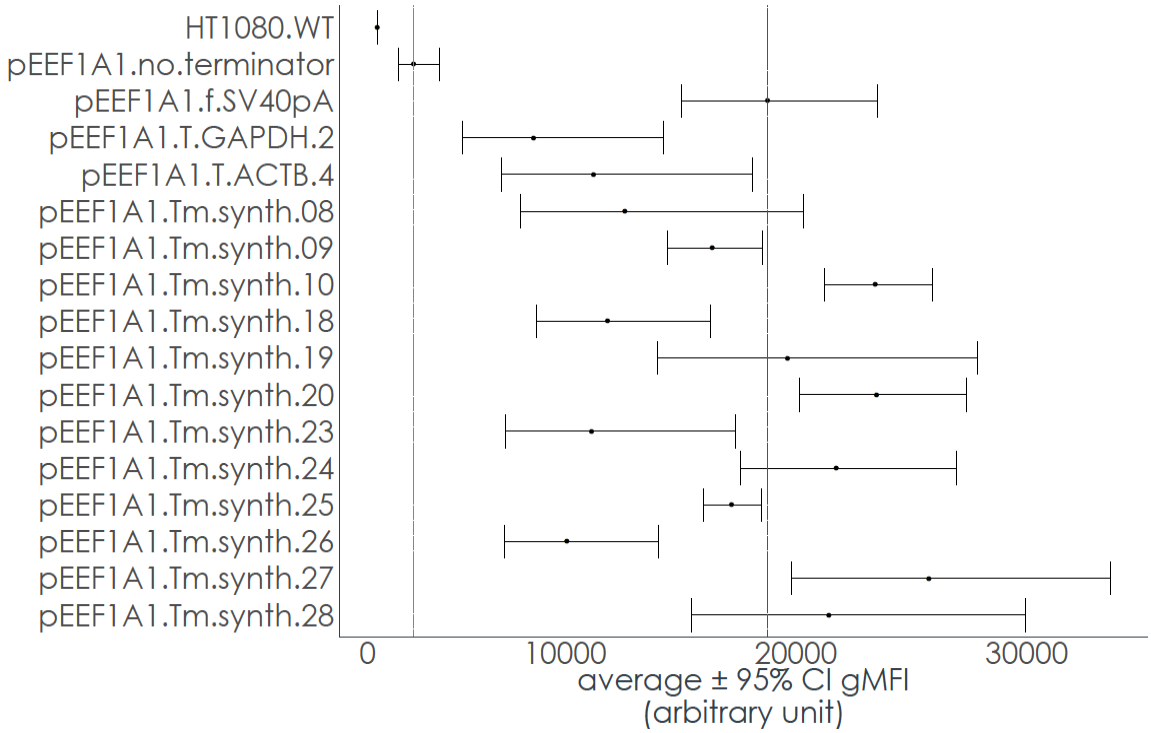


Expression reported as average specific productivity in pg/cell/day. Error bars represent the 95% CI of the specific productivity from at least 3 independent transfections.

When we replaced the moderate strength promoter (pEIF4A1.636) with the strong endogenous promoter (pEEF1A1.1356), our dynamic range of hrGFP expression was reduced to approximately 3.5-fold, essentially half of the range measured previously with our moderate strength promoter (**Figure 6-5**). However, a reduction in the expression range is expected since the mRNA generation rate of stronger promoters can overwhelm the degradation rate, consequently saturating the system. This pairing also revealed some additional context dependency on gene expression, most notably the weaker gene

expression associated with the endogenous terminators T.GAPDH.2 and T.ACTB.4 (**Figure 6-5**), mimicking the same observations with the SEAP reporter (**Figure 6-4**). However, under the context of a strong promoter (reflected by the marked increase in gMFI, compare **Figures 6-2** and **6-5**), 8 of the 12 synthetic terminators were comparable to the SV40 terminator based on gene expression levels (**Figure 6-5**).

Figure 6-5: Transient hrGFP expression 48h post-transfection in HT1080 cells when terminators are coupled to a strong endogenous promoter



Expression of various terminators with a strong promoter (pEEF1A1.1356) as average geometric mean fluorescence intensity (gMFI) in arbitrary units. Error bars represent the 95% CI of the average gMFI from at least 3 independent transfections.

6.3.4. Dissecting the key composition of mammalian terminators

The variable influence on gene expression by the terminator sequence affirmed the ability of the 3' UTR to tune gene expression and suggested that the key elements of these terminators can be dissected to reveal their contributions to this regulation. We compared

hrGFP and SEAP expression within a subset of these synthetic terminators, specifically interrogating the effects of spacer 2 (between the PAS and the polyadenylation site), the polyadenylation site, and the DSE sequence independently and between their interactions. This subset is comprised of Tm.synth.8-10, 18-20, and 23-28 (**Table 6-2**), which contains permutations of these three key terminator elements.

Analysis of the transient hrGFP expression driven by the moderate strength promoter (pEIF4A1.636) using ANOVA followed by Tukey's HSD post-hoc testing revealed that the DSE can significantly impact terminator functionality and gene expression ($p < 0.001$). Synthetic DSE sequences were clearly functional from the expression data (**Figures 6-2** to **6-5**) and they were stronger than the DSE consensus (**Table 6-4**). A broader ANOVA across the entire terminator subset revealed that the interactions between these three key elements in the terminator significantly impact gene expression (**Table 6-5**). Likewise, the same analyses conducted on the SEAP expression data and hrGFP expression data driven by the strong endogenous promoter revealed the same critical impact of the DSE and the interaction of these three key elements on gene expression (**Table 6-5**).

Table 6-4: Differential expression between a particular DSE and the DSE consensus from ANOVA analysis with Tukey's HSD post-hoc testing based on hrGFP expression driven by the pEIF4A1.636 promoter.

pairwise comparison	**differential expression**	**$p_{adjusted}$**
Levitt-consensus	0.140	0.035
MC4R-consensus	0.220	0.002
CstF64.RRM-consensus	0.271	0.000
consensus+Levitt-consensus	0.570	0.000

Table 6-5: ANOVA analysis with Tukey's HSD post-hoc testing of the spacer 2, polyadenylation site, and DSE impact on gene expression in conjunction with two promoter strengths.

pEIF4A1.636-hrGFP		pEIF4A1.636-SEAP		pEEF1A1.1356-hrGFP	
Factor/interaction	**p-value**	**Factor/interaction**	**p-value**	**Factor/interaction**	**p-value**
spacer.2	0.033	spacer.2	0.000	spacer.2	0.116
PA.site	0.919	PA.site	0.755	PA.site	0.539
DSE	<2.20E-16	DSE	3.351E-08	DSE	1.08E-12
spacer.2:PA.site	9.96E-11	spacer.2:PA.site	4.94E-05	spacer.2:PA.site	0.527
spacer.2:DSE	0.349	spacer.2:DSE	0.022	spacer.2:DSE	0.035
PA.site:DSE	0.779	PA.site:DSE	0.503	PA.site:DSE	9.01E-05
spacer.2:PA.site:DSE	2.52E-10	spacer.2:PA.site:DSE	1.23E-03	spacer.2:PA.site:DSE	0.361

Analysis corresponds to hrGFP expression driven by the pEIF4A1.636 promoter, SEAP expression driven by the pEIF4A1, and hrGFP expression driven by the pEEF1A1.1356 promoter.

We briefly investigated the interactions between these key elements that could guide subsequent terminator design. Based on the expression data, the DSE itself can drastically impact the resulting gene expression in a sequence-dependent manner and the DSE consensus sequence alone is sufficient for terminator activity (**Figures 6-2**, **6-4**, and **6-5**). While the DSE consensus was functional as a terminator element, we observed significantly stronger gene expression with other synthetic DSEs, suggesting that a consensus sequence may not be optimal (**Table 6-4**). These observations were recapitulated between the endogenous terminators and its synthetic variants with different DSEs (compare T.EEF1A1.1, T.EEF1A1.2, Tm.synth.11, Tm.synth.12, and Tm.synth.13, **Figure 6-2**), and the resulting gene expression again did not correlate with the DSE sequence length. Interestingly, the synthetic DSEs in this terminator context yielded stronger gene expression than the endogenous *EEF1A1* DSEs, yet the same effect was not observed with *GAPDH*-derived sequences (compare T.GAPDH.2, Tm.synth.01, Tm.synth.02, and Tm.synth.03, **Figure 6-2**).

Likewise, changing the endogenous USE sequence to the consensus 4x repeat did not result in a consistent impact on gene expression based on *EEF1A1*-derived (compare

T.EEF1A1.2, T.EEF1A1.4, and Tm.synth.14) and *GAPDH*-derived (compare T.GAPDH.1, T.GAPDH.2, and Tm.synth.04) sequences (**Figure 6-2**). Collectively, the data corroborates previous bioinformatics analyses suggesting that the DSE has a greater impact on polyadenylation processing than the USE[318]. In addition, changing the endogenous polyadenylation site to the dinucleotide consensus sequence had no measurable impact on gene expression (**Table 6-5**), further highlighting the suboptimal nature of a consensus sequence of a terminator element. Although, certain combinations of the spacer between the PAS and polyadenylation site (spacer 2), the polyadenylation site, and the DSE had significant impact on gene expression, this impact was masked when expression was driven by a strong promoter (**Table 6-5**). These results revealed that the impact of the terminator on gene expression can be tailored through these particular elements.

6.4. CONCLUSIONS

In this work, we generated and characterized the impact of 12 endogenous and 30 synthetic terminator variants on their ability to modulate gene expression. Critically, all of these variants were functional, and many variants required significantly reduced sequence space compared to a common, strong viral terminator. For the initial evaluation of these terminators, we opted to use a moderate strength promoter in order to observe a greater dynamic range in gene expression due to changes in the terminator. As a regulatory element, the terminator should ideally function in a predictable manner regardless of the preceding coding sequence or the relative strength of the promoter driving that particular transcript. However, our subsequent characterizations of a subset of these endogenous and synthetic terminators with an alternate reporter and stronger promoter showed that these terminators do exhibit some context dependency (compare **Figure 6-2**, **6-4**, and **6-5**). The

relative measured gene expression levels between terminators were inconsistent when the protein reporter was changed, but a large dynamic range in expression was maintained. This is expected since the terminator elements should fine-tune gene expression, whereas the promoter element responsible for generating mRNA was unchanged. This is further corroborated by the compression of this dynamic range when a subset of our terminators were coupled to a strong endogenous promoter (**Figure 6-5**). Previous work in our group reported similar findings in *S. cerevisiae*, where the terminator exerted greater influence on gene expression when paired with lower-strength promoters[307].

The premise of engineering novel terminators for high gene expression predominantly aims to replace the viral-derived components with endogenous and synthetic counterparts that are comparable in terms of their impact on gene expression. Based on the work for replacing the promoter with synthetic *de novo* designs (Chapter 4) and with endogenous variants (Chapter 5), the replacement of viral terminators with the endogenous and synthetic terminators characterized in this work enables a full conversion of viral-derived regulatory sequences to purely endogenous and synthetic sequences. This replacement was not detrimental to gene expression, but instead we achieved comparable or stronger expression using an endogenous promoter and terminator (**Figure 6-3**). However, much like the promoter engineering and characterization described in Chapters 4 and 5, terminator engineering still requires empirical characterization in mammalian cells even when using a generic structure as a scaffold based on the context dependency observed in this work. The dissection of several key elements in terminators affirmed the interaction between these elements (**Table 6-5**), but design rules for predictive terminator design to achieve a certain expression level were not evident from this work. Similar to previous efforts in yeast[305, 307], this work expanded the set of characterized terminators available for tuning gene expression. Ultimately, the ability to use a terminator as an

additional regulatory element to modulate transgene expression facilitates the transition from mammalian cell engineering for solely high transgene expression to engineering metabolic pathways and other applications that require precise control of gene expression.

Chapter 7: Conclusions and Major Findings

The engineering of mammalian host cells over the past several decades predominantly focused on industrial biotechnology and basic research applied to human health and disease. The advent of recombinant protein production in model organisms and mammalian cell types drastically changed the landscape of healthcare, simultaneously exposing the complex cellular and molecular biology of mammalian cells. The clinical and industrial successes of these bioengineered products fundamentally relied on many of the genetic elements governing gene expression that were derived from virology work. The extensive and prolonged use of these viral-derived elements revealed a critical limitation in addition to their inherent immunogenic potential: the sequences were eventually recognized as foreign elements and silenced[226, 229, 319], reducing the robustness required of viable industrial processes. Furthermore, many of cellular engineering efforts focused primarily on increasing gene expression or enabling applications for high gene expression with little regard for potential applications.

The work described in this dissertation aims to shift the mindset away from treating mammalian cells solely as recombinant protein factories and facilitates the exploration of their complex metabolism and cell biology for other therapeutic applications and medical needs. This was accomplished by developing the approaches and methods that enable extensive control of gene expression in these cell types while removing the dependence on viral-derived elements for these tasks. This work can be immediately applied to current protein production applications in mammalian cells, but most importantly, the sufficiently generic methods and approaches can greatly benefit the growing number of metabolic engineering and immune cell engineering applications.

By using an unbiased genome integration approach, our work in Chapter 2 revealed that the gene locus/integration site can act as one of the levers for tuning expression. We identified ten recombinant human cell lines with stable, single-copy, high-level heterologous gene expression and subsequently their corresponding integration loci. These results indicated the importance of both non-protein coding regions and exonic regions for heterologous gene expression. Expression maps for each of these loci also demonstrated negligible perturbations caused to the surrounding genes by our transgene integration. Lastly, we demonstrated that *de novo*, targeted integration at one of the identified "transcriptional hot-spots" resulted in superior expression compared to the standard, illegitimate (random) integration approach. This work provided a much needed cataloging of potential genomic hot spots for gene expression and an easily adaptable approach to seek out other loci that can be linked together with emerging genome editing tools. Relating to immediate applications, the targeting of these advantageous integration loci can significantly reduce the time, labor and materials associated with cell line development for therapeutic protein production. This approach (and the specific targets identified) can be extended to other mammalian cell lines used for industrial protein production, including CHO and HEK293 and help speed their development processes. Given our measured 20-fold dynamic range in transcriptional activity based on this limited set of integration sites, an extensive characterization of additional loci can enable facile exploitation of the integration site as an additional layer of gene regulation.

To truly utilize an integration locus for controlling gene expression, clonal cell populations must be easily generated and isolated that contain the desired integration. The work in Chapter 3 set the groundwork for improving the success of specific, targeted integrations. We verified a locus co-targeting approach[181] for validating gRNA target sites and corroborated the coordination of these gRNA with Cas9 to induce targeted DSBs by

Sanger sequencing and qPCR methods. A thorough characterization of this integration process is essential for incorporating pathways and complex expression programs associated with metabolic engineering and immune cell engineering in addition to gaining insight into the critical parameters that govern this process for cell line development.

In our work, we explored two avenues to engineer promoters, the key element that drives transcription. First, we demonstrated the utility of an 'omics guided workflow to create novel, synthetic promoters for a mammalian cell host in Chapter 4. By applying this generic workflow, we can develop a set of *de novo* promoters to properly tune the levels of protein intermediates in a metabolic pathway through adjusting the TFBS composition and arrangement. Moreover, these synthetic promoters had an overall reduced sequence footprint compared to reference promoters, thus increasing the overall utility for applications such as (heterologous) metabolic pathway designs[6] and gene delivery vector designs[253], akin to efforts in reducing the regulatory sequence footprint in a conventional metabolic engineering host[254]. The synthetic promoters designed in a single iteration of our workflow for high expression were comparable to a strong viral-derived and endogenous promoter, highlighting the potential of replacing viral-derived elements. Importantly, replacing viral-derived elements with these diverse, synthetic sequences would potentially reduce their susceptibility of epigenetic silencing and increase their long-term stability, which are two key attributes required in metabolic engineering and immune cell engineering applications. Ultimately, this workflow incorporates contemporary high-throughput methodologies to construct functional promoter elements that can facilitate a myriad of applications ranging from the study of fundamental processes to immediate use in large-scale industrial processes.

Second, we explored a more traditional approach of evaluating putative regulatory regions derived from segments of endogenous sequences to create hybrid promoters in

Chapter 5. We found that the core promoter element can be rationally engineered and expanded with repeats to achieve a certain level of functionality in a specific cell type, but the major regulation in gene expression was due to the enhancer region, corroborating the findings in Chapter 4. Our data revealed that the concomitant introns found in endogenous promoters correlated with strong expression, yet they were not necessary for strong expression. This was exemplified by our 546-bp hybrid promoter (ACTBe.182-pEEF1A1.356) containing sequences derived from *ACTB* and *EEF1A1* that was comparable to the 660-bp viral promoter, which was further improved by adding the *GAPDH* intron 1 (ACTBe.182-pEEF1A1.350.Gi1). Interestingly, the 182-bp *ACTB*-derived enhancer did not impact gene expression relative to its variant containing the *GAPDH* intron 1 (pEEF1A1.350.Gi1), indicating that the enhancer and intron were not purely additive in their effect on gene expression. Subsequent hybrid promoters containing short endogenous introns or a synthetic intron showed that the intronic region can further modulate promoter activity. Nonetheless, we created a set of hybrid promoters with expression levels ranging from minimally detectable expression to strong expression characteristic of a viral promoter. Some of these hybrid promoters were comparable to if not stronger than a strong viral promoter, expanding the ability to tune gene expression without viral-derived elements. Critically, this work highlighted the strong impact on gene expression by the intronic region in mammalian cells, and their impact warrants further characterization of the mechanisms responsible.

Lastly, many of the previous efforts around terminators focused on understanding the critical key elements that comprise this essential regulatory region in mammalian cells and other eukaryotic systems. Single, one-off, synthetic variants were reported for mammalian cells[315] and yeast[320] yet terminators were not systematically studied to harness their potential for modulating gene expression levels until this work in mammalian cells

and recently in yeast[305, 307]. We designed and evaluated a panel of endogenous and synthetic terminators, many of which exhibited a similar impact on gene expression as a commonly used viral terminator. Under the control of a weak promoter, the terminator can impact net gene expression by 11-fold while under the control of a strong promoter, this impact is reduced to 3.5-fold. Nonetheless, our data suggests that terminators can still significantly impact the resulting expression levels regardless of promoter strength. Finally, we examined a full replacement of viral-derived components for both the promoter and terminator and found that our endogenous and synthetic variants were comparable at regulating gene expression, and one particular variant was 1.9-fold stronger than the fully viral-derived combination. Similar to the TFBS composition impact on promoter strength, our analysis of the key elements that comprise terminators revealed that they are the drivers for the differential impact on the final gene expression. Therefore, our data indicates that terminators are additional tools for fine-tuning expression in mammalian cells, and we can rationally engineer them by specifying the key elements that comprise terminators.

Collectively, by selecting the transgene integration site, specifying an appropriate promoter strength for transcriptional activity, and regulating that transcript with an engineered terminator, we have an unprecedented ability to systematically control gene expression in mammalian cells. Without doubt, these approaches and methods enabled the replacement of viral-derived elements and can be readily adapted to a variety of mammalian cell types used in current applications. The multiplexing capability of these approaches and methods, along with the conversion away from viral-derived sequences, moves mammalian cell engineering forward in the metabolic engineering and immune cell engineering fields where precise and prolonged gene expression programs are required for their success.

In the broader sense, this work addresses several key challenges facing established and nascent mammalian cell engineering applications as therapeutics. The influence of the integration site on gene expression highlighted in our work strictly emphasizes the need for targeted transgene integrations into the host cell genome instead of relying on integrations into random loci. Not only can targeted integrations expedite cell line development for therapeutic protein production by reducing screening efforts, it can also build in additional quality assurance and quality control into the process from a federal regulation perspective ("Quality by Design"). Our work described in Chapter 2 established a straightforward approach to identify potential integration sites, which can be combined with high-throughput sequencing technologies to build a larger catalog of beneficial integration sites. Concurrently, adoption of our endogenous/synthetic promoter and terminator elements from Chapters 4-6 can culminate into a transgene design capable of driving extremely high protein expression without the use of viral-derived elements, potentially escaping epigenetic silencing of the transgene over time. Furthermore, this avoidance of viral-derived elements with robust activity is especially critical to *in vivo* applications such as immunotherapies/adoptive cell therapies so that a predictable medical benefit in patients can be maintained. Thus, the work described in this dissertation has immediate impact on a clinically- and commercially-relevant industry and emerging technologies.

Chapter 8: Proposals for future work

This work expanded the ability to fine tune gene expression in mammalian hosts, and in doing so, created endogenous and synthetic genetic elements that can replace viral-derived regulatory elements with comparable impact on gene expression. In this chapter, additional considerations and unanswered questions along with potential future studies in these topics are discussed.

In Chapter 2, we established a set of genomic loci that function as transcriptional hot-spots through our genome-wide screening approach. We isolated clonal populations with stable, strong hrGFP expression through FACS, with the underlying assumption that the clones with the highest expression levels correspond to integrations sites that favor transcription. We verified that the clonal populations contained a single copy of our reporter and identified the transgene integration sites in these clonal populations with low-throughput methods. Even within our set of characterized integration sites, two of the loci were unplaced, clearly highlighting a challenge in our approach and suggesting opportunities for subsequent studies.

For example, we can exploit the maturation of high-throughput sequencing (HTS) technologies to improve our success in identifying integration sites. While it would still be difficult to identify the integrations sites that are located in heavily repeated sequences such as short and long interspersed sequences (SINEs and LINEs)[321] with HTS techniques, it would be possible to gain additional insight into these integration sites with a large enough sample set. If high expression cell populations contain preferential integration sites, the classification or other attributes of these loci can facilitate subsequent *de novo* targeting of favorable sites. Given the high transcriptional activity associated with these identified loci, preferential recovery of integrations at these sites may also be biased by the

local chromatin remodeling at these loci as a result of DNA repair[322]. In addition, extensive cataloguing of these sites may reveal loci that are prone to DSBs with implications on cellular health[323]. Lastly, these transcriptional hot-spots can be evaluated in other cell lines to confirm that the impact on gene expression is maintained.

In Chapter 3, we laid most of the groundwork to enable a thorough assessment of targeted transgene integration into common mammalian host cells. A highly selective targeted integration and screening process would effectively isolate cell populations with the desired transgene(s) integrated at the target locus while an efficient process would incorporate as much of the donor template as possible. By taking an approach that should reduce the formation of DSBs and remove unwanted integrations at off-target loci, we would be able achieve a highly selective and efficient process. We verified that Zeocin™ is a highly selective agent for common mammalian host cells but we have yet to confirm the expression cytosine deaminase-uracil phosphoribosyltransferase fusion gene serves as an effective negative selection marker that responds to 5-fluorocytosine[206, 209] for removing integrations at undesired loci. An effective negative selection agent is imperative to this combination approach and requires thorough characterization.

Previous work demonstrated that multiple gRNAs targeting the same locus drastically improved gene silencing[79], suggesting that we can exploit this approach to increase the propensity for DSBs at the target site and subsequently improve the frequency of successful integrations through DNA repair. This putative improvement can still be combined with high-fidelity variants of the Cas9 nuclease responsible for DSB formation. Alternatively, the delivery of the Cas9 enzyme can be supplied either as mRNA or in its protein form[324] to reduce potential off-target activity due to the prolonged Cas9 nuclease expression when supplied as an expression vector. Interestingly, it was reported that HDR is favored over NHEJ in transcriptionally active chromatin[325], hinting that the combined

positive and negative selection approach is viable. Still, it is unclear whether the recovered cell populations from coupled selections would arise from a single cell or many cells that survived the selection pressures, warranting additional studies to fully investigate the integration efficiencies with this approach.

In Chapter 4, we established a workflow based on a Design-Build-Test paradigm that enabled *de novo* generation of promoters capable of driving high levels of gene expression. Through a single iteration of our workflow, we achieved our goal of designing synthetic promoters that were capable of driving high gene expression and were comparable to a strong viral-derived promoter. However, there are several avenues to explore and dissect our synthetic designs. For one, the workflow essentially represents a bottom-up approach to building functional genetic elements, but we can now apply a top-down approach to remove the transcription factor binding sites systematically to understand which of those building blocks were essential. This would address whether the number of putative sites available or the specific combination of sites were necessary for our observations. Furthermore, once the necessary and sufficient building blocks are identified, an exhaustive library of sequences containing permutations of those building blocks can explore the spacing and arrangement relationship between them in depth.

In Chapter 5, we explored additional promoter designs that could be viable for regulating gene expression in mammalian hosts. This chapter considered rational designs based on known synthetic sequences or mimicking traditional promoter characterization by examining a segment of the 5' UTR of highly expressed genes. However, superfluous sequences are likely included in promoter designs that simply take a segment of DNA sequence to confer promoter activity, which was common practice until *de novo* designs. Thus, employing a rational and a library screening approach in conjunction can identify novel sequences that would exhibit the desired functional characteristics, as demonstrated

previously in our research group in *S. cerevisiae*[248, 254]. Applying this approach to the core promoter region is particularly important to building fully synthetic regulatory elements for mammalian hosts and replacing all viral-derived sequences. In Chapters 4 and 5, the promoter engineering work primarily focused around designs for the proximal and distal enhancer regions. The multiple core promoter exploration along with the rational core promoter work here and previously reported in literature[267] suggest that particular element is highly tunable despite its essential interactions with basal RNA polymerase II machinery. Lastly, the minimal improvements in gene expression may not fully reflect the promoter strength if the translational machinery is saturated due to high levels of mRNA. Therefore, it would be an important confirmation to measure transcript levels with qPCR to verify whether the promoter strength was impacted by our designs.

In Chapter 6, we exploited the native structure of mammalian terminators to isolate putative endogenous terminators and to guide our synthetic designs. This work affirmed that the terminator can also tune gene expression in a sequence dependent manner, and we are able to achieve levels of gene expression using fully endogenous/synthetic elements that are comparable to expression driven by viral-derived elements. While these components showed minor context dependency (differences based on preceding coding sequence and promoter strength), it is crucial to demonstrate that these designs are functional in other mammalian cell types. Since our terminator designs are derived from endogenous and/or consensus sequences, we expect comparable behavior in other mammalian cell types based on the highly conserved machinery[326].

Critically, it is implied that the differential expression is likely due to the termination efficiency (lack of terminator read-through/stopping transcription) based on the comparison of gene expression between any terminator used and our no-terminator control. However, with our diverse set of terminator designs and their distinct permutations

of key elements within the terminator, it would be interesting to characterize whether the measured gene expression correlates with termination efficiency and/or the mRNA half-life. These results would reveal mechanistic insight into terminator function and simultaneously offer additional design parameters. For example, it would be possible to design a terminator with some termination inefficiency to enable polycistronic messages or that selectively tunes mRNA abundance with a specific half-life.

In closing, the studies demonstrated in this work provide some of the stepping stones necessary to truly push mammalian cell engineering away from being solely therapeutic protein factories and towards a new era of applications such as metabolic engineering and immunotherapy. We have yet to fully tap their potential in these applications where precise control of gene expression is inherently and critically tied to their success. In doing so, these future directions can uncover a deeper understanding of the cellular and molecular biology governing these cells, catalyzing new iterations of design approaches and methods to address ongoing medical needs.

Chapter 9: Materials and Methods

9.1. COMMON METHODS USED IN THIS WORK

9.1.1. Plasmid construction

Plasmids were generated by standard molecular biology techniques: restriction enzyme digestion, agarose gel DNA electrophoresis extraction, ligation, Gibson Assembly. Alternatively, plasmids were assembled using the In-Fusion cloning kit (Clontech).

9.1.2. Growth and media conditions

A suspension-adapted and serum-free HT1080 cell line, provided by Shire Pharmaceuticals and established from ATCC CCL-121, was used for all experiments. Cells were grown in CD 50/50 media, passaged every 48–72 h and seeded at 3 x 10^5 viable cells/mL. CD 50/50 contains 50% CD-CHO and 50% CD-293 (Invitrogen) with 4 mM glutamine and pH of 7.2. Cell viability, concentration, and size were measured using a Beckman Coulter ViCell. Shake flasks were maintained at 37°C, 5% CO_2, humidity above 80% and 125 rpm. Plasmids were propagated in *Escherichia coli* DH10β or Stellar strains using lysogeny broth with ampicillin at 37°C.

9.1.3. HT1080 transient transfections

To establish transient expression cultures, three batches of 12 x 10^6 viable cells were re-suspended in RPMI-1640 media (0.75 mL per cuvette) and transfected each with 50 μg of plasmid DNA (harboring our dual-reporter expression cassette with various promoters) using a 4 mm electroporation cuvette and Gene Pulser XCell (BioRad) at 350 V and 950 μF of capacitance.

9.1.4. Quantification of GFP expression

GFP expression profiles were determined using flow cytometry. 1.0 – 2.0 x10^6 cells were pelleted and suspended in sterile 1x DPBS 48-h post-transfection. These cell suspensions were analyzed using the BD LSRII Fortessa (UT Institute of Cellular and Molecular Biology Core Facility) with the following parameters: 340V FSC, 176V SSC, GFP voltages vary from 180V to 260V to obtain a suitable distribution within the quantification window. 25,000-30,000 events were collected from a subpopulation that represented live cells (as determined by gating WT cells) and these live cells were divided between GFP^- or GFP^+ by setting a threshold such that 1% of the WT cells are considered GFP^+ as an approximation of autofluorescence. All other samples were analyzed using the same gate for live cell subpopulation and threshold for GFP^+ cells.

9.1.5. Quantification of SEAP expression

The supernatant obtained from centrifugation to pellet cells was sampled for analysis using the NovaBright™ Secreted Placental Alkaline Phosphatase (SEAP) Enzyme Reporter Gene Chemiluminescent Detection System 2.0 (Invitrogen).

9.1.6. Genomic DNA extraction

Genomic DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega) from 3.5 x 10^6 cells per sample. This gDNA can be used to confirm the number of transgene copies or further digested with a restriction enzyme to facilitate qPCR analysis.

9.1.7. Data analysis

Data analysis for expression data conducted using MS Excel and R: The R Project for Statistical Computing versions 3.0.0 to 3.3.0.

9.2. CHAPTER 2 SPECIFIC

The methods corresponding to this chapter can be found in a previously authored publication: Cheng, J. K., Lewis, A. M., Kim, D. S., Dyess, T., & Alper, H. S. (2016). Identifying and retargeting transcriptional hot spots in the human genome. *Biotechnol J*, 11(8), 1100–1109. DOI: 10.1002/biot.201600015.

9.2.1. Plasmid Construction

Several key plasmid designs were used as shown in **Figure 2-1**. The pIRES-hrGFP plasmid (**Figure 2-1A**) was constructed through modification of pIRES-hrGFP-1a (Stratagene). The Zeocin™ resistance gene was amplified from pSV40-zeo2 (Invitrogen). The pHL-GFP plasmid (**Figure 2-1B**) was provided by Shire Pharmaceuticals. The pAML-Zeo plasmid (**Figure 2-1C**) was constructed as previously described [25]. pCMV-hrGFP-IRES-puro (pGP), pCMV-puro-IRES-hrGFP (pPG), pCMV-SEAP-IRES-puro (pSP), and pCMV-puro-IRES-SEAP (pPS) plasmids (**Figure 2-1D**) were derived from pAML-Zeo plasmid by replacing the Zeocin™ resistance gene for the puromycin resistance gene, and the hrGFP gene for the secreted alkaline phosphatase (SEAP) gene. This final set of vectors (**Figure 2-1D**) was designed for loci retargeting. The detailed construction of these plasmids, including specific primers, is described further in **Supplementary Material and Methods** of the publication[175].

9.2.2. Cell line development

Plasmid DNA was extracted from 150 mL of *E.coli* DH10β culture using the Qiagen HiSpeed Maxi Prep kit, digested at 37°C, and purified by phenol-chloroform extraction. To establish stable, green fluorescent protein (GFP) expressing cells, three batches of 12 x 10^6 viable cells were re-suspended in RPMI-1640 media (0.75 mL per

cuvette) and transfected with 50 μg of pIRES-hrGFP DNA using a 4 mm electroporation cuvette and Gene Pulser XCell (BioRad) at 350 V and 950 μF of capacitance. Transfection efficiencies were typically 80%. Cells were transferred to CD 50/50 media and recovered 48 hours before selection pressure was applied. Concentrations of 50, 100 and 250 μg/ml of Zeocin™ were used to establish the HT1080 libraries. Selective pressure was maintained until culture viability was above 90%.

9.2.3. Sterile FACS to isolate high hrGFP expression population

After stable cell selection was completed (approximately 15-25 days), cells were prepared for flow cytometry sorting and analysis. For each sort, 300,000 cells of the top 10-15% of the population (based on GFP expression) were isolated using a FACSAria. This population was transferred to a six well plate and split every 24-48 hours, expanding the population until another sort was feasible. This process was iterated twice to ensure stringent selection and sustained expression.

9.2.4. Single Cell Cloning

Single cell clones were established from the Zeocin™-resistant cell culture pools using dilution cloning. GFP fluorescence profiles of each clone were examined using flow cytometry. The copy number of GFP integrants was determined for each clone as previously described[41]. Detailed single cell cloning procedures are described further in **Supplementary Materials and Methods** of the publication[175].

9.2.5. Methods for identifying integration loci

Low-throughput methodologies for identifying integration loci rely on approaches that both isolate and amplify genomic DNA adjacent to the transgene. We utilized three primary approaches to identify the integration sites in our high expression clones: TAIL PCR, inverse PCR and plasmid recovery. TAIL PCR utilizes three interlaced PCR

reactions to amplify genomic fragments adjacent to the integrated transgene. Long primers, specific to the integrated sequence flanking the gDNA, along with an arbitrary, degenerate primer of 12-16 base pairs in length are used in each PCR reaction. Based on previous reports, we adapted a methodology that uses three interlaced PCR reactions to enrich the flanking genomic DNA fragment [327-330]. This methodology was used to successfully identify the integration loci for clones A, B, C, H, I and J, and further details for each clone can be found in **Supplementary Material** of the publication[175]. A second approach, inverse PCR, was adapted from previous reports[149, 150, 161, 162, 331]. This approach was used to successfully identify the integration locus for clone E and further details are discussed in **Supplementary Material** of the publication[175]. The third approach, plasmid recovery, provided better capture and recovery of genomic fragments, thereby increasing our coverage of the human genome. This approach was used to successfully identify the integration loci for clones D, F and G, and further details are discussed in **Supplementary Material** of the publication[175].

9.2.6. RNA extraction, cDNA synthesis, and quantitative PCR

Whole cell RNA was extracted using the RiboPure kit (Ambion) at 5 x 10^6 cells per sample. RNA was converted to cDNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Relative mRNA expression for genes of interest was measured and compared to a common housekeeping gene, *RPS11*. The primer pairs for each gene can be found in **Supplementary Material, Table S2,** (1-56) of the publication[175] and were designed using PrimerExpress software to ensure consistent primer lengths, amplicon lengths, GC content and melting temperatures, and minimal secondary structures across all primer pairs. Roche SYBR Green 2x master mix was used to prepare samples in triplicate and a standard deviation of less than 0.5 C_T units was imposed to

ensure consistent primer efficiency across pairs. The ViiA™ 7 Real-Time PCR System (Applied Biosystems) and software was used to run qPCR and analyze results. Melt-curves were reviewed for all primer pairs to ensure formation of a single product and no random binding and amplification occurred. The comparative C_T method was used to normalize measurements relative to *RPS11*, a highly expressed ribosomal gene. After normalizing to RPS11, relative mRNA expression across genes of interest was calculated using the comparative C_T method. A log fold change greater than 2 was considered statistically significant.

9.2.7. Genomic DNA extraction and copy number quantification

Genomic DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega) from 3.5 x 10^6 cells per sample and used to confirm the number of transgene copies. Roche SYBR Green 2x master mix was used to prepare samples in triplicate. The ViiA™ 7 Real-Time PCR System (Applied Biosystems) and software was used to run RT-PCR and analyze results. Copy number was measured based on a standard curve of each primer pair for a particular gene (**Supplementary Material, Table S2,** 78-83 of the publication[175]) compared to a common housekeeping gene, *RPPH1*.

9.2.8. Site-Specific Retargeting

The hCas9 plasmid, previously constructed by the Church group [144], was obtained from Addgene (41815). Two gRNA constructs were designed following recommendations from the Church group [144] and used to generate stable cell lines expressing hrGFP and SEAP (**Figure 1d**). The construction of these gRNA sequences, Grik1A and Grik1B, are described in detail in **Supplementary Material and Methods** of the publication[175].

9.3. CHAPTER 3 SPECIFIC

9.3.1. MIC_{75} determination

The maximum inhibitory concentrations for a particular selective agent was determined by subjecting healthy parental cells to a range of concentrations. Typically, 3 cultures were tested at each concentration and compared to the parental cells maintained under standard conditions. The impact of the selection pressure is monitored 6-12 days until the cultures reached a vital minimum viability threshold.

9.3.2. Sequence recovery by TOPO TA Cloning and Sanger sequencing

Genomic DNA from Cas9 edited cell populations were extracted using the standard method. The gDNA from various cell populations were amplified with region-specific primers to recover the target locus and subcloned into the pCR™4-TOPO® TA vector following manufacturer's protocol. The resulting products were transformed into *E. coli* TOP10 cells and plated on LB + 50 µg/mL Kanamycin + 20 µg/mL X-Gal plates. White colonies were selected and grown in liquid culture containing 50 µg/mL Kanamycin or 100 µg/mL ampicillin overnight. The plasmids from the confluent cultures were isolated by miniprep and submitted for Sanger sequencing using the T3 primer.

9.3.3. Quantitative PCR with custom designed probes

Using the same gDNA extracted for sequence recovery by TOPO TA Cloning and Sanger sequencing (described above), the edited regions were analyzed using quantitative PCR with custom design probes for the target loci. Prior to qPCR, 1 µg of gDNA was digested with *NruI* in 50 µL reaction (per gDNA) to reduce structure. The digested gDNA were analyzed with the iTaq™ Universal Probes Supermix (Bio-Rad) based on manufacturer's thermal cycling protocol. 600nM of each primer and 300nM of each probe was used per reaction.

9.4. CHAPTER 4 SPECIFIC

The methods pertaining to this chapter can be found in a previously authored publication: Cheng, J., & Alper, H. S. (2016). Transcriptomics-guided design of synthetic promoters for a mammalian system. *ACS Synth Biol*. Article ASAP. Publication Date (Web): June 7, 2016. DOI: 10.1021/acssynbio.6b00075.

9.4.1 Processing of gene expression data

Microarray data from the Illumina platform representing the growth of HT1080 WT cells under representative (industrial) bioreactor conditions was transformed with a logarithmic function in MS Excel and modeled using a 3-component Gaussian Mixture Model (GMM) using MATLAB software (Mathworks, R2014a). The GMM was generated by an expectation-maximization algorithm[332] to identify parameters (μ, σ, π) that described the 3 Gaussian components comprising the model. The parameters were randomly seeded (within the bounds of the values in the data set) 10 times independently, and each "run" of the algorithm had at least 50 iterations to identify converged parameter values for the entire log-transformed data set at each time point (t_1-t_4). Using the final parameters from each component, false-positive and false-negative estimates were calculated at a particular threshold expression value (*th*) using equations found in the **Supporting Information** of the publication[240]. In addition, the probability of that a particular expression value (*e.g.* median value of data set) could belong to a specific expression group (*i.e.* component of GMM representing high, moderate, or low expression group) can be determined using these finalized parameters.

9.4.2. Annotation and analysis of candidate promoter sequences

The JASPAR database[233] containing TFBS consensus sequences was adapted (full list of sequences used found in **Supporting Information Table S2** of the publication[240])

as a feature list in ApE, A plasmid Editor, v2.0.47 software. This procedure enables rapid annotation of any sequence length in both sense and anti-sense TFBS orientations. The list of putative sites in a particular sequence of interest is stored as a CSV file using MS Excel and then parsed using R: The R Project for Statistical Computing to tabulate all putative sites across promoter regions for comparison. The total putative sites are subsequently analyzed in MS Excel to determine TFBS frequencies.

Similarly, 2000-bp chromosome regions were annotated using the same process as candidate promoter sequences. The chromosome regions were selected at random (MS Excel, random number generator) and the number of chromosome regions selected reflected the relative size of the chromosomes: 10 regions were selected as a minimum (corresponding to the size of chromosome 21) and the number of regions were increased based on the chromosome's size relative to chromosome 21 (up to 53 regions for chromosome 1, **Supporting Information Table S5** of the publication[240]).

9.4.3. Plasmid construction

Plasmids were constructed by a combination of standard molecular biology techniques (restriction enzyme cloning, Gibson assembly, In-Fusion HD cloning kit). The essential elements required for plasmid propagation in *E.coli* (origin, beta-lactamase) from pUC19 were sub-cloned with the conventional SV40 late terminator and multiple cloning sites (*EcoRI*, *NotI*, *NheI*, *PmeI*, *XbaI*, and *NsiI*) using Gibson assembly to generate the base vector (**Supporting Information Figure S2a** of the publication[240]). Subsequently, the SEAP, IRES, and hrGFP sequences were sub-cloned from other plasmids into these cloning sites (**Supporting Information Figure S2b** of the publication[240]). The region 5' to the start codon of the human CMV immediate-early gene (M60321.1, nucleotides 1-2105) was synthesized and supplied in a pUC57 vector (GenScript); this entire length is

considered the "reference" promoter in this work. Promoter variants were synthesized as dsDNA gBlocks (IDT). The reference promoter and its variants were cloned between the *NotI* and *NheI* sites immediately 5' of the SEAP CDS. To evaluate the synthetic proximal promoter/enhancer region, sequences were synthesized as Ultramer oligonucleotides and annealed (IDT) and cloned between the *NotI* and *EcoRI* (5' of *NotI*) sites. For direct comparison to the reference promoter and its variants, the hCMV IE core promoter region (joining exon 1 and 2: 1110-1264…2089-2105, 172-bp total) was retained from the reference promoter and cloned between the *NotI* and *NheI* sites. Alternatively, the minimal (**Supporting Information Figure S2c** of the publication[240]) and full-length variants of the promoter derived from the *EEF1A1* gene (EF1a) were sub-cloned from genomic DNA extracted from HT1080 WT cells using primers found in (**Supporting Information Table S6** of the publication[240]). Plasmid DNA was extracted from 150 mL of *DH10β* culture using the Qiagen HiSpeed Maxi Prep kit or the Zymo Research ZymoPURE Maxi Prep kit, and further purified by ethanol precipitation.

Synthetic promoter elements used in variants synth.v1, v2, and v3 (**Supporting Information Figures S2d-S2f** of the publication[240]) were cloned between the *EcoRI* and *NotI* sites 5' of the core promoter (between the *NotI* and *NheI* sites) using primers found in **Supporting Information Table S6** of the publication[240]. hCMV IE promoter variants were generated using the "synthprom" algorithm[248] with 45% GC content to approximate the overall human genome and reference promoter sequence and dsDNA fragments were constructed as gBlocks gene fragments (IDT). These fragments were amplified using primers found in **Supporting Information Table S6** of the publication[240] and cloned between the *NotI* and *NheI* sites in the dual-reporter expression vector (**Supporting Information Figure S2b** of the publication[240]). Three variants were generated to vary TFBS arrangement found in the hCMV IE promoter: native, sequential, and random

variants. The native variant (**Supporting Information Figure S2g** of the publication[240]) contained the same TFBS arrangement as the reference sequence. The sequential variant (**Supporting Information Figure S2h** of the publication[240]) contained TFBSs arranged in a grouped manner (*e.g.* multiple instances of a TFBS were placed adjacent to each other or sequentially from 5' to 3'). Lastly, the random variant (**Supporting Information Figure S2i** of the publication[240]) contained TFBSs arranged randomly.

9.4.4. Transfections

Transient transfections for HT1080 were conducted based on protocol described above. Transient transfections for HEK293 were conducted by nucleofection using the Nucleofector™2b device (Lonza Biologics) with the Nucleofector® Kit V (Lonza Biologics).

9.5. CHAPTER 5 SPECIFIC

None.

9.6. CHAPTER 6 SPECIFIC

None.

Appendix A: Primers and gBlocks® Gene Fragments

A.1. COMMON PRIMERS USED IN THIS WORK

Table A1: Primers used for Sanger sequencing to confirm plasmid construction

Primer	notes	sequence
S1	pCMV variant 5' ⇒ 3'	ATCCGCTAGCGATTACGCCAAGCTC
S2	region between HgHB/IRES 3' ⇒ 5'	AGTCGTCGAGGAATTGCTATTATTT
S3	HgHB 5' ⇒ 3'	CAGAAGCGCGGCCGTCTGGACCGAT
S4	SV40 late poly(A) 5' ⇒ 3'	GCTTTATTTGTGAAATTTGTGATGCTATTGC
S5	IRES 5' ⇒ 3'	ACATGCTTTACATGTGTTTAGTCGA
S6	SV40 late poly(A) 5' ⇒ 3'	AAAAAAATGCTTTATTTGTGAAATTTGTGATGC
S7	IRES 3' ⇒ 5'	CAATATGGTGGAAAATAACATATAGACAAACGCAC
S8	pSEAP2-basic plasmid 5' ⇒ 3'	GGTACCGAGCTCTTACGCGTGCTAG
S9	SV40 late poly(A) 3' ⇒ 5'	GCATTTTTTTCACTGCATTCTAGTTGTGGT
S10	F1 origin 3' ⇒ 5'	GTTCTTTAATAGTGGACTCTTGTTCCAAACTGG
S11	SEAP 5' ⇒ 3'	AGGTGGAGGCCGAAAGTACATGTTT
S12	ColE1 origin	TTCGCCACCTCTGACTTGAGC
S13	core pCMV 5' ⇒ 3'	AGGTCTATATAAGCAGAGCTCGTTTAGTGA
S14	core pSV40 5' ⇒ 3'	CCTCTGAGCTATTCCAGAAGTAGTGAGG
S15	hrGFP 5' ⇒ 3'	GCACCGCCTGGAGAAGACCT
S16	mStraw 5' ⇒ 3'	CATCGTCGGCATCAAGTTGGACATC
S17	hrGFP 3' ⇒ 5'	CTCCAGGTTCACCTTGAAGCTCATGAT
S17b	hrGFP 3' ⇒ 5'	GTTCACCTTGAAGCTCAT
S18	pCMV 5' ⇒ 3'	CCCATTGACGCAAATGGGCG
S19	AmpR 5' ⇒ 3'	CTCTCAAGGATCTTACCGCTGTTGA
S20	TB 3' ⇒ 5'	TGTTTTGATGGAGAGCGTATGTTAG
S21	ColE1 origin 5' ⇒ 3'	AGGATTAGCAGAGCGAGGTATGTAG
S22	AmpR 3' ⇒ 5'	GCCCTCCCGTATCGTAGTTATC
S23	AmpR 5' ⇒ 3'	GCTCGTCGTTTGGTATGGCTTC
S24	SEAP 3' ⇒ 5'	TTCTCCTCCTCAACTGGGATGATG
S25	CMV enhancer 3' ⇒ 5'	CGTAAGTTATGTAACGCGGAACTCCA
S26	CMViA 3' ⇒ 5'	ATGGATTAGTAATGGAAAGTATCGTCACC
S27	ref pCMV 5' ⇒ 3'	GCTATTGGCCATTGCATACGTT
S28	CMV enhancer 5' ⇒ 3'	AATCAACGGGACTTTCCAA
S29	CMV iA 5' ⇒ 3'	CAGACATAATAGCTGACAGAC
S30	TB 5' ⇒ 3'	ACTAACATACGCTCTCCATCAAA
S31	pCLUAP1 5' ⇒ 3'	CTTCATTTGCATATTAGCCATCC
S32	pGAPDH.2000 5' ⇒ 3'	GAAAGGCAATCCCAGAAAGG

Table A1, continued:

Primer	notes	sequence
S33	pPGK1 5' ⇒ 3'	GCTTCTGGGAATAACATCACC
S34	Grik1B 3' homology 5' ⇒ 3'	TTACCTAGGACTTCCACCTAAT
S35	fcy1.fur1 5' ⇒ 3'	CTTCAAAGTGGGACCAGAAG
S36	fcy1.fur1 5' ⇒ 3'	CTACACCATCATCAGGAACAAG
S37	HBB terminator 5' ⇒ 3'	GCATCTGGATTCTGCCTAATAA
S38	pVIM 5' ⇒ 3'	GGGACTACAGAACACCTACA
S39	pEEF1A1 5' ⇒ 3'	CGGAGCTGAGAGTAATTCATAC
S40	pTPT 5' ⇒ 3'	CCAAGGGTTGCATTCTTACTC
S41	pTUBA1B 5' ⇒ 3'	GGGTGGTTCCCTAACATTC
S43	pUBC 5' ⇒ 3'	CTGCCACGTCAGACGAA
S44	pGAPDH.2500 5' ⇒ 3'	CTACCAGCATTTGTGGGAA
S45	pGAPDH.3000 5' ⇒ 3'	ATTAGCCGGGCGTATTG
S46	pF2R.2000 5' ⇒ 3'	GCCTCTGATCGTACTTTCTC
S47	GAPDH exon 1 3' ⇒ 5'	CTGCGGGCTCAATTTATAG
S48	Grik1B 3' homology 3' ⇒ 5'	ggatgcgtgtgacaaagatag
S49	hCas9 5' ⇒ 3'	GCCTGTCTGAGTTGGATAAAG
S50	hCas9 5' ⇒ 3'	GACACCACCATAGACAGAAAG
S51	hCas9 5' ⇒ 3'	GAGACAAGCAGAGTGGAAAG
S52	hCas9 5' ⇒ 3'	TCGAGGAAGTCGTGGATAAG
S53	hCas9 5' ⇒ 3'	ACGGCCTGTTTGGTAATC
S54	hCas9 5' ⇒ 3'	GATTCTCCTACAGTCGCTTAC
S55	hCas9 5' ⇒ 3'	GCAGATATACCCGCAGAAAG
S56	hCas9 5' ⇒ 3'	CAACTGCCTGAGAAGTACAA
S57	hCas9 3' ⇒ 5'	CCTTGTACTCGTCCGTAATG
S58	GAPDH exon 1 5' ⇒ 3'	CTATAAATTGAGCCCGCAG
S58b	GAPDH exon 1 5' ⇒ 3'	CCCGGTTTCTATAAATTGAGC
S59	pACTB.2387 5' ⇒ 3'	CGTGACTGTTACCCTCAAA
S60	pACTB.2387 5' ⇒ 3'	CGTTCCGAAAGTTGCCTT
S61	hSpCas9 (PX165) 3' ⇒ 5'	CCCTTAAACCCTTACCTCTTG
S62	hSpCas9 (PX165) 3' ⇒ 5'	GCTGTCCACCAGTTTCTT
S63	hSpCas9 (PX165) 5' ⇒ 3'	GAGCAAGAGGTAAGGGTTTAAG
S64	hSpCas9 (PX165) 5' ⇒ 3'	CATCTACCACCTGAGAAAGAAA
S65	hSpCas9 (PX165) 5' ⇒ 3'	CTGCCTGAGAAGTACAAAGAG
S66	hSpCas9 (PX165) 5' ⇒ 3'	CATACCACGATCTGCTGAAA
S67	hSpCas9 (PX165) 5' ⇒ 3'	GATCGAAGAGGGCATCAAAG
S68	hSpCas9 (PX165) 5' ⇒ 3'	AAGAGCGAGCAGGAAATC
S69	hSpCas9 (PX165) 5' ⇒ 3'	GCCACTATGAGAAGCTGAAG
S70	hSpCas9 (PX165) 5' ⇒ 3'	CATTGTCTGAGTAGGTGTCATT

Table A1, continued:

Primer	notes	sequence
S71	hSpCas9 (PX165) CAG promoter 3' ⇒ 5'	CTCACCTCGACCATGGTAATA
S72	hSpCas9 (PX165) CAG promoter 5' ⇒ 3'	CCTTATGGGACTTTCCTACTTG
S73	IRES-hrGFP/mStraw junction 3' ⇒ 5'	CTTGCTCACCATCATTATCATCG
S74	EMCV IRES (middle) 5' ⇒ 3'	CCACGTTGTGAGTTGGATAG

Table A2: Primers for constructing base expression vector/plasmid

Primer	notes	sequence
g187	bac-transcription blocker, 5' ⇒ 3'	ACATTTCCCCGAAAAGTGCC**ACCTGACGTCGATATCAATAAA**
g188	transcription blocker-base+pA, 3' ⇒ 5'	TTGCGGCCGCTTTTTTCCTT**CGGAATTCCGCCTTAATTAAG**
g189	transcription blocker-base+pA, 5' ⇒ 3'	ATTTCTCTCCTTAATTAAGG**CGGAATTCCGAAGGAAAAAAG**
g190	base+pA-bac, 3' ⇒ 5'	TGGCCTTTTGCTGGCCTTTTGCTC**ACATGTTACCACATTTGTAGAGG**
g191	base+pA-bac, 5' ⇒ 3'	AGTAAAACCTCTACAAATGTGGTA**ACATGTGAGCAAAAGGCCAGCAA**
g192	bac-transcription blocker, 3' ⇒ 5'	AAATAAAGATATTTTATTGATATC**GACGTCAGGTGGCACTTTTCG**
207b	replace EcoRV with AscI, 5' ⇒ 3'	TTGGCGCGCCAATAAAATATCTTTATTTTCATTACATCTGTGTGTTGGTTTTTTGTG
208b	replace EcoRV with AscI, 3' ⇒ 5'	TTGGCGCGCCAAAGGTGGCACTTTTCGGGG

A.2. CHAPTER 2 SPECIFIC

Relevant primer sequences can be found in the Supplementary Information of our publication: Cheng, J. K., Lewis, A. M., Kim, D. S., Dyess, T., & Alper, H. S. (2016). Identifying and retargeting transcriptional hot spots in the human genome. *Biotechnol J*, 11(8), 1100–1109. DOI: 10.1002/biot.201600015.

A.3. CHAPTER 3 SPECIFIC

Table A3: PCR primers for creating expression vectors with homology regions

Primer	notes	sequence
293	bac - TB...SV40pA (AscI) for adding site homology regions, 5' ⇒ 3'	GAGCAAAAGGCCAGCAAAAGG
294	bac - TB...SV40pA (AscI) for adding site homology regions, 3' ⇒ 5'	TACCACATTTGTAGAGGTTTTACTTGCTTT
295	(GRIK1B) 5' homology to TB...SV40pA (AscI) vector, 5' ⇒ 3'	CAAATAGGGGTTCCGCGCA
296	(GRIK1B) 5' homology to TB...SV40pA (AscI) vector, 3' ⇒ 5'	CACACAAAAAACCAACACACAGATGTA
297	(GRIK1B) 3' homology to TB...SV40pA (AscI) vector, 5' ⇒ 3'	GGAGGTGTGGGAGGTTTTTTAAAGC
298	(GRIK1B) 3' homology to TB...SV40pA (AscI) vector, 3' ⇒ 5'	CAGCAACGCGGCCTTTTTACG
538	TB-rpCMV.660...SV40pA for adding site homology regions, 5' ⇒ 3'	AATAAAATATCTTTATTTTCATTACATCTGTGTGTTGG
539	bac - {fcy:fur} - (AscI) for adding site homology regions, 3' ⇒ 5'	GGCGCGCCAAAGGTGG
299	fcy1:fur1 gBlock A 5' ⇒ 3'	CTAAAGTATATATGAGTAAACTTGGTCTGACAGG
300	fcy1:fur1 gBlock A 3' ⇒ 5'	TCCCTCAAAGTTCTCATTGGTGTC
301	fcy1:fur1 gBlock B 5' ⇒ 3'	CAGAAGCAGATTGTGGAGACTGA
302	fcy1:fur1 gBlock B 3' ⇒ 5'	ACTGGGGCCAGATGGTAAG
g380	GRIK1B 3' homology - fcy1:fur1 A 5' ⇒ 3'	gtcaggcatatggtatagaaaatttttccaggatGAGTAATTCATACAAAAG GACTCGCC
g381	GRIK1B 5' homology - fcy1:fur1 B 3' ⇒ 5'	tttcagaaccatggcctgtggtcacttaatGGCGCGCCTGCCAAGTGCATT AGCTGTTTG

Table A4: gBlocks® Gene Fragments (IDT) for creating expression vectors with homology regions

fragment	sequence
AAVS1.T2 gRNA	TGTACAAAAAAGCAGGCTTTAAAGGAACCAATTCAGTCGACTGGATCCGGT ACCAAGGTCGGGCAGGAAGAGGGCCTATTTCCCATGATTCCTTCATATTTG CATATACGATACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTA AACACAAAGATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTT GGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTATCATATGCTTA CCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGG ACGAAACACCGGGGCCACTAGGGACAGGATGTTTTAGAGCTAGAAATAGC AAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTC GGTGCTTTTTTTCTAGACCCAGCTTTCTTGTACAAAGTTGGCATTA

Table A4, continued:

fragment	sequence
GRIK1B 5' homology	CAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTTTGGCGCGC Cattaagtgaccacaggccatggttctgaaaacaaatttaaaacaaatcaaaaggaaattatgaatttgaaattctgcagtgag gcatgctatgtagttaatactaatgtgtaccactaggcctggaccagtataattccaaatgtatgccttgtctcttcctttaggatca tgatttttccccgtggactattcgaaaaacatctaatacactccaaattcagctctacattttgttcactaatttgttcagccttgcaat ccctttcatgcattacaaaatcgctagaccaccacttagactccttctaaccatctgtccatacaaacccttgaaggcaagtcac ccctctttttgtgtctcccacaggacctaacgacacaatatcttgacattgtagttactatgcaacaagtgtgtgtcacattcgtcg ctacttggtggggtaaatacttgggatgagtgaacctttctttacagttcatctagttgttcagtaatatggaaatgaatgaaaact cagaagaattcagtagacatagagctgaggagtacaattccagctgaccaatttgcctataaataagaagtctggtaaatattaa cgtctcaaggctctttgcactttgcccaccaaagctgtaagtttatgactatggcctttactcttttggtttgaagcttttgtgattaat tgttgggttttctgcattgctcagcaggttttggttttgctggtgtcctgtgtgtttctctgatatttacagactgtttctgttaactacca caggtttacacAATAAAATATCTTTATTTTCATTACATCTGTGTGTTGGTTTTTTGT GTG
GRIK1B 3' homology	GGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTAC TCGAGccatgcaaccctgactcagacgtggtggaaaacaattttactttactaaatagtttctggtttggagttggagctctc atgcagcaaggtacaccggttgcctatctttgtcacacgcatcccacagtgtgctgacaggctttcatgctggtaaactccacc atgaagctagagggtaggatgaaacaccatctcaggaacttaagaaaatagcaacggactttctagccggttttgaagatgaa gcaaaaccaatgacctatatggaactcccttgctatatatctaaaatagcacttctatcaagcaataaatacaaattttcctgtggg tagcaaaagataaggtaccttagaaatcgtttacaaatttgcatagatggagatgttggtcttctccttctcagccataactccacg agtcagcagtttaaaggagcgggatctaaatcctttggagattaaattcactgagattcttccaattgccaaccttacctaggact tccacctaattactactctggcaaaccctactgagattcatttcagcctcccagtaacttttcacaatctccctcaattggtaaccat agcaacaataaactgtaggagttggggaaataaaatttggccatagaaatgtaatatattttatacatttcctatgaattcggggat tctgcttttgtagtgggggaaggggagctattttaacctttggaagaactccatttctttgtgatttttttttcagcttcaatgtcag gcatatggtatagaaaatttttccaggatGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTA AAAAGGCCGCGTTGCTG
AAVS1.T2 5' homology	CAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTTTGGCGCGC Cctgctttctctgacctgcattctctcccctgggcctgtgccgctttctgtctgcagcttgtggcctgggtcacctctacggctgg cccagatccttccctgccgcctccttcaggttccgtcttcctccactccctcttcccctttgctctctgctgtgttgctgcccaaggat gctctttccggagcacttccttctcggcgctgcaccacgtgatgtcctctgagcggatcctccccgtgtctgggtcctctccgg gcatctctcctccctcacccaaccccatgccgtcttcactcgctgggttccctttttccttctccttctggggcctgtgccatctctcg tttcttaggatggccttctccgacggatgtctcccttgcgtcccgcctcccctccttgttaggcctgcatcatcaccgtttttctggac aaccccaaagtaccccgtctccctggctttagccacctctccatcctcttgctttctttgcctggacaccccgttctcctgtggatt cgggtcacctctcactcctttcatttgggcagctcccctaccccccttacctctctagtctgtgctagctcttccagccccctgtca tggcatcttccaggggtccgagagctcagctagtcttcttcctccaacccgggcccctatgtccacttcaggacagcatgtttgc tgcctccagggatcctgtgtccccgagctgggaccaccttatattcccagggccggttaatgtggctctggttctgggtactttt atctgtcccctccaccccacagtAATAAAATATCTTTATTTTCATTACATCTGTGTGTTGG TTTTTTGTGTG
AAVS1.T2 3' homology	GGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTAC TCGAGtgacagaaaagccccatccttaggcctcctccttcctagtctcctgatattgggtctaacccccacctcctgttagg cagattccttatctggtgacacaccccatttcctggagccatctctctccttgccagaacctctaaggtttgcttacgatggagc cagagaggatcctgggagggagagcttggcaggggggtgggagggaaggggggatgcgtgacctgcccggttctcagt ggccaccctgcgctaccctctcccagaacctgagctgctctgacgcggccgtctggtgcgtttcactgatcctggtgctgcag cttccttacacttcccaagaggagaagcagtttggaaaaacaaaatcagaataagttggtcctgagttctaactttggctcttcac ctttctagtccccaatttatattgttcctccgtgcgtcagttttacctgtgagataaggccagtagccagccccgtcctggcaggg ctgtggtgaggaggggggtgtccgtgtggaaaactccctttgtgagaatggtgcgtcctaggtgttcaccaggtcgtggccg cctctactccctttctctttctccatccttctttccttaaagagtccccagtgctatctgggacatattcctccgcccagagcagggt cccgcttccctaaggccctgctctgggcttctgggtttgagtccttggcaagcccaggagaggcgctcaggcttccctgtccc ccttcctcgtccaccatctcatgcccctggctctcctgcccctccctacaGAGCAAAAGGCCAGCAAAAG GCCAGGAACCGTAAAAAGGCCGCGTTGCTG

Table A4, continued:

fragment	sequence
Cg.Grik1.A 5' homology	CAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTTTGGCGCGC Cacgggtttaaattccattccatttaaacttttgacttgtgaatttatgactaagaaagaaaagatgggcatgtctatctatggaaat gcatttgttaaaacaagtctgagttcttgtgaaggtccctgttaactcgacaaggtttagagtcatctaggagacacatctcttgac atgtctgtgaagatgtttcctgacaggcttggatgaggagggaaggccctgaaggtgagatgcatcttcctataggctggggc tgtagacttaataaaaaggaaaaagcaatacaaaagtggacaaagttagtagggcacagagttcagtgttctttgtttccttact gcagatacactgtgactgttgccccaagttcctgtccccaagctccccacacctgatggaatatactcttaactatgccactgca aacccgttagcccttaaattgcttctaatcagttatttgattacatcaatgagtaggtaactaatataggctccttattcttcactctttt caggaaaatggtgtgtccattctttggaagaagtgtattcatggcaaagaagctccatttagaccctatttctgtcttcttttgagaa ctggctagaaactctattgctgttgttttctattcctgtgtgtttcatcctaccatctaccttcatggtggagttcaccagcacgaacg cttatcagcagactgtgggatgcaagtgacaaaatgggcaaccggtgtaccttgctgcatgagagctccaactccaaaccag aaactatttagcaaagtgaaattAATAAAATATCTTTATTTTCATTACATCTGTGTGTTGGT TTTTTGTGTGAATCGATAGTACTAACATACGCTCTCCATC
Cg.Grik1.A 3' homology	GGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTAC TCGAGgtgtaaacctgtgataattgacagaaacagtttgtaaacatcagatacacagatgaaccagaggaaaagctgctg agcaatgcagaaaacccaacaactgagcagaagcaattgcttctggattgggaaagtcactgggcaaatactcactagcttct tacttcagagtcaaacccattagtggctaggatggcactcctcagacctctatccactgaattctgagtttccaatcattcacatttt actgagcaactagacaactataaagaaaagttcctgcatctcaagttttgactctagcaaggtacaagagaggcatggaatga atgattacatggttgagcttttatgttgctagttgctgtgagggtcaggtgctctccagagacattatggttccatgaacagacact atacagagtctagcatttcacaaataatgtagaataaagaaaaaaaatgaacaaaaaagaagggatggatttgagagtatatc gcttatttcaaagatactatatagaatccttcaatgaaaaggcaaattttaaatttaggagtatcatggcccaggtataggtatata ctatatgtactaatatagtataatggcattaacttcaaggctcacagtgggaaattagaagctcatgcttccctttgatatattatta attggatggttttctgaattatgccctgaagtcatataatcagcaactctggttaggtgtgcaaagtgttttctgtggcaccattatt gctctacctccaagaagccatggctgtaagtttctggGAGCAAAAGGCCAGCAAAAGGCCAGGAA CCGTAAAAAGGCCGCGTTGCTG
fcy1:fur1 A	GATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTT TAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGGAGTAATTC ATACAAAAGGACTCGCCCCTGCCTTGGGGAATCCCAGGGACCGTCGTTAAA CTCCCACTAACGTAGAACCCAGAGATCGCTGCGTTCCCGCCCCCTCACCCG CCCGCTCTCGTCATCACTGAGGTGGAGAAGAGCATGCGTGAGGCTCCGGTG CCCGTCAGTGGGCAGAGCGCACATCGCCCACAGTCCCCGAGAAGTTGGGG GGAGGGGTCGGCAATTGAACCGGTGCCTAGAGAAGGTGGCGCGGGGTAAA CTGGGAAAGTGATGTCGTGTACTGGCTCCGCCTTTTTCCCGAGGGTGGGGG AGAACCGTATATAAGTGCAGTAGTCGCCGTGAACGTTCTTTTTATGGTCAC AGGAGGCATGGCTTCAAAGTGGGACCAGAAGGGCATGGACATTGCCTATG AGGAGGCTGCTCTGGGCTACAAGGAGGGAGGGGTCCCAATTGGTGGCTGC CTCATCAACAACAAGGATGGCAGTGTCCTGGGCAGGGGCCACAACATGAG GTTCCAGAAGGGCAGTGCCACCCTGCATGGGGAGATCAGCACCCTGGAGA ACTGTGGCAGGCTGGAGGGCAAGGTCTACAAGGACACCACTCTGTACACCA CCCTCAGCCCTTGTGACATGTGCACAGGGGCCATCATCATGTATGGCATTCC CAGGTGTGTGGTGGGAGAGAATGTCAACTTCAAGTCAAAAGGAGAGAAGT ACCTCCAGACCAGGGGCCATGAGGTGGTTGTGGTGGATGATGAGAGGTGC AAGAAGATTATGAAGCAGTTCATTGATGAGAGACCCCAGGACTGGTTTGAG GACATTGGGGAGGCCTCTGAGCCCTTCAAGAATGTGTACCTCCTCCCCCAG ACCAACCAACTCCTGGGACTCTACACCATCATCAGGAACAAGAACACCACC AGGCCAGACTTCATCTTCTACAGTGACAGGATCATCAGGCTCCTGGTGGAG GAGGGCCTCAACCACCTCCCTGTGCAGAAGCAGATTGTGGAGACTGACACC AATGAGAACTTTGAGGGA

Table A4, continued:

fragment	sequence
fcy1:fur1 B	CAGAAGCAGATTGTGGAGACTGACACCAATGAGAACTTTGAGGGAGTGTCT TTCATGGGCAAGATTTGTGGGGTGTCCATTGTGAGGGCTGGGGAGAGCATG GAGCAGGGCCTGAGGGACTGTTGCAGGAGTGTGAGGATTGGCAAGATCCT GATCCAGAGGGATGAGGAGACTGCCCTGCCCAAGCTGTTCTATGAGAAGCT CCCTGAAGACATCTCTGAGAGGTATGTCTTCCTCCTGGACCCCATGCTGGCA ACTGGAGGCTCTGCAATCATGGCCACTGAGGTGCTCATCAAGAGGGGAGTC AAGCCTGAGAGGATCTACTTCCTCAACCTCATCTGCTCAAAGGAGGGCATT GAGAAGTACCATGCTGCCTTCCCTGAAGTGAGGATTGTCACTGGGGCTCTG GACAGGGGCCTGGATGAGAACAAGTACCTGGTCCCTGGCCTGGGAGACTTT GGGGACAGATACTACTGTGTCTAAGCTCGCTTTCTTGCTGTCCAATTTCTAT TAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGA AGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGC AATGATGTATTTAAATTATTTCTGAATATTTTACTAAAAAGGGAATGTGGG AGGTCAGTGCATTTAAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGG GAAAATACACTATATCTTAAACTCCATGAAAGAAGGTGAGGCTGCAAACA GCTAATGCACTTGGCATTACCAATGCTTAATCAGTGAGGCACCTATCTCAG CGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGAT AACTACGATACGGGAGGGCTTACCATCTGGCCCCAGT

Table A5: PCR primers for amplifying target regions

Primer	notes	sequence
Grik1-gDNA-F	inside *GRIK1* 3' homology, 3' ⇒ 5'	GAAATGGAGTTCTTCCAAAGGTTAAAATAGCTCCCCCTTC
Grik1-gDNA-R	inside *GRIK1* 5' homology, 5' ⇒ 3'	GGTTTGAAGCTTTTGTGATTAATTGTTGGGTTTTCTGCAT
593	AAVS1 confirmation_F[144]	TATATTCCCAGGGCCGGTTA
594	AAVS1 confirmation_R[144]	ACAGGAGGTGGGGGTTAGAC
527	*HPRT1* confirmation_F[181]	GATGCTCACCTCTCCCACAC
528	*HPRT1* confirmation_R1[181]	ACATCCATGGGACTTCTGCC
528b	*HPRT1* confirmation_R2	CCGCAACCAGCCTCAATATG
D.641	Cg.Grik1_F confirmation (edited region), 5' ⇒ 3'	AGAATTCAGTGGATAGAGGTCTGAGGAGTG
D.642	Cg.Grik1_R confirmation (edited region), 3' ⇒ 5'	CAGGAAAATGGTGTGTCCATTCTTTGGAAG
D.643	Cg.Hprt1.1_F confirmation (edited region), 5' ⇒ 3'	GCCTTCAATGCCCGGCTTTATATGTTTTTC
D.644	Cg.Hprt1.1_R confirmation (edited region), 3' ⇒ 5'	CTCACAAGGTAAGCGACAATCTATCGAAGG

Table A6: PCR primers for preparing gRNA expression vectors with Gibson Assembly

Primer	notes	sequence
if426	(GRIK1B) gRNA into hCas9 MfeI site 5' ⇒ 3'	GCTTGACCGACAATTTGTACAAAAAAGCAGGCTTTAAAG
if427	(GRIK1B) gRNA into hCas9 MfeI site 3' ⇒ 5'	ATTCTTCATGCAATTTAATGCCAACTTTGTACAAGAAAG

Table A6, continued:

Primer	notes	sequence
if552	(AAVS1.T2) gRNA into hSpCas9 (FZ lab) NotI site, 5' ⇒ 3'	GGGGTTCCTGCGGCCTGTACAAAAAAGCAGGCTTTA
if553	(AAVS1.T2) gRNA into hSpCas9 (FZ lab) NotI site, 3' ⇒ 5'	TGCTGGGGAGCGGCCTAATGCCAACTTTGTACAAGA
if558	PX165 - EcoRI - hSpCas9(GC), 5' ⇒ 3'	TCAGCGAGCTCTAGGAATTCTCACACCTTCCTCTTCTTC TTGGGG
if559	PX165 - AgeI - hSpCas9(GC), 3' ⇒ 5'	TTTTTTCAGGTTGGACCGGTATGGGCGGTAGGCGTGTAC
gR.586	GRIK1A gRNA	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACAC **CGCTATTTTAGATATATAGCA**
gR.587	GRIK1A gRNA	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAA **CTGCTATATATCTAAAATAGC**
gR.674	GRIK1C gRNA	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACAC **CGCAGGTTTACACCCTACGAG**
gR.675	GRIK1C gRNA	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAA **CCTCGTAGGGTGTAAACCTGC**
gR.595	AAVS1.T1 gRNA	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACAC **CGGTCCCCTCCACCCCACAGT**
gR.596	AAVS1.T1 gRNA	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAA **CACTGTGGGGTGGAGGGGACC**
g525	HPRT1 gRNA_F	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACAC **CGAAGTAATTCACTTACAGTC**
g526	HPRT1 gRNA_R	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAA **CGACTGTAAGTGAATTACTTC**
gR.Cg540	CHO Grik1 gRNA_F1	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACAC **CGTTTTCCACCACGTCTGAGT**
gR.Cg541	CHO Grik1 gRNA_R1	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAA **CACTCAGACGTGGTGGAAAAC**
gR.Cg548	CHO Grik1 gRNA_F2	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACAC **CGTGGGGATTGTACCACTCGT**
gR.Cg549	CHO Grik1 gRNA_R2	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAA **CACGAGTGGTACAATCCCCAC**
gR.Cs534	CHO Hprt1 gRNA_F1	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACAC **CGTTTGTGTCATTAGTGAAAC**
gR.Cs535	CHO Hprt1 gRNA_R1	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAA **CGTTTCACTAATGACACAAAC**
gR.Cs536	CHO Hprt1 gRNA_F2	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACAC **CGGGGTTGTACCGCTTGACCA**
gR.Cs537	CHO Hprt1 gRNA_R2	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAA **CTGGTCAAGCGGTACAACCCC**

Table A7: Quantitative PCR primers and probes for target regions

Primer	notes	sequence
PT-RPPH1-F	*RPPH1* housekeeping gene	AGTGAGTTCAATGGCTGAGG
PT-RPPH1-R	*RPPH1* housekeeping gene	GGCGGAGGAGAGTAGTCT
Q-RPPH1 probe	*RPPH1* housekeeping gene probe probe with FAM/ZEN/IBFQ	TTGGGTTATGAGGTCCCCTGCG

Table A7, continued:

Primer	notes	sequence
Q-GRIK1-F1	GRIK1 edited region (GRIK1B)	GGTTTTCTGCATTGCTCAGC
Q-GRIK1-R1	GRIK1 edited region (GRIK1B)	TCCAACTCCAAACCAGAAACT
Q-GRIK1-NHEJ insensitive probe	NHEJ-insensitive probe with HEX/ZEN/IBFQ (GRIK1B)	TGCTGGTGTCCTGTGTGTTTCTCT
Q-GRIK1-NHEJ sensitive probe	NHEJ-sensitive probe with FAM/ZEN/IBFQ (GRIK1B)	CGAGTGGTATAACCCCCACCCATGCA
Q-AAVS1-F1	AAVS1 edited region (AAVS1.T2)	CTTCAGGACAGCATGTTTGC
Q-AAVS1-R1	AAVS1 edited region (AAVS1.T2)	AGACCCAATATCAGGAGACTAGG
Q-AAVS1-NHEJ insensitive probe	NHEJ-insensitive probe with HEX/ZEN/IBFQ (AAVS1.T2)	ACCCAGAACCAGAGCCACATTAACC
Q-AAVS1-NHEJ sensitive probe	NHEJ-sensitive probe with FAM/ZEN/IBFQ (AAVS1.T2)	AGGATTGGTGACAGAAAAGCCCCATCC
Q-Cg.Actb-F1	*C. griseus Actb* housekeeping gene	GCATCCACGAAACTACATTCAA
Q-Cg.Actb-R1	*C. griseus Actb* housekeeping gene	AGCAATGCCTGGGTACAT
Q-Cg.Actb probe	*C. griseus Actb* housekeeping gene probe with FAM/ZEN/IBFQ	AACACAGTGCTGTCTGGTGGTACC
Q-Cg.Grik1-F1	*C. griseus Grik1* edited region	CTGATGTTTACAAACTGTTTCTGTC
Q-Cg.Grik1-R1	*C. griseus Grik1* edited region	AACCGGTGTACCTTGCT
Q-Cg.Grik1-NHEJ insensitive probe	NHEJ-insensitive probe with HEX/ZEN/IBFQ (Cg.Grik1)	TGGTTTGGAGTTGGAGCTCTCATGC

Table A7, continued:

Primer	notes	sequence
Q-Cg.Grik1-NHEJ sensitive probe	NHEJ-sensitive probe with FAM/ZEN/IBFQ (Cg.Grik1)	ctacgagtggtacaatccccaccc

A.4. CHAPTER 4 SPECIFIC

Relevant primer sequences can be found in the Supplementary Information of our publication: Cheng, J., & Alper, H. S. (2016). Transcriptomics-guided design of synthetic promoters for a mammalian system. *ACS Synth Biol*. Article ASAP. Publication Date (Web): June 7, 2016. DOI: 10.1021/acssynbio.6b00075.

A.5. CHAPTER 5 SPECIFIC

Table A8: Primers for constructing UCP core promoter and CMV enhancer variants

Primer	notes	sequence
a61	Assembly PCR primer for UCP (20nt match to a62) 5' ⇒ 3'	CTAGCTAGCGCGCGCCTATATAAGTTTGTTTCGTTTAGTGAACCGTCAGA
a62	Assembly PCR primer for UCP (20nt match to a61, 17nt match to a63) 3' ⇒ 5'	TCGGCGTCTCCAGAGAATCTGACGGTTCACTAAACGA
a63	Assembly PCR primer for UCP (17nt match to a62) 5' ⇒ 3'	TTCTCTGGAGACGCCGAGCCGAGCGGTCAGACCTCCATAGAAGCGGCCGCAAAAGGAAAA
f64	Flanking primer for NheI-UCP-NotI 5'⟹3'	CTAGCTAGCGCGCGCCTATA
f65	Flanking primer for NheI-UCP-NotI 3'⟹5'	TTTTCCTTTTGCGGCCGCTT
72	AflII-CMV enhancer 5' ⇒ 3'	AGACCCCTTAAGCGCGTTACATAACTTACGGTAAATGGC
73	NheI-CMV enhancer 3' ⇒ 5'	AACTAGCTAGCCAAAACAAACTCCCATTGACGTCAA
74	AflII-CMV enhancer "long" 5' ⇒ 3'	AGACCCCTTAAGTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGC
75b	NheI-CMV enhancer full 3' ⇒ 5'	AACTAGCTAGCCCCACCGTACACGCCTACCG

Table A9: Primers for multiple core promoter work

Primer	notes	sequence
193	NotI-core pCMV, 5' ⇒ 3'	TTTTCCTTTTGCGGCCGCAGGTCTATATAAGCAGAGCTCGTTTA
194	NheI-core pCMV, 3' ⇒ 5'	CTAGCTAGCGATCTGACGGTTCACTAAACGAGCTCTGCTTATATAGACCT

Table A9, continued:

Primer	notes	sequence
197	EcoRI-cpCMV no TSS-NotI, 5' ⇒ 3'	CGGAATTCAGGTCTATATAAGCAGAGCTCGTTTAGTGAACCGGCGGCCGCAAAAGGAAAA
198	NotI-cpCMV no TSS-EcoRI, 3' ⇒ 5'	TTTTCCTTTTGCGGCCGCCGGTTCACTAAACGAGCTCTGCTTATATAGACCTGAATTCCG
199	EcoRI-cpCMV, 5' ⇒ 3'	CCGGAATTCAGGTCTATATAAGCAGAGCTCGTTTAGTGAAC
200	NotI-cpCMV, 3' ⇒ 5'	TTTTCCTTTTGCGGCCGCGATCTGACGGTTCACTAAACGAGCTCTGCTTA
201c	PacI-BamHI-cpCMV, 5' ⇒ 3'	CCTTAATTAAGGCGCGGATCCAGGTCTATATAAGCAGAGCTCG
202b	EcoRI-cpCMV no TSS, 3' ⇒ 5'	CCGGAATTCCGGTTCACTAAACGAGCTCTGCTTATATAGACCT
203b	EcoRI-cpCMV, 3' ⇒ 5'	CCGGAATTCGATCTGACGGTTCACTAAACGAGCTCTGCTTATATAGACCT

Table A10: Primers for hybrid promoters

Primer	notes	sequence
243	NotI-pEEF1A1.1356, 5' ⇒ 3'	ATAAGAATGCGGCCGCGAGTAATTCATACAAAAGGACTCGCCCCTGC
310b	NotI-pEEF1A1.1048 5' ⇒ 3'	ATAAGAATGCGGCCGCGGGGGAGAACCGTATATAAGTGCAGTAGTC
306-1	pEEF1A1 (exon 2) 3' ⇒ 5'	TTTGGCTTTTAGGGGTAGTTTTCACGACAC
307	NotI-pCLUAP1.2000 5' ⇒ 3'	ATAAGAATGCGGCCGCATCATATATCTGATAAGGGTCTCATAGGCA
308	NotI-pCLUAP1.500 5' ⇒ 3'	ATAAGAATGCGGCCGCCTGCAACCTCTGCCTGCC
309	NheI-pCLUAP1 3' ⇒ 5'	CTAGCTAGCAACCAATGGACCCTTGGACG
311	NotI-pTPT.2000 5' ⇒ 3'	ATAAGAATGCGGCCGCGGGCAACAGAGCAAGACCCCATCTG
g311	GA pTPT.2000 5' ⇒ 3'	ATTAAGGCGGAATTCCGAAGGAAAAAAGCGGCCGCGGGCAACAGAGCAAGACCCCATCTG
g311b	GA CMVe-pTPT.2000 5' ⇒ 3'	AAATGGGCGGTAGGCGTGTACGGTGGGGCGGCCGCGGGCAACAGAG
312	NotI-pTPT.500 5' ⇒ 3'	ATAAGAATGCGGCCGCCGGACTCAGCGGTGCCCC
313	NheI-pTPT 3' ⇒ 5'	CTAGCTAGCGCGCCTCCGGAAGCGACG
g313	GA pTPT 3' ⇒ 5'	TAGCCTCAGGCCCAGCAGCAGCAGCAGCAGCAGCATGCTAGCGCGCCTCCGGAAGCGACG
g313b	GA promoter 3' ⇒ 5'	CAGCAGCAGCAGCAGCAGCAGCATGCTAGC
g314	GA pTUBA1B.2000 5' ⇒ 3'	GAAGGAAAAAAGCGGCCGCACAATATACCCTTTTTTTTTTTTGGAGATGGAGTTTCGGTC
g314b	GA CMVe-pTUBA1B.2000 5' ⇒ 3'	GGTAGGCGTGTACGGTGGGGCGGCCGCACAATATACCCTTTTTTTTTTTTGGAGATGGAG
315	NotI-pTUBA1B.500 5' ⇒ 3'	ATAAGAATGCGGCCGCGCTTGGGCACTACTCTTCGAATGACTGAAATACTATTTGC
g315	GA pTUBA1B.500 5' ⇒ 3'	GAAGGAAAAAAGCGGCCGCGCTTGGGCACTACTCTTCGAATGACTGAAATACTATTTGC
g315b	GA CMVe-pTUBA1B.500 5' ⇒ 3'	GCAAATGGGCGGTAGGCGTGTACGGTGGGGCGGCCGCGCTTGGGCACTACTCTTCG
g315c	GA CMVe-pTUBA1B.500i 5' ⇒ 3'	CAAATGGGCGGTAGGCGTGTACGGTGGGGCGGCCGCGGGGTGGGGTCTG

Table A10, continued:

Primer	notes	sequence
316	NheI-pTUBA1B 3' ⇒ 5'	CTAGCTAGCTCCGGCCCCGCGCGC
g316	GA pTUBA1B 3' ⇒ 5'	CTGTAGCCTCAGGCCCAGCAGCAGCAGCAGCAGCATGCTAGCTCCGGCCCCGCGCGC
317b	NotI-pGGA1.2000 5' ⇒ 3'	ATAAGAATGCGGCCGCAAATTAGCCAGGCATGATGGCGCATGCCTG
317-1	pGGA1 5' ⇒ 3'	AGTCTCTCAAAGCAACACATTTGCTCATGG
318b	NotI-pGGA1.500 5' ⇒ 3'	ATAAGAATGCGGCCGCGCCTTCATCCTACTTTTTCTCCCTGAATATCTTTTCGCACC
319b	NheI-pGGA1 3' ⇒ 5'	CTAGCTAGCTTAAAAGGGCGATAAGCTACATCCTCATGTACCTTGGCC
319-1	pGGA1 3' ⇒ 5'	GGACACTCACTGATTCGCGCCTCC
320	NotI-pLAIR1.2000 5' ⇒ 3'	ATAAGAATGCGGCCGCATGCCATTGCACTCCAGC
320b	NotI-pLAIR1.2000 5' ⇒ 3'	ATAAGAATGCGGCCGCATGCCATTGCACTCCAGCCTGGGTGACAGAG
321	NotI-pLAIR1.500 5' ⇒ 3'	ATAAGAATGCGGCCGCAAAAATTTCTTTAAATTGGCCTTTGG
321b	NotI-pLAIR1.500 5' ⇒ 3'	ATAAGAATGCGGCCGCAAAAATTTCTTTAAATTGGCCTTTGGAAATTTACCAGCAGTGTG
322	NheI-pLAIR1 3' ⇒ 5'	CTAGCTAGCAGACATAGCGGGTGTCATAGATG
322b	NheI-pLAIR1 3' ⇒ 5'	CTAGCTAGCAGACATAGCGGGTGTCATAGATGTGAAAAAGCTTCTGCTATACCAG
322c	NheI-pLAIR1.L 3' ⇒ 5'	CTAGCTAGCCTTCTGTCGCGGATGCAACCCTGGAAGGAAG
323b	NotI-pUBC-1778 5' ⇒ 3'	ATAAGAATGCGGCCGCCCCTGTTGGCATCAAGTAGGACC
323c	NotI-pUBC-1778 5' ⇒ 3'	ATAAGAATGCGGCCGCCCCTGTTGG
323-1	pUBC 5' ⇒ 3'	GGAGGCCATTTTCTTCCCTGTCACC
323-2	pUBC 5' ⇒ 3'	ggaggccattttcttccctgtcacctcagtg
324	NotI-pUBC-1310 5' ⇒ 3'	ATAAGAATGCGGCCGCGCGGGGTTCGCGACC
324b	NotI-pUBC-1310 5' ⇒ 3'	ATAAGAATGCGGCCGCGCGGGGTTC
325	NheI-pUBC 3' ⇒ 5'	CTAGCTAGCTGTCTAACAAAAAAGCCAAAAACG
325b	NheI-pUBC 3' ⇒ 5'	CTAGCTAGCTGTctaacaaaaaagccaaaaacggcc
325-1	pUBC 3' ⇒ 5'	tcaagcgcaggaccaagtgcagagtggactc
326-1	pUBC 3' ⇒ 5'	TCAAGCGCAGGACCAAGTGCAGAG
326	NotI-pVIM.2000 5' ⇒ 3'	ATAAGAATGCGGCCGCTTTAAGTCACCCAATTCAACTGAC
327	NotI-pVIM.500 5' ⇒ 3'	ATAAGAATGCGGCCGCGTAACCTGCAGTACCCCCTGC
328	NheI-pVIM 3' ⇒ 5'	CTAGCTAGCGATTAATGAGTTCAAATGAAAGGCAATTATG
332b	NotI-pF2R.2000 5' ⇒ 3'	ATAAGAATGCGGCCGCGAAGGAAGGAAGGAAGAGAGGGAGGGAGGAAGGAAGGAAG
332-1	pF2R-3000 5' ⇒ 3'	CAAGCCTCAAAAAATACAACTGGCAGAAATATTCCTAGAAGG
332-1b	pF2R-3070 5' ⇒ 3'	caagcctcaaaaaatacaactggcagaaatattcctagaaggaatcc
333b	NotI-pF2R.500 5' ⇒ 3'	ATAAGAATGCGGCCGCACTCAAGGGCCCTTTCTCATTTAGGGGCAAC
334b	NheI-pF2R 3' ⇒ 5'	CTAGCTAGCAGCGAGGAAGGGCGCCCTCC

Table A10, continued:

Primer	notes	sequence
334c	NheI-pF2R.L 3' ⇒ 5'	CTAGCTAGCTGTCCCGGGCTCTGCGCGG
334-1	pF2R-3070 3' ⇒ 5'	agggtgcccacgggtaagatcagggtccaag
362	NotI-pEIF4A1.776 5' ⇒ 3'	ATAAGAATGCGGCCGCGTGGTGGTCTTCCTTAAGGGGCTTCAAATTAGTG
363	NotI-pEIF4A1.500 5' ⇒ 3'	ATAAGAATGCGGCCGCCTGGGGAGGGGACCAACCAGGATTC
364	NheI-pEIF4A1 3' ⇒ 5'	CTAGCTAGCCGCCTGCCGGCGCTCAG
364b	NheI-pEIF4A1ex 3' ⇒ 5'	CTAGCTAGCGATCCTTAGAAACTAGGGCGGAGTGCCCGCC
365	NotI-pPGK1.2000 5' ⇒ 3'	ATAAGAATGCGGCCGCATATATTTTCCCAGCACTGTGAATAAAAGTCAGTTGAATGAG
366	NotI-pPGK1.1500 5' ⇒ 3'	ATAAGAATGCGGCCGCACAGCCCTTTCCCCTTCTTGCTG
367	NotI-pPGK1.1000 5' ⇒ 3'	ATAAGAATGCGGCCGCGGCAGCGGACAAGGTGAACCC
368	NotI-pPGK1.500 5' ⇒ 3'	ATAAGAATGCGGCCGCGGACCTGGGCCTCTTCCAACTTC
368b	NotI-pPGK1.500 5' ⇒ 3'	ATAAGAATGCGGCCGCGGACCTGGGCCTCTTCCAACTTCTGAGAGG
369	NheI-pPGK1 3' ⇒ 5'	CTAGCTAGCACCCCGCCTCCCGCA
369b	NheI-pPGK1.L 3' ⇒ 5'	CTAGCTAGCTTTGGAAATACAGCTGGGGAGAGAGGTCGG
370	NotI-pGAPDH.2000 5' ⇒ 3'	ATAAGAATGCGGCCGCCAAGGGTTGCTTTCTGCCGTG
371	NotI-pGAPDH.1500 5' ⇒ 3'	ATAAGAATGCGGCCGCGAATAGCTGAGTCAGAGGTGGGGC
372	NotI-pGAPDH.1000 5' ⇒ 3'	ATAAGAATGCGGCCGCGTGCCCAAGACCTCTTTTCCCAC
373	NotI-pGAPDH.500 5' ⇒ 3'	ATAAGAATGCGGCCGCGATGGGGAGGGGGAAGTGG
374	NheI-pGAPDH 3' ⇒ 5'	CTAGCTAGCCTCGGACCCCGCCTCC
374b	NheI-pGAPDH (intron) 3' ⇒ 5'	CTAGCTAGCGGTGTCTGAGCGATGTGGCTC
375	NotI-pPGK1 5' ⇒ 3'	ATAAGAATGCGGCCGCGGGTTGGGGTTGCGCCTTTTC
if498	EcoRI - EEF1A1 enhancer (308bp), 5' ⇒ 3'	AATTAAGGCGGAATTCGAGTAATTCATACAAAAGGACTCGC
if498b	EcoRI - EEF1A1 enhancer (156bp), 5' ⇒ 3'	AATTAAGGCGGAATTCGGCTCCGGTGCCCGTCAG
if499	pEIF4A1.500 - NotI - EEF1A1 enhancer (308), 3' ⇒ 5'	CCTCCCCAGGCGGCCGCACCCTCGGGAAAAAGGC
if499b	pEIF4A1.500 - NotI - EEF1A1 enhancer (156), 3' ⇒ 5'	CCTCCCCAGGCGGCCGCACCCTCGGGAAAAAGGCGGAG
if500	pGAPDH.500.i1 - NotI - EEF1A1 enhancer, 3' ⇒ 5'	CTCCCCATCGCGGCCGCACCCTCGGGAAAAAGGC
if503	pUBC.1310 - NotI - EEF1A1 enhancer (308), 3' ⇒ 5'	GAACCCCGCGCGGCCGCACCCTCGGGAAAAAGGC
if503b	pUBC.1310 - NotI - EEF1A1 enhancer (156), 3' ⇒ 5'	GAACCCCGCGCGGCCGCACCCTCGGGAAAAAGGCGGAG
if507	EcoRI - EIF4A1 enhancer (450bp), 5' ⇒ 3'	AATTAAGGCGGAATTCCTGGGGAGGGGACCAACCAGG
if508	pEEF1A1.1048 - NotI - EIF4A1 enhancer (450bp), 3' ⇒ 5'	TTCTCCCCCGCGGCCGCGTGAAGGCCCGCCCCGC

Table A10, continued:

Primer	notes	sequence
if519	EcoRI - ACTBe.182, 5' ⇒ 3'	AATTAAGGCGGAATTCAGGCGGCCAAC
if520	pEEF1A1.1048 - NotI - ACTBe.182, 3' ⇒ 5'	TTCTCCCCCGCGGCCGCTC
if521	NotI - pACTB.2387, 5' ⇒ 3'	AGGAAAAAAGCGGCCGCGTTCCATGTCCTTATATGGACTCATCTTTGCCTATTGCGACAC
if522	NheI - pACTB.2387, 3' ⇒ 5'	GCAGCAGCATGCTAGCGGTGAGCTGCGAGAATAGCCGGGCGCG
if550	EcoRI - LMO1 enhancer (553bp), 5' ⇒ 3'	AATTAAGGCGGAATTCGTAGGGGTTGGAGTTCAG
if551	pEEF1A1.1048 - NotI - LMO1 enhancer (553bp), 3' ⇒ 5'	TTCTCCCCCGCGGCCGCAGGGGCTCTGTAGTCTCC
if554	EcoRI - MMse176.1751, 5' ⇒ 3'	AATTAAGGCGGAATTCGAGTTAAAAAGAGGTGATTTACAGTGCTATTTGAGAAGG
if555	pEEF1A1.1048 - NotI - MMse176.1751, 3' ⇒ 5'	TTCTCCCCCGCGGCCGCCTTCTCTCTGATATTAAACGGCTTCCAAATGC
if556	EcoRI - MMse117.3270, 5' ⇒ 3'	AATTAAGGCGGAATTCCCTCAATTTCTCTAATCTCTTTCATAGGCTCCTG
if557	pEEF1A1.1048 - NotI - MMse117.3270, 3' ⇒ 5'	TTCTCCCCCGCGGCCGCCCATCTCTACTAAAAATACAAAAATTAGCCGGGC
560	MMse176.1751, 5' ⇒ 3'	GAGTTAAAAAGAGGTGATTTACAGTGCTATTTGAGAAGGGG
561	MMse176.1751, 3' ⇒ 5'	CTTCTCTCTGATATTAAACGGCTTCCAAATGCAAGC
562	MMse117.3270, 5' ⇒ 3'	CCTCAATTTCTCTAATCTCTTTCATAGGCTCCTGCACTG
563	MMse117.3270, 3' ⇒ 5'	CCATCTCTACTAAAAATACAAAAATTAGCCGGGCATGGTG
if564	EcoRI - pACTB.1173, 5' ⇒ 3'	AATTAAGGCGGAATTCAGGCGGCCAACGCCAAAAC
if565	NheI - pACTB.1173, 3' ⇒ 5'	GCAGCAGCATGCTAGCGGTGAG
if566	pRPL41.398 - NotI - ACTBe.182, 3' ⇒ 5'	AGCACCTATGCGGCCGCTCGAGCCATAAAAGG
if567	pGAPDH.500 - NotI - ACTBe.182, 3' ⇒ 5'	CTCCCCATCGCGGCCGCTCGAGCCATAAAAGG
if568	pEEF1A1.350 - NotI - ACTBe.182, 3' ⇒ 5'	GAATTACTCGCGGCCGCTCGAGCCATAAAAGG
if569	NotI - pRPLP2.479, 5' ⇒ 3'	AGGAAAAAAGCGGCCGCTCACTTCCGGAACTGCTGCCCTTCGCCTTTG
if570	NheI - pRPLP2.479, 3' ⇒ 5'	GCAGCAGCATGCTAGCCCTGAGGCGGGCGGAGAGGACGCGAC
if571	NotI - pEEF1A1.350, 5' ⇒ 3'	AGGAAAAAAGCGGCCGCGAGTAATTCATACAAAAGGACTCGC
if572	NheI - pEEF1A1.350, 3' ⇒ 5'	GCAGCAGCATGCTAGCAAAAAGAACGTTCACGGCG
if588	pRPLP2.479 - NotI - ACTBe.182, 3' ⇒ 5'	cggaagtgaGCGGCCGCTCGAGCCATAAAAGG
if676	pEEF1A1.204/1204, 5' ⇒ 3'	AGGAAAAAAGCGGCCGCGGCTCCGGTGCCCGTC

Table A11: Primers for other hybrid promoters with introns

Primer	notes	sequence
i264	iS1, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgttactaa ctttttttcttccatttca
i265	iS1, 3' ⇒ 5'	TTTGGCTTTTAGGGGTAGTTTTCACGACACctgaaatggaag aaaaaaagttagtaacCT
g224	EcoRI-NotI-reference cpCMV-iA, 5' ⇒ 3'	GGCGGAATTCCGAAGGAAAAAAGCGGCCGCAGGTCT ATATAAGCAGAGCTCGTTTAGTGA
g225	iA-hrGFP junction, 3' ⇒ 5'	AGGCCGGTGTTCTTCAGGATCTGCTTGCTCACCATCGT GTCAAGGACGGTGActg
g226	iA-SEAP junction, 3' ⇒ 5'	GCTGTAGCCTCAGGCCCAGCAGCAGCAGCAGCAGCA GCATCGTGTCAAGGACGGTGActg
g458	core CMV, 5' ⇒ 3'	AGGTCTATATAAGCAGAGCTCGTTTAGTGAACCG
g430	SEAP towards NheI site 3' ⇒ 5'	CAGCAGCAGCAGCAGCAGCAG
g430b	SEAP towards NheI site 3' ⇒ 5'	CTCAGGCCCAGCAGCAGCAGC
g431	rpCMV.660.Gi1 5' ⇒ 3'	TATATAAGCAGAGCTCGTTTAGTGAACCGGGCTGGGA CTGGCTGAGCCTG
g432	pF2R.500.Gi1 5' ⇒ 3'	GGGGAGGGGGCGCCGAGCGGCTCCAGCGCGGCTGGG ACTGGCTGAGCCTG
g433	pGGA1.500.Gi1 5' ⇒ 3'	AAACTGATTTATTTTCGTCATTTTCACAGGGCTGGGAC TGGCTGAGCCTG
g434	pLAIR1.500.Gi1 5' ⇒ 3'	TCAGTTTTGCTCCGTTCCTGACCCTGGTAGGCTGGGAC TGGCTGAGCCTG
g485	pEIF4A1.500.Gi1, 5' ⇒ 3'	TCCAATGGTGCCTGCGGGCCGGAGCGACTAGGCTGGG ACTGGCTGAGCCTG
g506	pEEF1A1.204.Gi1 5' ⇒ 3'	GTGCAGTAGTCGCCGTGAACGTTCTTTTTGGCTGGGA CTGGCTGAGCCTG
645	MST1L.i9, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtgagtc cctggtgctcccggccccgccagGTGTCGTGAAAACTACCCCTAA AAGCCAAA
646	AQP12A.i1, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtgggtg cgggtgcagctcgggccctgctgcctggagGTGTCGTGAAAACTACCC CTAAAAGCCAAA
647	GUCY2EP.i10, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtgaccc ctgatcacgggccgcctgagtgcgttggctgatgaagGTGTCGTGAAAACT ACCCCTAAAAGCCAAA
648	NDOR1.i12, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtgtgtat gctcaggggctgggaaaggaggggagggagctccggctcagGTGTCGTGAA AACTACCCCTAAAAGCCAAA
649	SAMD14.i8, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtgggtt gggggtgaactctcctcagtggcatgcacctttttctgtctagGTGTCGTGAAA ACTACCCCTAAAAGCCAAA
650	LCA10.i2, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtagggc acacaggcctgtctcctcgaccctcacctgccagtcctggccagGTGTCGTGAA AACTACCCCTAAAAGCCAAA
651	PDHB.i6, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtggcct tcagcctgtgcctatagtcttcagagggcccaatggtgcctcagcagGTGTCGTG AAAACTACCCCTAAAAGCCAAA
652	SYT14P1.i1, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgttaatttt tttttaaaagttcatttttcatttgcttctgcttttgcctgatccaaagagGTGTCGTGA AAACTACCCCTAAAAGCCAAA

Table A11, continued:

Primer	notes	sequence
653	SIGLEC6.i1, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtgagtg ggccaggggagaggtgccgtggggctgggccgagctgaccctcatgtctccatagGT GTCGTGAAAACTACCCCTAAAAGCCAAA
654	SIGLEC17P.i1, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtgagtg gccctggggagaggggccgtggggatgagcccatctgaccctcatgtctccacagGT GTCGTGAAAACTACCCCTAAAAGCCAAA
655	ESRP2.i6, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtgagtg tgggagcaggggctgggggtgacaacagctgaatagGTGTCGTGAAAAC TACCCCTAAAAGCCAAA
656	HNRNPH1.i6, 5' ⇒ 3'	CTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtaaggt aagaattgaatttctcagttgaaggatgcttacactcttgtccatctagGTGTCGTGA AAACTACCCCTAAAAGCCAAA

Table A12: gBlocks® for hybrid promoters

fragment	sequence
ACTBe.182	AATTAAGGCGGAATTCAGGCGGCCAACGCCAAAACTCTCCCTCCTCCTCTT CCTCAATCTCGCTCTCGCTCTTTTTTTTTTTCGCAAAAGGAGGGGAGAGGGG GTAAAAAAATGCTGCACTGTGCGGCGAAGCCGGTGAGTGAGCGGCGCGGG GCCAATCAGCGTGCGCCGTTCCGAAAGTTGCCTTTTATGGCTCGAGCGGCC GCGGGGGAGAA
RPL41-regs (5' regulatory element from *RPL41*)	AGGAAAAAAGCGGCCGCATAGGTGCTGACGTTTAAATAACACAGCGTCCTC ATACTAAATCTGGGGGGGAACTGGTAACTCGAAAACCAAATACTCGGTCTT CCGAAAGAACTAACTCAACCTACCCTTCTACAAGAGGGTCCGAAAACCACT GTTACGCCCATTGGGTAGCCCCGCCCTTGGGGGGGGCAAAGGGCGTGAAA GCGGAAGTGACGACACCCGGCGCTCCATTAAATAGCCGTAGACGGAACTTC GCCTTTCTCTCGGCCTTAGCGCCATTTTTTTGGGTGAGTGTTTTTTGGTTCCT GCGTTGGGATTCCGTGTACAATCCATAGACATCTGACCTCGGCACTTAGCA TCATCACAGCAAACTAACTGTAGCCTTTCTCTCTTTCCCTGTAGAAACCTCT GCGCCGCTAGCATGCTGCTGCCCCCTGGGCAGCCTGCACGAGTGGGTGTAA ACCGCTAGCTTGTTGCACCGTGGAGGCCACAGGAGCAGAAACATGGAATG CCAGACGCTGGGGATGCTGGTACAAGTTGTGGGACTGCATGCTACTGTCTA GAGCTTGTCTCAATGGATCTAGAACTTCATCGCCCTCTGATCGCCGATCACC TCTGAGACCCACCTTGCTCATAAACAAAATGCCCATGTTGGTCCTCTGCCCT GGACCTGTGACATTCTGGACTATTTCTGTGTTTATTTGTGGCCGAGTGTAAC AACCATATAATAAAtcacctcttccgctgttttagctgaagaattaaatCAtcttgtctattaTGTTTTTTA TGGTTCCATCGGGTGGGGGTTTTCTGTCATTAGAGTTTGCCCTGTCACTACC TGTGCTATGGAGGGTATGAGCAAAAGGCCAGCAAAAGGCCAGGAACC
LMO1e.553[281]	GTAGGGGTTGGAGTTCAGCCTGTTTCCCCTCCAATGTTGTTCCCCCCACATC CTGAGACTTAGGGGTGACCCTGGGTTGAGTGGACTGGTTTATTCTGCTGGG CCCAGCGCATGCATCTGAGTGTGTGCCCAGGCGTGCGTGTCGGCGCAAACA TCATCCATTGTGAAATATCAGTGTTTTCATGGGTGAGTAGTAATTACTGGGT AATGCTTTAAAACCTTTCCTGAAGGAGCGCAAAGCCATTTTTTTCTAAAGTC AGGAGTACATTAAAAGGATTACCATGTAGATTTGATTTTTAGATAACACTA AAATGGATCCCAAATGGACTTCAGCAAAGGGATGCTATCTCCTTAATGGAA AGTGCATGGCCCGAGGCTCAGGTCCCAGAGCCAGGCTGgggaaggagggagggaag aggtgtctgcaggggggcaggctggcagattgggtgggggctaggtgggaatggggaaggcagagcaggagggagg gCCTGGACCCTGTGGGGAGCTTATCCCTCCATCTGGGGAGCAGGAGACTAC AGAGCCCCT

A.6. CHAPTER 6 SPECIFIC

Table A13: Primers for creating expression vectors

Primer	notes	sequence
if492	NotI - promoter, 5' ⇒ 3'	AGGAAAAAAGCGGCCGC
if493	hrGFP - NheI - pEIF4A1.636, 3' ⇒ 5'	TGCTCACCATGCTAGCGATCCTTAGAAACTAGGG
494	hrGFP - AscI - bac elements, 5' ⇒ 3'	GAGTGGGTGTAAGGCGCGCCGAGCAAAAGGCCAGCAAAAG
495	hrGFP - AscI - bac elements, 3' ⇒ 5'	GGCCTTTTGCTCGGCGCGCCTTACACCCACTCGTGCAG
682	SEAP, 5' ⇒ 3'	ATGCTGCTGCTGCTGCTGCTGCTGG
683	SEAP, 3' ⇒ 5'	TCATGTCTGCTCGAAGCGGCCGGC
g684	SEAP (30bp) - NheI - pEIF4A1.636, 3' ⇒ 5'	CAGGCCCAGCAGCAGCAGCAGCAGCAGCATGCTAGCGATCCTTAGAAACTAG
g685	SEAP (30bp) - bac - TB - pEIF4A1.636, 5' ⇒ 3'	GGGGCGGCCGGCCGCTTCGAGCAGACATGAGAGCAAAAGGCCAGCAAAAG
284b	NotI-pCMV, 5' ⇒ 3'	AAGGAAAAAAGCGGCCGC
if691	hrGFP - NheI - rpCMV.660, 3' ⇒ 5'	TGCTCACCATGCTAGCGTCTTCTATGGAGGTC
if681	hrGFP - NheI - EEF1A1 exon 2, 3' ⇒ 5'	GCTCACCATGCTAGCTTTGGCTTTTAGGGGTAGTTTTC
g684	SEAP (30bp) - NheI - pEIF4A1.636, 3' ⇒ 5'	CAGGCCCAGCAGCAGCAGCAGCAGCAGCATGCTAGCGATCCTTAGAAACTAG
g685	SEAP (30bp) - bac - TB - pEIF4A1.636, 5' ⇒ 3'	GGGGCGGCCGGCCGCTTCGAGCAGACATGAGAGCAAAAGGCCAGCAAAAG
g686	SEAP (30bp) - f.SV40pA - bac - TB - pEIF4A1.636, 5' ⇒ 3'	GGGGCGGCCGGCCGCTTCGAGCAGACATGACAGACATGATAAGATACATTGATGAGTT
g687	SEAP (30bp) - T.GAPDH.2 - bac - TB - pEIF4A1.636, 5' ⇒ 3'	GGGGCGGCCGGCCGCTTCGAGCAGACATGATGTAGACCCCCTTGAAGAGGG
g688	SEAP (30bp) - T.ACTB.4 - bac - TB - pEIF4A1.636, 5' ⇒ 3'	GGGGCGGCCGGCCGCTTCGAGCAGACATGATTGCTTTCGTGTAAATTATGTAATGC
g689	SEAP (28bp) - 4x USE consensus - GAPDH spacer.2 - bac - TB - pEIF4A1.636, 5' ⇒ 3'	GGCGGCCGGCCGCTTCGAGCAGACATGATGTAATGTAATGTAATGTAAAATAAAgtaccc
g690	SEAP (26bp) - 4x USE consensus - EEF1A1 spacer.2 - bac - TB - pEIF4A1.636, 5' ⇒ 3'	CGGCCGGCCGCTTCGAGCAGACATGATGTAATGTAATGTAATGTAAAATAAAtgaatttt

Table A14: Primers for creating endogenous terminator variants

Primer	notes	sequence
g496	hrGFP (30bp) - SV40pA (late), 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAACAGACATGATAAGATACATTGATGAG
g497	bac elements (30bp) - SV40pA (late), 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCTACCACATTTGTAGAGGTTTTACTTG
g465	hrGFP (30bp)-minSV40pA.1, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGTAAccattataagctgcAATAAAcaagttaacaacAAcaattgc
g466	GRIK1 3' homology (30bp)-minSV40pA.1, 3' ⇒ 5'	ttccaccacgtctgagtcagggttgcatggGAAACATAaaatgaatgcaattgTTgttgttaacttgTTTATTgcagc

Table A14, continued:

Primer	notes	sequence
g466b	ColE1 homology (30bp)-minSV40pA.1, 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCGAAACATAaaatgaatgcaattgTTgttgttaacttgTTTATTgcagc
a467	hrGFP-minSV40pA.2, assembly 1 (26bp), 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGTAAccattataagctgcAATAAAcaagt
a468	minSV40pA.2, assembly 2 (26, 28bp), 3' ⇒ 5'	CTGAAACATAaaatgaatgcaattgTTgttgttaacttgTTTATTgcagcttataatggT
a469	minSV40pA.2-GRIK1 3' homology, assembly 3 (28, 20bp), 5' ⇒ 3'	cAAcaattgcattcatttTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTccat
a469b	minSV40pA.2-ColE1 homology, assembly 3 (28, 20bp), 5' ⇒ 3'	cAAcaattgcattcatttTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTGAGC
a470	minSV40pA.2-GRIK1 3' homology, assembly 4 (20 bp), 3' ⇒ 5'	ttccaccacgtctgagtcagggttgcatggAAAAAACCTCCCACAC
a470b	minSV40pA.2-ColE1 homology, assembly 4 (20 bp), 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCAAAAAACCTCCCACAC
g589	hrGFP-SV40pA.3, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAACAGACATGATAAGATACATTGATGAGTTTG
g590	ColE1 origin-SV40pA.3, 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCAAAAAACCTCCCACACCTCC
g591	hrGFP-SV40pA.4, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGTGATGCTATTGCTTTATTTGTAACC
g592	ColE1 origin-SV40pA.4, 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCTACCACATTTGTAGAGGTTTTACTTGC
a473	hrGFP-TE.1, assembly 1 (25bp), 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGGTATtcattacaaacttgctcactacAA
a474	TE.1, assembly 2 (25, 23bp), 3' ⇒ 5'	AAtgccacttcatcTTaaagcttaaaattcaTTTATTgtagtgagcaagtttgtaatgaA
a475	TE.1-GRIK1 3' homology, assembly 3 (23, 18bp), 5' ⇒ 3'	aagctttAAgatgaagtggcaTTTCTTTTccatgcaaccctgactcagacgtggtggaa
a475b	TE.1-ColE1 homology, assembly 3 (23, 17bp), 5' ⇒ 3'	aagctttAAgatgaagtggcaTTTCTTTTGAGCAAAAGGCCAGCAAAAGGCCAGGAACC
a476	GRIK1 3' homology end overlap, assembly 4 (18bp), 3' ⇒ 5'	ttccaccacgtctgagtc
a476b	ColE1 homology, assembly 4 (17bp), 3' ⇒ 5'	GGTTCCTGGCCTTTTGC
g529	ColE1 origin-TE.3, 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCAAAAGAAAtgccacttcatcTTaaagctt
g530	hrGFP-TE.3/TE.4, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGTGAAACCCAGTGTCTTAGACAAC
g531	ColE1 origin-TE.2/TE.4, 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCTAAACAAAAAAAGCCAAGCACTACCTTG
a477	hrGFP-TG.1, assembly 1 (24bp), 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGTCATGTACcatcAATAAAgtaccctgtg
a478	TG.1, assembly 2 (24, 23bp), 3' ⇒ 5'	GACCctagaataagacaggacaagTAactggttgagcacagggtacTTTATTgatgGTAC
a479	TG.1-GRIK1 3' homology, assembly 3 (23, 17bp), 5' ⇒ 3'	ttgtcctgtcttattctagGGTCTGGGGCAGAGGGGAGGGAAGCTGGGCTTGTGTCccat
a479b	TG.1-ColE1 homology, assembly 3 (23, 17bp), 5' ⇒ 3'	ttgtcctgtcttattctagGGTCTGGGGCAGAGGGGAGGGAAGCTGGGCTTGTGTCGAGC
a480	TG.1-GRIK1 3' homology, assembly 4 (17bp), 3' ⇒ 5'	ttccaccacgtctgagtcagggttgcatggGACACAAGCCCAG

Table A14, continued:

Primer	notes	sequence
a480b	TG.1-ColE1 homology, assembly 4 (17bp), 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCGACACAAG CCCAG
g532	hrGFP-TG.2, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGTAGAC CCCTTGAAGAGGGGAG
g533	ColE1 origin-GAPDH DSE, 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCGACACAAG CCCAGCTTCCCT
g636	hrGFP-T.ACTB.f, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAAGCGGACT ATGACTTAGTTGCGTTACAC
g637	ColE1 ori-T.ACTB, 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCGTTGGCCC CCACTGCCC
g638	hrGFP-T.ACTB, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATTGCTTT CGTGTAAATTATGTAATGCAAAA
g639	ColE1 ori-T.ACTB, 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCCCCCCAAC CCAGCCACAC
g640	ColE1 ori-T.ACTB.f, 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCTCCCATAG GTGAAGGCAAAGGC

Table A15: Primers for creating synthetic terminator variants

Primer	notes	sequence
573	ColE1 origin-DSE consensus-TGAPDH, 3' ⇒ 5'	GCTGGCCTTTTGCTCAAAACACActagaataagacaggacaagT Aactggttgag
574	Levitt consensus-TGAPDH, 3' ⇒ 5'	CACACAAAAAACCAACACACAGActagaataagacaggacaag TAactggttgag
575	Levitt+DSE consensus-TGAPDH, 3' ⇒ 5'	CACACAAAAAACCAACACACAGAAAAACACActagaataa gacaggacaagTAactggttgag
576	TGAPDH long USE, 5' ⇒ 3'	TGTAGACCCCTTGAAGAGGGG
g577	ColE1 origin-DSE consensus, 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCAAAACACA
g578	ColE1 origin-Levitt DSE consensus, 3' ⇒ 5'	GGTTCCTGGCCTTTTGCTGGCCTTTTGCTCCACACAAA AAACCAACACACAGA
579	4x USE consensus-TGAPDH, 5' ⇒ 3'	GTGGGTGTAATGTAATGTAATGTAATGTAAcatcAATAA AgtaccctgtgctcaaccagtTActtgtcct
g580	hrGFP-4x USE consensus-TGAPDH, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGTAATG TAATGTAATGTAAcatc
g581	hrGFP-4x USE consensus-GAPDH spacer 2, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGTAATG TAATGTAATGTAAAATAAAgtac
582	4x USE consensus-poly(A) signal-GAPDH spacer 2-GAPDH poly(A) site, 5' ⇒ 3'	TGTAATGTAATGTAATGTAAAATAAAgtaccctgtgctcaacca gtTA
583	ColE1 origin-DSE consensus-GAPDH poly(A) site-GAPDH spacer 2-poly(A) signal, 3' ⇒ 5'	CCTTTTGCTCAAAACACATAactggttgagcacagggtacTTTAT TT
584	Levitt consensus-GAPDH poly(A) site-GAPDH spacer 2-poly(A) signal, 3' ⇒ 5'	CAAAAAACCAACACACAGATAactggttgagcacagggtacTTT ATTT
585	Levitt+DSE consensus-GAPDH poly(A) site-GAPDH spacer 2-poly(A) signal, 3' ⇒ 5'	CAAAAAACCAACACACAGAAAAACACATAactggttgagca cagggtacTTTATTT
g612	ColE1 origin-DSE consensus-TEEF1A1, 3' ⇒ 5'	GCTGGCCTTTTGCTCAAAACACAtgccacttcatcTTaaagcttaa aattc

Table A15, continued:

Primer	notes	sequence
g613	Levitt consensus-TEEF1A1, 3' ⇒ 5'	CACACAAAAAACCAACACACAGAtgccacttcatcTTaaagcttaaaattc
g614	Levitt+DSE consensus-TEEF1A1, 3' ⇒ 5'	CACACAAAAAACCAACACACAGAAAAACACAtgccacttcatcTTaaagcttaaaattc
f615	hrGFP-terminator flanking primer, 5' ⇒ 3'	CCCCTGGGCAGCCTG
g616	4x USE consensus-TEEF1A1, 5' ⇒ 3'	GTGGGTGTAATGTAATGTAATGTAATGTAAtcattacaaacttgctcactacAATAAAtg
g617	hrGFP-4x USE consensus-TEEF1A1, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGTAATGTAATGTAATGTAAtca
618	4x USE consensus-poly(A) signal-EEF1A1 spacer 2-EEF1A1 poly(A) site, 5' ⇒ 3'	TGTAATGTAATGTAATGTAAAATAAAtgaattttaagctttAAT
619	ColE1 origin-DSE consensus-EEF1A1 poly(A) site-EEF1A1 spacer 2-poly(A) signal, 3' ⇒ 5'	CCTTTTGCTCAAAACACATTaaagcttaaaattcaTTTATTTTACATTAC
620	Levitt consensus-EEF1A1 poly(A) site-EEF1A1 spacer 2-poly(A) signal, 3' ⇒ 5'	CAAAAAACCAACACACAGATTaaagcttaaaattcaTTTATTTTACATTAC
621	Levitt+DSE consensus-EEF1A1 poly(A) site-EEF1A1 spacer 2-poly(A) signal, 3' ⇒ 5'	CAAAAAACCAACACACAGAAAAACACATTaaagcttaaaattcaTTTATTTTACATTAC
g622	hrGFP-4x USE consensus-EEF1A1 spacer 2, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGTAATGTAATGTAATGTAAAATAAAtga
657	4x USE consensus-poly(A) signal-ACTB spacer 2-ACTB poly(A) site, 5' ⇒ 3'	TGTAATGTAATGTAATGTAAAATAAAagtgcacaccttaaaaatGA
658	Levitt+DSE consensus-ACTB poly(A) site-ACTB spacer 2-poly(A) signal, 3' ⇒ 5'	CAAAAAACCAACACACAGAAAAACACATCatttttaaggtgtgcactTTTATTT
659	4x USE consensus-poly(A) signal-ACTB spacer 2-poly(A) site, 5' ⇒ 3'	TGTAATGTAATGTAATGTAAAATAAAagtgcacaccttaaaaatCA
660	Levitt+DSE consensus-poly(A) site-ACTB spacer 2-poly(A) signal, 3' ⇒ 5'	CAAAAAACCAACACACAGAAAAACACATGatttttaaggtgtgcactTTTATTT
g661	hrGFP-4x USE consensus-ACTB spacer 2, 5' ⇒ 3'	CCCCTGGGCAGCCTGCACGAGTGGGTGTAATGTAATGTAATGTAATGTAAAATAAAagtg
662	4x USE consensus-poly(A) signal-GAPDH spacer 2-poly(A) site, 5' ⇒ 3'	TGTAATGTAATGTAATGTAAAATAAAgtaccctgtgctcaaccagtCA
663	ColE1 origin-DSE consensus-poly(A) site-GAPDH spacer 2-poly(A) signal, 3' ⇒ 5'	CCTTTTGCTCAAAACACATGactggttgagcacagggtacTTTATTT
664	Levitt consensus-poly(A) site-GAPDH spacer 2-poly(A) signal, 3' ⇒ 5'	CAAAAAACCAACACACAGATGactggttgagcacagggtacTTTATTT
665	Levitt+DSE consensus-poly(A) site-GAPDH spacer 2-poly(A) signal, 3' ⇒ 5'	CAAAAAACCAACACACAGAAAAACACATGactggttgagcacagggtacTTTATTT
666	4x USE consensus-poly(A) signal-EEF1A1 spacer 2-poly(A) site, 5' ⇒ 3'	TGTAATGTAATGTAATGTAAAATAAAtgaattttaagctttCAT
667	ColE1 origin-DSE consensus-poly(A) site-EEF1A1 spacer 2-poly(A) signal, 3' ⇒ 5'	CCTTTTGCTCAAAACACATGaaagcttaaaattcaTTTATTTTACATTAC

Table A15, continued:

Primer	notes	sequence
668	Levitt consensus-poly(A) site-EEF1A1 spacer 2-poly(A) signal, 3' ⇒ 5'	CAAAAAACCAACACACAGATGaaagcttaaaattcaTTTATTTTACATTAC
669	Levitt+DSE consensus-poly(A) site-EEF1A1 spacer 2-poly(A) signal, 3' ⇒ 5'	CAAAAAACCAACACACAGAAAAACACATGaaagcttaaaattcaTTTATTTTACATTAC
677	ColE1 origin-MC4R DSE-poly(A) site-GAPDH spacer 2-poly(A) signal, 3' ⇒ 5'	CCTTTTGCTCAATGCTTATGAATAACACGTGactggttgagcacagggtacTTTATTT
678	ColE1 origin-#7 DSE (CstF-64 RRM)-poly(A) site-GAPDH spacer 2-poly(A) signal, 3' ⇒ 5'	CCTTTTGCTCCAACACACACAACTGactggttgagcacagggtacTTTATTT

Table A16: gBlocks® for terminators

fragment	sequence
TE-TG (3' UTR from *EEF1A1* and *GAPDH*)	TGTGAAACCCAGTGTCTTAGACAACTGTGGCTTGAGCACCACCTGCTGGTATTCATTACAAACTTGCTCACTACAATAAATGAATTTTAAGCTTTAAGATGAAGTGGCATTTCTTTTAACAGTTACTATGTTGGAATTGGTTACAAATTTTGGAGTGGATTTCAAAAGTGAGAGCTAACTTCAGTTGATTTCAAGGTAGTGCTTGGCTTTTTTTGTTTATGTAGACCCCTTGAAGAGGGGAGGGGCCTAGGGAGCCGCACCTTGTCATGTACCATCAATAAAGTACCCTGTGCTCAACCAGTTACTTGTCCTGTCTTATTCTAGGGTCTGGGGCAGAGGGGAGGGAAGCTGGGCTTGTGTC

Appendix B: Sequences

B.1. COMMON SEQUENCES USED IN THIS WORK

Table B1: Coding sequences of reporter proteins and antibiotic resistance

gene	sequence
humanized *Renilla* GFP (hrGFP): 720-bp	ATGGTGAGCAAGCAGATCCTGAAGAACACCGGCCTGCAGGAGATCATGAGCTTCAAGGTGAAC CTGGAGGGCGTGGTGAACAACCACGTGTTCACCATGGAGGGCTGCGGCAAGGGCAACATCCTG TTCGGCAACCAGCTGGTGCAGATCCGCGTGACCAAGGGCGCCCCCCTGCCCTTCGCCTTCGACA TCCTGAGCCCCGCCTTCCAGTACGGCAACCGCACCTTCACCAAGTACCCCGAGGACATCAGCG ACTTCTTCATCCAGAGCTTCCCCGCCGGCTTCGTGTACGAGCGCACCCTGCGCTACGAGGACGG CGGCCTGGTGGAGATCCGCAGCGACATCAACCTGATCGAGGAGATGTTCGTGTACCGCGTGGA GTACAAGGGCCGCAACTTCCCCAACGACGGCCCCGTGATGAAGAAGACCATCACCGGCCTGCA GCCCAGCTTCGAGGTGGTGTACATGAACGACGGCGTGCTGGTGGGCCAGGTGATCCTGGTGTA CCGCCTGAACAGCGGCAAGTTCTACAGCTGCCACATGCGCACCCTGATGAAGAGCAAGGGCGT GGTGAAGGACTTCCCCGAGTACCACTTCATCCAGCACCGCCTGGAGAAGACCTACGTGGAGGA CGGCGGCTTCGTGGAGCAGCACGAGACCGCCATCGCCCAGCTGACCAGCCTGGGCAAGCCCCT GGGCAGCCTGCACGAGTGGGTGTAA
secreted alkaline phosphatase (SEAP): 1560-bp	ATGCTGCTGCTGCTGCTGCTGGGCCTGAGGCTACAGCTCTCCCTGGGCATCATCCCAGTTG AGGAGGAGAACCCGGACTTCTGGAACCGCGAGGCAGCCGAGGCCCTGGGTGCCGCCAAGAAG CTGCAGCCTGCACAGACAGCCGCCAAGAACCTCATCATCTTCCTGGGCGATGGGATGGGGGTG TCTACGGTGACAGCTGCCAGGATCCTAAAAGGGCAGAAGAAGGACAAACTGGGGCCTGAGAT ACCCCTGGCCATGGACCGCTTCCCATATGTGGCTCTGTCCAAGACATACAATGTAGACAAACAT GTGCCAGACAGTGGAGCCACAGCCACGGCCTACCTGTGCGGGGTCAAGGGCAACTTCCAGACC ATTGGCTTGAGTGCAGCCGCCCGCTTTAACCAGTGCAACACGACACGCGGCAACGAGGTCATC TCCGTGATGAATCGGGCCAAGAAAGCAGGGAAGTCAGTGGGAGTGGTAACCACCACACGAGT GCAGCACGCCTCGCCAGCCGGCACCTACGCCCACACGGTGAACCGCAACTGGTACTCGGACGC CGACGTGCCTGCCTCGGCCCGCCAGGAGGGGTGCCAGGACATCGCTACGCAGCTCATCTCCAA CATGGACATTGACGTGATCCTAGGTGGAGGCCGAAAGTACATGTTTCGCATGGGAACCCCAGA CCCTGAGTACCCAGATGACTACAGCCAAGGTGGGACCAGGCTGGACGGGAAGAATCTGGTGCA GGAATGGCTGGCGAAGCGCCAGGGTGCCCGGTATGTGTGGAACCGCACTGAGCTCATGCAGGC TTCCCTGGACCCGTCTGTGACCCATCTCATGGGTCTCTTTGAGCCTGGAGACATGAAATACGAG ATCCACCGAGACTCCACACTGGACCCCTCCCTGATGGAGATGACAGAGGCTGCCCTGCGCCTG CTGAGCAGGAACCCCCGCGGCTTCTTCCTCTTCGTGGAGGGTGGTCGCATCGACCATGGTCATC ATGAAAGCAGGGCTTACCGGGCACTGACTGAGACGATCATGTTCGACGACGCCATTGAGAGGG CGGGCCAGCTCACCAGCGAGGAGGACACGCTGAGCCTCGTCACTGCCGACCACTCCCACGTCT TCTCCTTCGGAGGCTACCCCCTGCGAGGGAGCTCCATCTTCGGGCTGGCCCCTGGCAAGGCCCG GGACAGGAAGGCCTACACGGTCCTCCTATACGGAAACGGTCCAGGCTATGTGCTCAAGGACGG CGCCCGGCCGGATGTTACCGAGAGCGAGAGCGGGAGCCCCGAGTATCGGCAGCAGTCAGCAG TGCCCCTGGACGAAGAGACCCACGCAGGCGAGGACGTGGCGGTGTTCGCGCGCGGCCCGCAG GCGCACCTGGTTCACGGCGTGCAGGAGCAGACCTTCATAGCGCACGTCATGGCCTTCGCCGCCT GCCTGGAGCCCTACACCGCCTGCGACCTGGCGCCCCCCGCCGGCACCACCGACGCCGCGCACC CGGGTTACTCTAGAGTCGGGGCGGCCGGCCGCTTCGAGCAGACATGA
Zeocin™ resistance: 375-bp	ATGGCCAAGTTGACCAGTGCCGTTCCGGTGCTCACCGCGCGCGACGTCGCCGGAGCGGTCGAG TTCTGGACCGACCGGCTCGGGTTCTCCCGGGACTTCGTGGAGGACGACTTCGCCGGTGTGGTCC GGGACGACGTGACCCTGTTCATCAGCGCGGTCCAGGACCAGGTGGTGCCGGACAACACCCTGG CCTGGGTGTGGGTGCGCGGCCTGGACGAGCTGTACGCCGAGTGGTCGGAGGTCGTGTCCACGA ACTTCCGGGACGCCTCCGGGCCGGCCATGACCGAGATCGGCGAGCAGCCGTGGGGGCGGGAGT TCGCCCTGCGCGACCCGGCCGGCAACTGCGTGCACTTCGTGGCCGAGGAGCAGGACTGA
puromycin resistance: 600-bp	ATGACCGAGTACAAGCCCACGGTGCGCCTCGCCACCCGCGACGACGTCCCCAGGGCCGTACGC ACCCTCGCCGCCGCGTTCGCCGACTACCCCGCCACGCGCCACACCGTCGATCCGGACCGCCAC ATCGAGCGGGTCACCGAGCTGCAAGAACTCTTCCTCACGCGCGTCGGGCTCGACATCGGCAAG GTGTGGGTCGCGGACGACGGCGCCGCGGTGGCGGTCTGGACCACGCCGGAGAGCGTCGAAGC GGGGGCGGTGTTCGCCGAGATCGGCCCGCGCATGGCCGAGTTGAGCGGTTCCCGGCTGGCCGC GCAGCAACAGATGGAAGGCCTCCTGGCGCCGCACCGGCCCAAGGAGCCCGCGTGGTTCCTGGC CACCGTCGGCGTCTCGCCCGACCACCAGGGCAAGGGTCTGGGCAGCGCCGTCGTGCTCCCCGG AGTGGAGGCGGCCGAGCGCGCCGGGGTGCCCGCCTTCCTGGAGACCTCCGCGCCCCGCAACCT CCCCTTCTACGAGCGGCTCGGCTTCACCGTCACCGCCGACGTCGAGGTGCCCGAAGGACCGCG CACCTGGTGCATGACCCGCAAGCCCGGTGCCTGA

Table B1, continued:

gene	sequence
hygromycin-B resistance: 1029-bp	ATGGGTAAAAAGCCTGAACTCACCGCGACGTCTGTCGAGAAGTTTCTGATCGAAAAGTTCGAC AGCGTCTCCGACCTGATGCAGCTCTCGGAGGGCGAAGAATCTCGTGCTTTCAGCTTCGATGTAG GAGGGCGTGGATATGTCCTGCGGGTAAATAGCTGCGCCGATGGTTTCTACAAAGATCGTTATGT TTATCGGCACTTTGCATCGGCCGCGCTCCCGATTCCGGAAGTGCTTGACATTGGGGAATTCAGC GAGAGCCTGACCTATTGCATCTCCCGCCGTGCACAGGGTGTCACGTTGCAAGACCTGCCTGAA ACCGAACTGCCCGCTGTTCTGCAGCCGGTCGCGGAGGCCATGGATGCGATCGCTGCGGCCGAT CTTAGCCAGACGAGCGGGTTCGGCCCATTCGGACCGCAAGGAATCGGTCAATACACTACATGG CGTGATTTCATATGCGCGATTGCTGATCCCCATGTGTATCACTGGCAAACTGTGATGGACGACA CCGTCAGTGCGTCCGTCGCGCAGGCTCTCGATGAGCTGATGCTTTGGGCCGAGGACTGCCCCGA AGTCCGGCACCTCGTGCACGCGGATTTCGGCTCCAACAATGTCCTGACGGACAATGGCCGCAT AACAGCGGTCATTGACTGGAGCGAGGCGATGTTCGGGGATTCCCAATACGAGGTCGCCAACAT CTTCTTCTGGAGGCCGTGGTTGGCTTGTATGGAGCAGCAGACGCGCTACTTCGAGCGGAGGCAT CCGGAGCTTGCAGGATCGCCGCGGCTCCGGGCGTATATGCTCCGCATTGGTCTTGACCAACTCT ATCAGAGCTTGGTTGACGGCAATTTCGATGATGCAGCTTGGGCGCAGGGTCGATGCGACGCAA TCGTCCGATCCGGAGCCGGGACTGTCGGGCGTACACAAATCGCCCGCAGAAGCGCGGCCGTCT GGACCGATGGCTGTGTAGAAGTACTCGCCGATAGTGGAAACCGACGCCCCAGCACTCGTCCGA GGGCAAAGGAATAA

Table B2: Base dual-expression transgene cassette for evaluating promoters used in Chapters 4 and 5

construct	sequence
SEAP-IRES-hrGFP-SV40 poly(A): 3097-bp	ATGCTGCTGCTGCTGCTGCTGGGCCTGAGGCTACAGCTCTCCCTGGGCATCATCCCAGTTG AGGAGGAGAACCCGGACTTCTGGAACCGCGAGGCAGCCGAGGCCCTGGGTGCCGCCAAGAAG CTGCAGCCTGCACAGACAGCCGCCAAGAACCTCATCATCTTCCTGGGCGATGGGATGGGGGTG TCTACGGTGACAGCTGCCAGGATCCTAAAAGGGCAGAAGAAGGACAAACTGGGGCCTGAGAT ACCCCTGGCCATGGACCGCTTCCCATATGTGGCTCTGTCCAAGACATACAATGTAGACAAACAT GTGCCAGACAGTGGAGCCACAGCCACGGCCTACCTGTGCGGGGTCAAGGGCAACTTCCAGACC ATTGGCTTGAGTGCAGCCGCCCGCTTTAACCAGTGCAACACGACACGCGGCAACGAGGTCATC TCCGTGATGAATCGGGCCAAGAAAGCAGGGAAGTCAGTGGGAGTGGTAACCACCACACGAGT GCAGCACGCCTCGCCAGCCGGCACCTACGCCCACACGGTGAACCGCAACTGGTACTCGGACGC CGACGTGCCTGCCTCGGCCCGCCAGGAGGGGTGCCAGGACATCGCTACGCAGCTCATCTCCAA CATGGACATTGACGTGATCCTAGGTGGAGGCCGAAAGTACATGTTTCGCATGGGAACCCCAGA CCCTGAGTACCCAGATGACTACAGCCAAGGTGGGACCAGGCTGGACGGGAAGAATCTGGTGCA GGAATGGCTGGCGAAGCGCCAGGGTGCCCGGTATGTGTGGAACCGCACTGAGCTCATGCAGGC TTCCCTGGACCCGTCTGTGACCCATCTCATGGGTCTCTTTGAGCCTGGAGACATGAAATACGAG ATCCACCGAGACTCCACACTGGACCCCTCCCTGATGGAGATGACAGAGGCTGCCCTGCGCCTG CTGAGCAGGAACCCCCGCGGCTTCTTCCTCTTCGTGGAGGGTGGTCGCATCGACCATGGTCATC ATGAAAGCAGGGCTTACCGGGCACTGACTGAGACGATCATGTTCGACGACGCCATTGAGAGGG CGGGCCAGCTCACCAGCGAGGAGGACACGCTGAGCCTCGTCACTGCCGACCACTCCCACGTCT TCTCCTTCGGAGGCTACCCCCTGCGAGGGAGCTCCATCTTCGGGCTGGCCCCTGGCAAGGCCCG GGACAGGAAGGCCTACACGGTCCTCCTATACGGAAACGGTCCAGGCTATGTGCTCAAGGACGG CGCCCGGCCGGATGTTACCGAGAGCGAGAGCGGGAGCCCCGAGTATCGGCAGCAGTCAGCAG TGCCCCTGGACGAAGAGACCCACGCAGGCGAGGACGTGGCGGTGTTCGCGCGCGGCCCGCAG GCGCACCTGGTTCACGGCGTGCAGGAGCAGACCTTCATAGCGCACGTCATGGCCTTCGCCGCCT GCCTGGAGCCCTACACCGCCTGCGACCTGGCGCCCCCCGCCGGCACCACCGACGCCGCGCACC CGGGTTACTCTAGAGTCGGGGCGGCCGGCCGCTTCGAGCAGACATGAGTTTAAACCCCCTCTCC CTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTGTCTAT ATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTGGCCCTGTCTT CTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAGGTCTGTTGAATGTC GTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGTCTGTAGCGACCCTTTGC AGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCCAAAAGCCACGTGTATAAGAT ACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGAGTTGGATAGTTGTGGAAAGAGTC AAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGAAGGATGCCCAGAAGGTACCCCATTGT ATGGGATCTGATCTGGGGCCTCGGTGCACATGCTTTACATGTGTTTAGTCGAGGTTAAAAAACG TCTAGGCCCCCCGAACCACGGGGACGTGGTTTTCCTTTGAAAAACACGATGATAATGTCTAGA ATGGTGAGCAAGCAGATCCTGAAGAACACCGGCCTGCAGGAGATCATGAGCTTCAAGGTGAAC CTGGAGGGCGTGGTGAACAACCACGTGTTCACCATGGAGGGCTGCGGCAAGGGCAACATCCTG TTCGGCAACCAGCTGGTGCAGATCCGCGTGACCAAGGGCGCCCCCCTGCCCTTCGCCTTCGACA TCCTGAGCCCCGCCTTCCAGTACGGCAACCGCACCTTCACCAAGTACCCCGAGGACATCAGCG ACTTCTTCATCCAGAGCTTCCCCGCCGGCTTCGTGTACGAGCGCACCCTGCGCTACGAGGACGG CGGCCTGGTGGAGATCCGCAGCGACATCAACCTGATCGAGGAGATGTTCGTGTACCGCGTGGA GTACAAGGGCCGCAACTTCCCCAACGACGGCCCCGTGATGAAGAAGACCATCACCGGCCTGCA GCCCAGCTTCGAGGTGGTGTACATGAACGACGGCGTGCTGGTGGGCCAGGTGATCCTGGTGTA CCGCCTGAACAGCGGCAAGTTCTACAGCTGCCACATGCGCACCCTGATGAAGAGCAAGGGCGT GGTGAAGGACTTCCCCGAGTACCACTTCATCCAGCACCGCCTGGAGAAGACCTACGTGGAGGA CGGCGGCTTCGTGGAGCAGCACGAGACCGCCATCGCCCAGCTGACCAGCCTGGGCAAGCCCCT GGGCAGCCTGCACGAGTGGGTGTAAATGCATCAGACATGATAAGATACATTGATGAGTTTGGA CAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTT TATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCATTCATTTTATGTT TCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTA

Table B2, continued:

construct	sequence
CMV enhancer-SEAP-IRES-hrGFP-SV40 poly(A): 3700-bp	GACCGCCATGTTGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGT TCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTCGTGACCGC CCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGA CTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGT GTATCATATGCCAAGTCCGGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTA TGCCCAGTACATGACCTTACGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCT ATTACCATGGTGATGCGGTTTTGGCAGTACACCAATGGGCGTGGATAGCGGTTTGACTCACGG GGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGG ACTTTCCAAAATGTCGTAATAACCCCGCCCCGTTGACGCAAATGGGCGGTAGGCGTGTACGGT GGGGCGGCCGCAAAAGGAAAACTAGCTAGCATGCTGCTGCTGCTGCTGCTGGGCCTGAGG CTACAGCTCTCCCTGGGCATCATCCCAGTTGAGGAGGAGAACCCGGACTTCTGGAACCGCGAG GCAGCCGAGGCCCTGGGTGCCGCCAAGAAGCTGCAGCCTGCACAGACAGCCGCCAAGAACCTC ATCATCTTCCTGGGCGATGGGATGGGGGTGTCTACGGTGACAGCTGCCAGGATCCTAAAAGGG CAGAAGAAGGACAAACTGGGGCCTGAGATACCCCTGGCCATGGACCGCTTCCCATATGTGGCT CTGTCCAAGACATACAATGTAGACAAACATGTGCCAGACAGTGGAGCCACAGCCACGGCCTAC CTGTGCGGGGTCAAGGGCAACTTCCAGACCATTGGCTTGAGTGCAGCCGCCCGCTTTAACCAGT GCAACACGACACGCGGCAACGAGGTCATCTCCGTGATGAATCGGGCCAAGAAAGCAGGGAAG TCAGTGGGAGTGGTAACCACCACACGAGTGCAGCACGCCTCGCCAGCCGGCACCTACGCCCAC ACGGTGAACCGCAACTGGTACTCGGACGCCGACGTGCCTGCCTCGGCCCGCCAGGAGGGGTGC CAGGACATCGCTACGCAGCTCATCTCCAACATGGACATTGACGTGATCCTAGGTGGAGGCCGA AAGTACATGTTTCGCATGGGAACCCCAGACCCTGAGTACCCAGATGACTACAGCCAAGGTGGG ACCAGGCTGGACGGGAAGAATCTGGTGCAGGAATGGCTGGCGAAGCGCCAGGGTGCCCGGTA TGTGTGGAACCGCACTGAGCTCATGCAGGCTTCCCTGGACCCGTCTGTGACCCATCTCATGGGT CTCTTTGAGCCTGGAGACATGAAATACGAGATCCACCGAGACTCCACACTGGACCCCTCCCTG ATGGAGATGACAGAGGCTGCCCTGCGCCTGCTGAGCAGGAACCCCCGCGGCTTCTTCCTCTTCG TGGAGGGTGGTCGCATCGACCATGGTCATCATGAAAGCAGGGCTTACCGGGCACTGACTGAGA CGATCATGTTCGACGACGCCATTGAGAGGGCGGGCCAGCTCACCAGCGAGGAGGACACGCTGA GCCTCGTCACTGCCGACCACTCCCACGTCTTCTCCTTCGGAGGCTACCCCCTGCGAGGGAGCTC CATCTTCGGGCTGGCCCCTGGCAAGGCCCGGGACAGGAAGGCCTACACGGTCCTCCTATACGG AAACGGTCCAGGCTATGTGCTCAAGGACGGCGCCCGGCCGGATGTTACCGAGAGCGAGAGCG GGAGCCCCGAGTATCGGCAGCAGTCAGCAGTGCCCCTGGACGAAGAGACCCACGCAGGCGAG GACGTGGCGGTGTTCGCGCGCGGCCCGCAGGCGCACCTGGTTCACGGCGTGCAGGAGCAGACC TTCATAGCGCACGTCATGGCCTTCGCCGCCTGCCTGGAGCCCTACACCGCCTGCGACCTGGCGC CCCCCGCCGGCACCACCGACGCCGCGCACCCGGGTTACTCTAGAGTCGGGGCGGCCGGCCGCT TCGAGCAGACATGAGTTTAAACCCCCTCTCCCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCG CTTGGAATAAGGCCGGTGTGCGTTTGTCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCA ATGTGAGGGCCCGGAAACCTGGCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCT CGCCAAAGGAATGCAAGGTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTG AAGACAAACAACGTCTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTG CCTCTGCGGCCAAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCA CGTTGTGAGTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGG GCTGAAGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATG CTTTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGTTT TCCTTTGAAAAACACGATGATAATGTCTAGAATGGTGAGCAAGCAGATCCTGAAGAACACCGG CCTGCAGGAGATCATGAGCTTCAAGGTGAACCTGGAGGGCGTGGTGAACAACCACGTGTTCAC CATGGAGGGCTGCGGCAAGGGCAACATCCTGTTCGGCAACCAGCTGGTGCAGATCCGCGTGAC CAAGGGCGCCCCCCTGCCCTTCGCCTTCGACATCCTGAGCCCCGCCTTCCAGTACGGCAACCGC ACCTTCACCAAGTACCCCGAGGACATCAGCGACTTCTTCATCCAGAGCTTCCCCGCCGGCTTCG TGTACGAGCGCACCCTGCGCTACGAGGACGGCGGCCTGGTGGAGATCCGCAGCGACATCAACC TGATCGAGGAGATGTTCGTGTACCGCGTGGAGTACAAGGGCCGCAACTTCCCCAACGACGGCC CCGTGATGAAGAAGACCATCACCGGCCTGCAGCCCAGCTTCGAGGTGGTGTACATGAACGACG GCGTGCTGGTGGGCCAGGTGATCCTGGTGTACCGCCTGAACAGCGGCAAGTTCTACAGCTGCC ACATGCGCACCCTGATGAAGAGCAAGGGCGTGGTGAAGGACTTCCCCGAGTACCACTTCATCC AGCACCGCCTGGAGAAGACCTACGTGGAGGACGGCGGCTTCGTGGAGCAGCACGAGACCGCC ATCGCCCAGCTGACCAGCCTGGGCAAGCCCCTGGGCAGCCTGCACGAGTGGGTGTAAATGCAT CAGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAAT GCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAA GTTAACAACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTT AAAGCAAGTAAAACCTCTACAAATGTGGTA

B.2. CHAPTER 2 SPECIFIC

Table B3: Constructs for transgene integration

construct	Sequence (transgene region only)
pCMV-hrGFP-IRES-puro (GP) variant: 2733-bp	TAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTT ACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCA ATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGT ATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTAT TGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTT CCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGT ACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGT CAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCC CCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTGGTTTA GTGAACCGTCAGATCCGCTAGCATGGTGAGCAAGCAGATCCTGAAGAACACCGGCCTGCAGGA GATCATGAGCTTCAAGGTGAACCTGGAGGGCGTGGTGAACAACCACGTGTTCACCATGGAGGG CTGCGGCAAGGGCAACATCCTGTTCGGCAACCAGCTGGTGCAGATCCGCGTGACCAAGGGCGC CCCCCTGCCCTTCGCCTTCGACATCCTGAGCCCCGCCTTCCAGTACGGCAACCGCACCTTCACC AAGTACCCCGAGGACATCAGCGACTTCTTCATCCAGAGCTTCCCCGCCGGCTTCGTGTACGAGC GCACCCTGCGCTACGAGGACGGCGGCCTGGTGGAGATCCGCAGCGACATCAACCTGATCGAGG AGATGTTCGTGTACCGCGTGGAGTACAAGGGCCGCAACTTCCCCAACGACGGCCCCGTGATGA AGAAGACCATCACCGGCCTGCAGCCCAGCTTCGAGGTGGTGTACATGAACGACGGCGTGCTGG TGGGCCAGGTGATCCTGGTGTACCGCCTGAACAGCGGCAAGTTCTACAGCTGCCACATGCGCA CCCTGATGAAGAGCAAGGGCGTGGTGAAGGACTTCCCCGAGTACCACTTCATCCAGCACCGCC TGGAGAAGACCTACGTGGAGGACGGCGGCTTCGTGGAGCAGCACGAGACCGCCATCGCCCAG CTGACCAGCCTGGGCAAGCCCCTGGGCAGCCTGCACGAGTGGGTGTAACTCGAGCCCCTCTCC CTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTGTCTAT ATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTGGCCCTGTCTT CTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAGGTCTGTTGAATGTC GTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGTCTGTAGCGACCCTTTGC AGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCCAAAAGCCACGTGTATAAGAT ACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGAGTTGGATAGTTGTGGAAAGAGTC AAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGAAGGATGCCCAGAAGGTACCCCATTGT ATGGGATCTGATCTGGGGCCTCGGTGCACATGCTTTACATGTGTTTAGTCGAGGTTAAAAAACG TCTAGGCCCCCCGAACCACGGGGACGTGGTTTTCCTTTGAAAAACACGATGATAATGGTCGAC ATGACCGAGTACAAGCCCACGGTGCGCCTCGCCACCCGCGACGACGTCCCCAGGGCCGTACGC ACCCTCGCCGCCGCGTTCGCCGACTACCCCGCCACGCGCCACACCGTCGATCCGGACCGCCAC ATCGAGCGGGTCACCGAGCTGCAAGAACTCTTCCTCACGCGCGTCGGGCTCGACATCGGCAAG GTGTGGGTCGCGGACGACGGCGCCGCGGTGGCGGTCTGGACCACGCCGGAGAGCGTCGAAGC GGGGGCGGTGTTCGCCGAGATCGGCCCGCGCATGGCCGAGTTGAGCGGTTCCCGGCTGGCCGC GCAGCAACAGATGGAAGGCCTCCTGGCGCCGCACCGGCCCAAGGAGCCCGCGTGGTTCCTGGC CACCGTCGGCGTCTCGCCCGACCACCAGGGCAAGGGTCTGGGCAGCGCCGTCGTGCTCCCCGG AGTGGAGGCGGCCGAGCGCGCCGGGGTGCCCGCCTTCCTGGAGACCTCCGCGCCCCGCAACCT CCCCTTCTACGAGCGGCTCGGCTTCACCGTCACCGCCGACGTCGAGGTGCCCGAAGGACCGCG CACCTGGTGCATGACCCGCAAGCCCGGTGCCTGAGGCGCGCCCAGACATGATAAGATACATTG ATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTG ATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCAT TCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTAC AAATGTGGTA

Table B3, continued:

construct	Sequence (transgene region only)
pCMV-puro-IRES-hrGFP (PG) variant: 2733-bp	TAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTT ACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCA ATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGT ATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTAT TGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTT CCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGT ACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGT CAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCC CCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTGGTTTA GTGAACCGTCAGATCCGCTAGCATGACCGAGTACAAGCCCACGGTGCGCCTCGCCACCCGCGA CGACGTCCCCAGGGCCGTACGCACCCTCGCCGCCGCGTTCGCCGACTACCCCGCCACGCGCCA CACCGTCGATCCGGACCGCCACATCGAGCGGGTCACCGAGCTGCAAGAACTCTTCCTCACGCG CGTCGGGCTCGACATCGGCAAGGTGTGGGTCGCGGACGACGGCGCCGCGGTGGCGGTCTGGAC CACGCCGGAGAGCGTCGAAGCGGGGGCGGTGTTCGCCGAGATCGGCCCGCGCATGGCCGAGTT GAGCGGTTCCCGGCTGGCCGCGCAGCAACAGATGGAAGGCCTCCTGGCGCCGCACCGGCCCAA GGAGCCCGCGTGGTTCCTGGCCACCGTCGGCGTCTCGCCCGACCACCAGGGCAAGGGTCTGGG CAGCGCCGTCGTGCTCCCCGGAGTGGAGGCGGCCGAGCGCGCCGGGGTGCCCGCCTTCCTGGA GACCTCCGCGCCCCGCAACCTCCCCTTCTACGAGCGGCTCGGCTTCACCGTCACCGCCGACGTC GAGGTGCCCGAAGGACCGCGCACCTGGTGCATGACCCGCAAGCCCGGTGCCTGACTCGAGCCC CTCTCCCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTT GTCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTGGCC CTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAGGTCTGTT GAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGTCTGTAGCGAC CCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCCAAAAGCCACGTGT ATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGAGTTGGATAGTTGTGGA AAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGAAGGATGCCCAGAAGGTACC CCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCTTTACATGTGTTTAGTCGAGGTTA AAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGTTTTCCTTTGAAAAACACGATGATAAT GGTCGACATGGTGAGCAAGCAGATCCTGAAGAACACCGGCCTGCAGGAGATCATGAGCTTCAA GGTGAACCTGGAGGGCGTGGTGAACAACCACGTGTTCACCATGGAGGGCTGCGGCAAGGGCA ACATCCTGTTCGGCAACCAGCTGGTGCAGATCCGCGTGACCAAGGGCGCCCCCCTGCCCTTCGC CTTCGACATCCTGAGCCCCGCCTTCCAGTACGGCAACCGCACCTTCACCAAGTACCCCGAGGAC ATCAGCGACTTCTTCATCCAGAGCTTCCCCGCCGGCTTCGTGTACGAGCGCACCCTGCGCTACG AGGACGGCGGCCTGGTGGAGATCCGCAGCGACATCAACCTGATCGAGGAGATGTTCGTGTACC GCGTGGAGTACAAGGGCCGCAACTTCCCCAACGACGGCCCCGTGATGAAGAAGACCATCACCG GCCTGCAGCCCAGCTTCGAGGTGGTGTACATGAACGACGGCGTGCTGGTGGGCCAGGTGATCC TGGTGTACCGCCTGAACAGCGGCAAGTTCTACAGCTGCCACATGCGCACCCTGATGAAGAGCA AGGGCGTGGTGAAGGACTTCCCCGAGTACCACTTCATCCAGCACCGCCTGGAGAAGACCTACG TGGAGGACGGCGGCTTCGTGGAGCAGCACGAGACCGCCATCGCCCAGCTGACCAGCCTGGGCA AGCCCCTGGGCAGCCTGCACGAGTGGGTGTAAGGCGCGCCCAGACATGATAAGATACATTGAT GAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGAT GCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCATTC ATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTACAA ATGTGGTA

Table B3, continued:

construct	Sequence (transgene region only)
pCMV-SEAP-IRES-puro (SP) variant: 3573-bp	TAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTT ACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCA ATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGT ATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTAT TGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTT CCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGT ACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGT CAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCC CCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTGGTTTA GTGAACCGTCAGATCCGCTAGCATGCTGCTGCTGCTGCTGCTGGGCCTGAGGCTACAGCTC TCCCTGGGCATCATCCCAGTTGAGGAGGAGAACCCGGACTTCTGGAACCGCGAGGCAGCCGAG GCCCTGGGTGCCGCCAAGAAGCTGCAGCCTGCACAGACAGCCGCCAAGAACCTCATCATCTTC CTGGGCGATGGGATGGGGGTGTCTACGGTGACAGCTGCCAGGATCCTAAAAGGGCAGAAGAA GGACAAACTGGGGCCTGAGATACCCCTGGCCATGGACCGCTTCCCATATGTGGCTCTGTCCAA GACATACAATGTAGACAAACATGTGCCAGACAGTGGAGCCACAGCCACGGCCTACCTGTGCGG GGTCAAGGGCAACTTCCAGACCATTGGCTTGAGTGCAGCCGCCCGCTTTAACCAGTGCAACAC GACACGCGGCAACGAGGTCATCTCCGTGATGAATCGGGCCAAGAAAGCAGGGAAGTCAGTGG GAGTGGTAACCACCACACGAGTGCAGCACGCCTCGCCAGCCGGCACCTACGCCCACACGGTGA ACCGCAACTGGTACTCGGACGCCGACGTGCCTGCCTCGGCCCGCCAGGAGGGGTGCCAGGACA TCGCTACGCAGCTCATCTCCAACATGGACATTGACGTGATCCTAGGTGGAGGCCGAAAGTACA TGTTTCGCATGGGAACCCCAGACCCTGAGTACCCAGATGACTACAGCCAAGGTGGGACCAGGC TGGACGGGAAGAATCTGGTGCAGGAATGGCTGGCGAAGCGCCAGGGTGCCCGGTATGTGTGG AACCGCACTGAGCTCATGCAGGCTTCCCTGGACCCGTCTGTGACCCATCTCATGGGTCTCTTTG AGCCTGGAGACATGAAATACGAGATCCACCGAGACTCCACACTGGACCCCTCCCTGATGGAGA TGACAGAGGCTGCCCTGCGCCTGCTGAGCAGGAACCCCCGCGGCTTCTTCCTCTTCGTGGAGGG TGGTCGCATCGACCATGGTCATCATGAAAGCAGGGCTTACCGGGCACTGACTGAGACGATCAT GTTCGACGACGCCATTGAGAGGGCGGGCCAGCTCACCAGCGAGGAGGACACGCTGAGCCTCGT CACTGCCGACCACTCCCACGTCTTCTCCTTCGGAGGCTACCCCCTGCGAGGGAGCTCCATCTTC GGGCTGGCCCCTGGCAAGGCCCGGGACAGGAAGGCCTACACGGTCCTCCTATACGGAAACGGT CCAGGCTATGTGCTCAAGGACGGCGCCCGGCCGGATGTTACCGAGAGCGAGAGCGGGAGCCCC GAGTATCGGCAGCAGTCAGCAGTGCCCCTGGACGAAGAGACCCACGCAGGCGAGGACGTGGC GGTGTTCGCGCGCGGCCCGCAGGCGCACCTGGTTCACGGCGTGCAGGAGCAGACCTTCATAGC GCACGTCATGGCCTTCGCCGCCTGCCTGGAGCCCTACACCGCCTGCGACCTGGCGCCCCCCGCC GGCACCACCGACGCCGCGCACCCGGGTTACTCTAGAGTCGGGGCGGCCGGCCGCTTCGAGCAG ACATGACTCGAGCCCCTCTCCCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAA GGCCGGTGTGCGTTTGTCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGC CCGGAAACCTGGCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGA ATGCAAGGTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACA ACGTCTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGC CAAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGAGT TGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGAAGGAT GCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCTTTACATGTG TTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGTTTTCCTTTGAAAA ACACGATGATAATGGTCGACATGACCGAGTACAAGCCCACGGTGCGCCTCGCCACCCGCGACG ACGTCCCCAGGGCCGTACGCACCCTCGCCGCCGCGTTCGCCGACTACCCCGCCACGCGCCACA CCGTCGATCCGGACCGCCACATCGAGCGGGTCACCGAGCTGCAAGAACTCTTCCTCACGCGCG TCGGGCTCGACATCGGCAAGGTGTGGGTCGCGGACGACGGCGCCGCGGTGGCGGTCTGGACCA CGCCGGAGAGCGTCGAAGCGGGGGCGGTGTTCGCCGAGATCGGCCCGCGCATGGCCGAGTTGA GCGGTTCCCGGCTGGCCGCGCAGCAACAGATGGAAGGCCTCCTGGCGCCGCACCGGCCCAAGG AGCCCGCGTGGTTCCTGGCCACCGTCGGCGTCTCGCCCGACCACCAGGGCAAGGGTCTGGGCA GCGCCGTCGTGCTCCCCGGAGTGGAGGCGGCCGAGCGCGCCGGGGTGCCCGCCTTCCTGGAGA CCTCCGCGCCCCGCAACCTCCCCTTCTACGAGCGGCTCGGCTTCACCGTCACCGCCGACGTCGA GGTGCCCGAAGGACCGCGCACCTGGTGCATGACCCGCAAGCCCGGTGCCTGAGGCGCGCCCAG ACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCT TTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTT AACAACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAA GCAAGTAAAACCTCTACAAATGTGGTA

Table B3, continued:

construct	Sequence (transgene region only)
pCMV-puro-IRES-SEAP (PS) variant: 3573-bp	TAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTT ACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCA ATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGT ATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTAT TGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTT CCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGT ACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGT CAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCC CCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTGGTTTA GTGAACCGTCAGATCCGCTAGCATGACCGAGTACAAGCCCACGGTGCGCCTCGCCACCCGCGA CGACGTCCCCAGGGCCGTACGCACCCTCGCCGCCGCGTTCGCCGACTACCCCGCCACGCGCCA CACCGTCGATCCGGACCGCCACATCGAGCGGGTCACCGAGCTGCAAGAACTCTTCCTCACGCG CGTCGGGCTCGACATCGGCAAGGTGTGGGTCGCGGACGACGGCGCCGCGGTGGCGGTCTGGAC CACGCCGGAGAGCGTCGAAGCGGGGGCGGTGTTCGCCGAGATCGGCCCGCGCATGGCCGAGTT GAGCGGTTCCCGGCTGGCCGCGCAGCAACAGATGGAAGGCCTCCTGGCGCCGCACCGGCCCAA GGAGCCCGCGTGGTTCCTGGCCACCGTCGGCGTCTCGCCCGACCACCAGGGCAAGGGTCTGGG CAGCGCCGTCGTGCTCCCCGGAGTGGAGGCGGCCGAGCGCGCCGGGGTGCCCGCCTTCCTGGA GACCTCCGCGCCCCGCAACCTCCCCTTCTACGAGCGGCTCGGCTTCACCGTCACCGCCGACGTC GAGGTGCCCGAAGGACCGCGCACCTGGTGCATGACCCGCAAGCCCGGTGCCTGACTCGAGCCC CTCTCCCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTT GTCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTGGCC CTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAGGTCTGTT GAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGTCTGTAGCGAC CCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCCAAAAGCCACGTGT ATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGAGTTGGATAGTTGTGGA AAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGAAGGATGCCCAGAAGGTACC CCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCTTTACATGTGTTTAGTCGAGGTTA AAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGTTTTCCTTTGAAAAACACGATGATAAT GGTCGACATGCTGCTGCTGCTGCTGCTGCTGGGCCTGAGGCTACAGCTCTCCCTGGGCATCATC CCAGTTGAGGAGGAGAACCCGGACTTCTGGAACCGCGAGGCAGCCGAGGCCCTGGGTGCCGCC AAGAAGCTGCAGCCTGCACAGACAGCCGCCAAGAACCTCATCATCTTCCTGGGCGATGGGATG GGGGTGTCTACGGTGACAGCTGCCAGGATCCTAAAAGGGCAGAAGAAGGACAAACTGGGGCC TGAGATACCCCTGGCCATGGACCGCTTCCCATATGTGGCTCTGTCCAAGACATACAATGTAGAC AAACATGTGCCAGACAGTGGAGCCACAGCCACGGCCTACCTGTGCGGGGTCAAGGGCAACTTC CAGACCATTGGCTTGAGTGCAGCCGCCCGCTTTAACCAGTGCAACACGACACGCGGCAACGAG GTCATCTCCGTGATGAATCGGGCCAAGAAAGCAGGGAAGTCAGTGGGAGTGGTAACCACCACA CGAGTGCAGCACGCCTCGCCAGCCGGCACCTACGCCCACACGGTGAACCGCAACTGGTACTCG GACGCCGACGTGCCTGCCTCGGCCCGCCAGGAGGGGTGCCAGGACATCGCTACGCAGCTCATC TCCAACATGGACATTGACGTGATCCTAGGTGGAGGCCGAAAGTACATGTTTCGCATGGGAACC CCAGACCCTGAGTACCCAGATGACTACAGCCAAGGTGGGACCAGGCTGGACGGGAAGAATCT GGTGCAGGAATGGCTGGCGAAGCGCCAGGGTGCCCGGTATGTGTGGAACCGCACTGAGCTCAT GCAGGCTTCCCTGGACCCGTCTGTGACCCATCTCATGGGTCTCTTTGAGCCTGGAGACATGAAA TACGAGATCCACCGAGACTCCACACTGGACCCCTCCCTGATGGAGATGACAGAGGCTGCCCTG CGCCTGCTGAGCAGGAACCCCCGCGGCTTCTTCCTCTTCGTGGAGGGTGGTCGCATCGACCATG GTCATCATGAAAGCAGGGCTTACCGGGCACTGACTGAGACGATCATGTTCGACGACGCCATTG AGAGGGCGGGCCAGCTCACCAGCGAGGAGGACACGCTGAGCCTCGTCACTGCCGACCACTCCC ACGTCTTCTCCTTCGGAGGCTACCCCCTGCGAGGGAGCTCCATCTTCGGGCTGGCCCCTGGCAA GGCCCGGGACAGGAAGGCCTACACGGTCCTCCTATACGGAAACGGTCCAGGCTATGTGCTCAA GGACGGCGCCCGGCCGGATGTTACCGAGAGCGAGAGCGGGAGCCCCGAGTATCGGCAGCAGT CAGCAGTGCCCCTGGACGAAGAGACCCACGCAGGCGAGGACGTGGCGGTGTTCGCGCGCGGC CCGCAGGCGCACCTGGTTCACGGCGTGCAGGAGCAGACCTTCATAGCGCACGTCATGGCCTTC GCCGCCTGCCTGGAGCCCTACACCGCCTGCGACCTGGCGCCCCCCGCCGGCACCACCGACGCC GCGCACCCGGGTTACTCTAGAGTCGGGGCGGCCGGCCGCTTCGAGCAGACATGAGGCGCGCCC AGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATG CTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAG TTAACAACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTA AAGCAAGTAAAACCTCTACAAATGTGGTA

B.3. CHAPTER 3 SPECIFIC

Table B4: *GRIK1* homology regions based on GRIK1B gRNA target site

homology	sequence
5' homology	attaagtgaccacaggccatggttctgaaaacaaatttaaaacaaatcaaaaggaaattatgaatttgaaattctgcagtgaggcatgctatgtagttaatactaat gtgtaccactaggcctggaccagtataattccaaatgtatgccttgtctcttccttttaggatcatgatttttccccgtggactattcgaaaaacatctaatacactcca aattcagctctacattttgttcactaatttgttcagccttgcaatccctttcatgcattacaaaatcgctagaccaccacttagactccttctaaccatctgtccatacа aacccctgaaggcaagtcacccctctttttgtgtctcccacaggacctaacgacacaatatcttgacattgtagttactatgcaacaagtgtgtgtcacattcgtc gctacttggtggggtaaatacttgggatgagtgaaccttctttacagttcatctagttgttcagtaatatggaaatgaatgaaaactcagaagaattcagtagaca tagagctgaggagtacaattccagctgaccaatttgcctataaataagaagtctggtaaatattaacgtctcaaggctctttgcactttgcccaccaaagctgtaa gtttatgactatggcctttactcttttggtttgaagcttttgtgattaattgttgggttttctgcattgctcagcaggttttggttttgctggtgtcctgtgtgtttctctgata tttacagactgtttctgttaactaccacaggtttacac
3' homology	ccatgcaaccctgactcagacgtggtggaaaacaatttttactttactaaatagtttctggtttggagttggagctctcatgcagcaaggtacaccggttgcctatct ttgtcacacgcatcccacagtgtgctgacaggctttcatgctggtaaactccaccatgaagctagagggtaggatgaaacaccatctcaggaacttaagaaaat agcaacggactttctagccggttttgaagatgaagcaaaaccaatgacctatatggaactcccttgctatatatctaaaatagcacttctatcaagcaataaatac aaattttcctgtgggtagcaaaagataaggtaccttagaaatcgtttacaaatttgcatagatggagatgttggtcttctccttctcagccataactccacgagtca gcagtttaaaggagcgggatctaaatcctttggagattaaattcactgagattcttccaattgccaaccttacctaggacttccacctaattactactctgcgcaaac cctactgagattcatttcagcctcccagtaacttttcacaatctccctcaattggtaaccatagcaacaataaactgtaggagttggggaaataaaatttggccata gaaatgtaatatattttatacatttcctatgaattcggggattctgcttttgtagtgggggaaggggagctattttaacctttggaagaactccatttctttgtgatttt ttttttcagcttcaatgtcaggcatatggtatagaaaatttttccaggat

Table B5: Target loci for editing/integration

region	sequence
HPRT1	gatgctcacctctcccacaccctttatagtttagggattgtatttccaaggtttctagactgagagccctttcatctttgctcattgacactctgtacccattaatcct ccttattagctccccttcaatggacacatgggtagtcagggtgcaggtctcagaactgtccttcaggttccaggtgatcaaccaagtgccttgtctgtagtgtcaa ctcattgctgccccttcctagtaatccccataatttagctctccatttcatagtctttccttgggtgtgttaaaagtgaccatggtacactcagcacggatgaaatgaa acagtgtttagaaacgtcagtcttctcttttgtaatgccctgtagtctctctgtatgttatatgtcacattttgtaattaacagCTTGCTGGTGAAAAGG ACCCCACGAAGTGTTGGATATAAGCCAGACTgtaagtgaattactttttttgtcaatcatttaaccatctttaacctaaaagagttttat gtgaaatggcttataattgcttagagaatatttgtagagaggcacatttgccagtattagatttaaaagtgatgttttcttttatctaaatgatgaattatgattctttttag TTGTTGGATTTGAAATTCCAGACAAGTTTGTTGTAGGATATGCCCTTGACTATAATGAATACTT CAGGGATTTGAATgtaagtaattgcttctttttctcactcatttttcaaaacacgcataaaaatttaggaaagagaattgttttctccttccagcacctcat aatttgaacagactgatggttcccattagtcacataaagctgtagtctagtacagacgtccttagaactggaacctggccaggctagggtgacacttcttgttgg ctgaaatagttgaacagctttaatatacaataattgttgcattattatttcagatgataaatgtggtcataagtaagaaataaatgatcgagtttagtctttaattcact gtcctttgaatacctgcctcttactctggaggcagaagtcccatggatgtgtttatgaacatggttgaggaagatttaggaagactgcaacagtacactacctaaa gcaggttttttactccatctttttttgccacgtacactggcctcccactttgatatgcttgaaattatctccttgatttgtctttcaaaactacatattgaggctggttgcg g
GRIK1	ggtttgaagcttttgtgattaattgttgggttttctgcattgctcagcaggttttggttttgctggtgtcctgtgtgtttctctgatatttacagactgtttctgttaactacc acaggtttacaccctacgagtggtataaccccccacccatgcaaccctgactcagacgtggtggaaaacaattttacttttactaaatagtttctggtttggagttgg agctctcatgcagcaaggtacaccggttgcctatctttgtcacacgcatcccacagtgtgctgacaggctttcatgctggtaaactccaccatgaagctagagg gtaggatgaaacaccatctcaggaacttaagaaaatagcaacggacttttctagccggttttgaagatgaagcaaaaccaatgacctatatggaactcccttgct atatatctaaaatagcacttctatcaagcaataaatacaaatttttcctgtgggtagcaaaagataaggtaccttagaaatcgtttacaaatttgcatagatggagatg ttggtcttctccttctcagccataactccacgagtcagcagtttaaaggagcgggatctaaatcctttggagattaaattcactgagattcttccaattgccaacctt acctaggacttccacctaattactactctggcaaaccctactgagattcatttcagcctcccagtaacttttcacaatctccctcaattggtaaccatagcaacaata aactgtaggagttggggaaataaaatttggccatagaaatgtaatatattttatacatttcctatgaattcggggattctgcttttgtagtgggggaaggggagct attttaacctttggaagaactccatttc
AAVS1= *PPP1R12C*	TATATTCCCAGGGCCGGTTAATGTGGCTCTGGTTCTGGGTACTTTTATCTGTCCCCTCCACCCCA CAGTGGGGCCACTAGGGACAGGATTGGTGACAGAAAAGCCCCATCCTTAGGCCTCCTCCTTCC TAGTCTCCTGATATTGGGTCTAACCCCCACCTCCTGT
C. griseus Hprt1	gccttcaatgcccggctttatatgtttttcaaagttacataataagagaaagaaaaaagtacactaatggaacatatttctgctttgttttcaaaatgtgttctcttcaa gttactggtccaaagtaccttgtttggtaggaacccagacaattctctaatgttgctcttacctctccagaatatatttttctaatgtgagtttctttttcttttgcagCAT ATTTGTGTCATTAGTGAAACTGGGAAAGCCAAATACAAAGCCTAAgatgagagttcaagttgaatctgcaaacac gaggagtcccattcatgttcccagtaaaattaccaagcattctagttctgcagccatctgcttagtacagctttttgcatgaacccttctaagaattttatggttttttattt ttagaaatgtcagttgctgcattcttaaactttttatttgcactatgagccttcgatagattgtcgcttaccttgtgag

Table B5, continued:

region	sequence
C. griseus Grik1	agaattcagtggatagaggtctgaggagtgccatcctagccactaatgggtttgactctgaagtaagaagctagtgagtatttgcccagtgactttcccaatcca gaagcaattgcttctgctcagttgttgggtttttctgcattgctcagcagcttttcctctggttcatctgtgtatctgatgtttacaaactgtttctgtcaattatcacaggtt tacaccctacgagtggtacaatccccacccgtgcaaccccgactcagacgtggtggaaaacaatttcactttgctaaatagtttctggtttggagttggagctct catgcagcaaggtacaccggttgcccattttgtcacttgcatcccacagtctgctgataagcgttcgtgctggtgaactccaccatgaaggtagatggtaggat gaaacacacaggaatagaaaacaacagcaatagagtttctagccagttctcaaaagaagacagaaatagggtctaaatggagcttctttgccatgaatacactt cttccaaagaatggacacaccatttttcctg

B.4. CHAPTER 4 SPECIFIC

Table B6: Reference enhancer and promoter regions of highly expressed genes in HT1080

source	sequence
CMV regulatory region, accession code M60321: 2105-bp	CTGCAGTGAATAATAAAATGTGTGTTTGTCCGAAATACGCGTTTTGAGATTTCTGTCGCCGACT AAATTCATGTCGCGCGATAGTGGTGTTTATCGCCGATAGAGATGGCGATATTGGAAAAATCGA TATTTGAAAATATGGCATATTGAAAATGTCGCCGATGTGAGTTTCTGTGTAACTGATATCGCCA TTTTTCCAAAAGTGATTTTTGGGCATACGCGATATCTGGCGATACGGCTTATATCGTTTACGGG GGATGGCGATAGACGACTTTGGCGACTTGGGCGATTCTGTGTGTCGCAAATATCGCAGTTTCGA TATAGGTGACAGACGATATGAGGCTATATCGCCGATAGAGGCGACATCAAGCTGGCACATGGC CAATGCATATCGATCTATACATTGAATCAATATTGGCAATTAGCCATATTAGTCATTGGTTATA TAGCATAAATCAATATTGGCTATTGGCCATTGCATACGTTGTATCTATATCATAATATGTACATT TATATTGGCTCATGTCCAATATGACCGCCATGTTGACATTGATTATTGACTAGTTATTAATAGTA ATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTA AATGGCCCGCCTCGTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCC CATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCC CACTTGGCAGTACATCAAGTGTATCATATGCCAAGTCCGGCCCCCTATTGACGTCAATGACGGT AAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTACGGGACTTTCCTACTTGGCAGTACA TCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACACCAATGGGCGTGG ATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTT TGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAATAACCCCGCCCCGTTGACGCAAATG GGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTCGTTTAGTGAACCGTCAGATC GCCTGGAGACGCCATCCACGCTGTTTTGACCTCCATAGAAGACACCGGGACCGATCCAGCCTC CGCGGCCGGGAACGGTGCATTGGAACGCGGATTCCCCGTGCCAAGAGTGACgtaagtaccgcctatagact ctataggcacacccctttggctcttatgcatgctatactgtttttggcttggggcctatacacccccgctccttatgctataggtgatggtatagcttagcctataggt gtgggttattgaccattattgaccactcccctattggtgacgatactttccattactaatccataacatggctctttgccacaactatctctattggctatatgccaata ctctgtccttcagagactgacacggactctgtatttttacaggatggggtcccatttattatttacaaattcacatatacaacaacgccgtccccgtgcccgcagt ttttattaaacatagcgtgggatctccacgcgaatctcgggtacgttccggacatgggctcttctccggtagcggcggagcttccacatccgagccctggtc ccatgcctccagcggctcatggtcgctcggcagctccttgctcctaacagtggaggccagacttaggcacagcacaatgcccaccaccaccagtgtgccgc acaaggccgtggcggtagggtatgtgtctgaaaatgagctcggagattgggctcgcaccgtgacgcagatggaagacttaaggcagcggcagaagaagaat gcaggcagctgagttgttgtattctgataagagtcagaggtaactcccgttgcggtgctgttaacggtggagggcagtgtagtctgagcagtactcgttgctgc cgcgcgcgccaccagacataatagctgacagactaacagactgttcctttccatgggtcttttctgcagTCACCGTCCTTGACACGTTATT GACCATTATTGACCACTCCCCTATTGGTGACGATACTTTCCATTACTAATCCATAACATGGCTCT TTGCCACAACTATCTCTATTGGCTATATGCCAATACTCTGTCCTTCAGAGACTGACACGGACTC TGTATTTTTACAGGATGGGGTCCCATTTATTATTTACAAATTCACATATACAACAACGCCGTCC CCCGTGCCCGCAGTTTTTATTAAACATAGCGTGGGATCTCCACGCGAATCTCGGGTACGTGTTC CGGACATGGGCTCTTCTCCGGTAGCGGCGGAGCTTCCACATCCGAGCCCTGGTCCCATGCCTCC AGCGGCTCATGGTCGCTCGGCAGCTCCTTGCTCCTAACAGTGGAGGCCAGACTTAGGCACAGC ACAATGCCCACCACCACCAGTGTGCCGCACAAGGCCGTGGCGGTAGGGTATGTGTCTGAAAAT GAGCTCGGAGATTGGGCTCGCACCGTGACGCAGATGGAAGACTTAAGGCAGCGGCAGAAGAA GATGCAGGCAGCTGAGTTGTTGTATTCTGATAAGAGTCAGAGGTAACTCCCGTTGCGGTGCTGT TAACGGTGGAGGGCAGTGTAGTCTGAGCAGTACTCGTTGCTGCCGCGCGCGCCACCAGACATA ATAGCTGACAGACTAACAGACTGTTCCTTTCCATGGGTCTTTTCTGCAGTCACCGTCCTTGACA CG

B.5. CHAPTER 5 SPECIFIC

Table B7: Core promoters

variant	sequence
CMV with TSS	AGGTCTATATAAGCAGAGCTCGTTTAGTGAACCGTCAGATC
CMV without TSS	AGGTCTATATAAGCAGAGCTCGTTTAGTGAACCG
UCP variant	gcgcgcctatataagtttgtttcgtttagtgaaccgtcAGATTCTCTGGAGACGCCGAGCCGAGCGGTCAGACCTCCATAGAA

Table B8: Additional promoter and enhancer regions

variant	sequence
EEF1A1. 1048 promoter	GGGGGAGAACCGTATATAAGTGCAGTAGTCGCCGTGAACGTTCTTTTTCGCAACGGGTTTGCCGCCAGAACACAGgtaagtgccgtgtgtggttcccgcgggcctggcctctttacgggttatggcccttgcgtgccttgaattacttccacgcccctggctgcagtacgtgattcttgatcccgagcttcgggttggaagtgggtgggagagttcgaggccttgcgcttaaggagccccttcgcctcgtgcttgagttgaggcctggcttgggcgctggggccgccgcgtgcgaatctggtggcaccttcgcgcctgtctcgctgctttcgataagtctctagccatttaaaatttttgatgacctgctgcgacgctttttttctggcaagatagtcttgtaaatgcgggccaagatctgcacactggtatttcggttttggggccgcgggcggcgacggggcccgtgcgtcccagcgcacatgttcggcgaggcggggcctgcgagcgcggccaccgagaatcggacgggggtagtctcaagctggccggcctgctctggtgcctggcctcgcgccgccgtgtatcgccccgccctgggcggcaaggctggcccggtcggcaccagttgcgtgagcggaaagatggccgcttcccggccctgctgcagggagctcaaaatggaggacgcggcgctcgggagagcgggcgggtgagtcacccacacaaaggaaaagggcctttccgtcctcagccgtcgcttcatgtgactccacggagtaccgggcgccgtccaggcacctcgattagttctcgagcttttggagtacgtcgtctttaggttgggggagggttttatgcgatggagtttccccacactgagtgggtggagactgaagttaggccagcttggcacttgatgtaattctccttggaatttgccctttttgagtttggatcttggttcattctcaagcctcagacagtggttcaaagttttttttcttccatttcagGTGTCGTGAAAACTACCCCTAAAAGCCAAA
ACTB. 1173 promoter	AGGCGGCCAACGCCAAAACTCTCCCTCCTCCTCTTCCTCAATCTCGCTCTCGCTCTTTTTTTTTTTTCGCAAAAGGAGGGGAGAGGGGGTAAAAAAATGCTGCACTGTGCGGCGAAGCCGGTGAGTGAGCGGCGCGGGGCCAATCAGCGTGCGCCGTTCCGAAAGTTGCCTTTTATGGCTCGAGCGGCCGCGGCGCGCCCTATAAAACCCAGCGGCGCGACGCGCCACCACCGCCGAGACCGCGTCCGCCCCGCGAGCACAGAGCCTCGCCTTTGCCGATCCGCCGCCCGTCCACACCCGCCGCCAGgtaagcccggccagccgaccggggcaggcggctcacggcccggccgcaggcggccgcggcccctcgcccgtgcagagccgccgtctgggccgcagcgggggcgcatgggggggaaccggaccgccgtgggggcgcgggagaagcccctgggcctccggagatgggggacaccccacgccagttcggaggcgcgaggccgcgctcgggaggcgcgctccgggggtgccgctctcggggcgggggcaaccggcggggtctttgtctgagccgggctcttgccaatggggatcgcagggtgggcgcggcggagccccgccaggcccggtgggggctgggcgccattgcgcgtgcgcgctggtcctttgggcgctaactgcgtgcgcgctgggaattggcgctaattgcgcgtgcgcgctgggactcaaggcgctaactgcgcgtgcgttctggggcccggggtgccgcggcctgggctggggcgaaggcgggctcggccggaaggggtggggtcgccgcggctcccgggcgcttgcgcgcacttcctgcccgagccgctggccgcccgagggtgtggccgctgcgtgcgcgcgcgccgacccggcgctgtttgaaccgggcggaggcggggctggcgcccggttgggaggggttggggcctggcttcctgccgcgcgccgcggggacgcctccgaccagtgtttgccttttatggtaataacgcggccggcccggcttcctttgtccccaatctgggcgcgcgccggcgcccccctggcggcctaaggactcggcgcgccggaagtggccagggcgggggcgacctcggctcacagcgcgcccggctattctcgcagCTCACC
EIF4A1. 872 extended 5'UTR promoter	GTGGTGGTCTTCCTTAAGGGGCTTCAAATTAGTGGATATGCTTAGCTCAGACCTTCCAGCCAGTCTCTTGAGACTAAAGGGTTCAGCTTTCCATCCCTGGCTCAGGCACTGCCAACACCTTGTCTTCACCCCAAACAAATCCCCGAGATGGGAGCAGAGAGCAGGAAGGAGGGAAAGTAGATAAGCCTCAAGAATAAGGGCATCCGAGAGGGAAGCTGGGGAACTGGACACAAGGGACTGGGGAGGGGACCAACCAGGATTCATGATAGTACCCCAAAGCCCTTTACAGTTTTCTTCCATCCCTCCACCATCCAGCCAGGGGAATCCTCCCATCCCTACGATATCGCTGTTGATTCCTTCATCCCTGGCACACGTCCAGGCAGTGTCGAATCCATCTCTGCTACAGGGGAAAACAAATAACATTTGAGTCCAGTGGAGACCGGGAGCAGAAGTAAAGGGAAGTGATAACCCCCAGAGCCCGGAAGCCTCTGGAGGCTGAGACCTCGCCCCCCTTGCGTGATAGGGCCTACGGAGCCACATGACCAAGGCACTGTCGCCTCCGCACGTGTGAGAGTGCAGGGCCCCAAGATGGCTGCCAGGCCTCGAGGCCTGACTCTTCTATGTCACTTCCGTACCGGCGAGAAAGGCGGGCCCTCCAGCCAATGAGGCTGCGGGGCGGGCCTTCACCTTGATAGGCACTCGAGTTATCCAATGGTGCCTGCGGGCCGGAGCGACTAGGAACTAACGTCATGCCGAGTTGCTGAGCGCCGGCAGGCGGGGCCGGGGCGGCCAAACCAATGCGATGGCCGGGGCGGAGTCGGGCGCTCTATAAGTTGTCGATAGGCGGGCACTCCGCCCTAGTTTCTAAGGATC

Table B8, continued:

variant	sequence
F2R.500.L extended 5'UTR promoter	ACTCAAGGGCCCTTTCTCATTTAGGGGCAACCCTGTCACTACATCATAAACTTTTAAATCCGTG ATCCCCACGTTACAAAAGCAGAAGTCCCTTTTAGACTTTTAGCGAAAACTGAACTTTGCCGGTG TCCCACACGGAGGGAGGGAGGACGGGAGGCCACGCCAGGGCTGCGGGGCTGCAGGGCGTGGA CGCATCCTGGCCGGGGCGTCCACTGTCGACGTCTCCACATCCCAGGAGGGTCGAGACGGCCGC GGGAAGCAGCCTGCGAGCCGTGCGGCCCCATTCCAAGGACCCCGCCAGTGTGAGTCACTGACA GCTTCGCGAATCAACGGTGCCCAGAGGAAAAAACTTCTCATTTGGACTTCTAGGCCCGGCAGT GGCCGGCGGCCAGGGCCCCCCAGTAGGGCAGGGCGGGGCGGGGCGGGGCGCACAGAGCCAGA GGGGCTTGCGAGCGGCGGCTGAGGGACCGCGGGGAGGGGGCGCCGAGCGGCTCCAGCGCAGA GACTCTCACTGCACGCCGGAGGGCGCCCTTCCTCGCTCGCGCCCGCGCGACCGCGCGCCCCAGT CCCGCCCCGCCCCGCTAACCGCCCCAGACACAGCGCTCGCCGAGGGTCGCTTGGACCCTGATCT TACCCGTGGGCACCCTGCGCTCTGCCTGCCGCGAAGACCGGCTCCCCGACCCGCAGAAGTCAG GAGAGAGGGTGAAGCGGAGCAGCCCGAGGCGGGGCAGCCTCCCGGAGCAGCGCCGCGCAGAG CCCGGGACA
LAIR1.500.L extended 5'UTR promoter	AAAAATTTCTTTAAATTGGCCTTTGGAAATTTACCAGCAGTGTGCTGGTAAAGTCTTGACAATC AGCTCTCTGAAAAAAAAAGCAAAAAGAAAAACAAAAAACAACCCCGACGTGTAGCATTTGCC GATTTCTCTGGTGTAAATACTCACAGCATGGCTTTGACATGAGTTTTACATTTGGTAAAAGCAA ATTGTGCCTACTTTGAATAGAAGGATTGGGACAGAGATATGGTTCTTGTCAGGCACTAATTAGG GAGTAAGGCTTGTCTAATATTGCCTTGGCTCTCAAGCAAAATAAAAAAAAAAAAAGTAACGTTT GGGAATCTGTGTTGCTTCCTCAGCCCCATCCTGGGTAAAATCGGAGACGTATACAGGGCAGGG AGAAGCTGTTTATTTCCGTGCCTGCGGCTGGAGCTTCTTAAGGATTTGAGCTGTGATGCTGGCA CCTGGCAGACCACATCCTGTGCGGTTTTCAGTTTTGCTCCGTTCCTGACCCTGGTATAGCAGAA GCTTTTTCACATCTATGACACCCGCTATGTCTTGGTAAACCCTGGAAGGGAAAGGAGGACAAG GTTAAAATACTGTTCCGAGGACCTGGTCTCTCCACAGCGCAGGCTGGAGGTGGCAGCCCGTGG AAAGCCAAGTTCATCCACCATCGGAGCCCAGGCCAGGCTGCCAAGGCTAATATTCAGGACAAA GCCAGGCACAGGTCGGGAATCCTATGAAGATGATCATCGTCCTGAGGTCTTCCTTCCAGGGTTG CATCCGCGACAGAAG
PGK1.500.L extended 5'UTR promoter	GGACCTGGGCCTCTTCCAACTTCTGAGAGGTCTCTATTACTAAGTAAGCCTTAAGAAGCAGAAT TCCATGAAGGGAGCTAGGAAACCAGGATTTTCCAAAAGGAGGTGGCATTTGCATTGATCCTGG TAGGGCAGCCTCGAATTCCACGGGGTTGGGGTTGCGCCTTTTCCAAGGCAGCCCTGGGTTTGCG CAGGGACGCGGCTGCTCTGGGCGTGGTTCCGGGAAACGCAGCGGCGCCGACCCTGGGTCTCGC ACATTCTTCACGTCCGTTCGCAGCGTCACCCGGATCTTCGCCGCTACCCTTGTGGGCCCCCCGG CGACGCTTCCTGCTCCGCCCCTAAGTCGGGAAGGTTCCTTGCGGTTCGCGGCGTGCCGGACGTG ACAAACGGAAGCCGCACGTCTCACTAGTACCCTCGCAGACGGACAGCGCCAGGGAGCAATGG CAGCGCGCCGACCGCGATGGGCTGTGGCCAATAGCGGCTGCTCAGCAGGGCGCGCCGAGAGC AGCGGCCGGGAAGGGGCGGTGCGGGAGGCGGGGTGTGGGGCGGTAGTGTGGGCCCTGTTCCT GCCCGCGCGGTGTTCCGCATTCTGCAAGCCTCCGGAGCGCACGTCGGCAGTCGGCTCCCTCGTT GACCGAATCACCGACCTCTCTCCCCAGCTGTATTTCCAAA
RPL41.398 promoter	ATAGGTGCTGACGTTTAAATAACACAGCGTCCTCATACTAAATCTGGGGGGGGAACTGGTAACT CGAAAACCAAATACTCGGTCTTCCGAAAGAACTAACTCAACCTACCCTTCTACAAGAGGGTCC GAAAACCACTGTTACGCCCATTGGGTAGCCCCGCCCTTGGGGGGGGCAAAGGGCGTGAAAGCG GAAGTGACGACACCCGGCGCTCCATTAAATAGCCGTAGACGGAACTTCGCCTTTCTCTCGGCCT TAGCGCCATTTTTTTGGGTGAGTGTTTTTTGGTTCCTGCGTTGGGATTCCGTGTACAATCCATAG ACATCTGACCTCGGCACTTAGCATCATCACAGCAAACTAACTGTAGCCTTTCTCTCTTTCCCTGT AGAAACCTCTGCGCC
RPLP2.479 promoter	tcacttccggaactgctgcccttcgcctttgccgcgaaaagtgcctgacgggggctgggaaggggggaggggggaaggagtcaacgactccctccccgtcgt cattggctgagcgcctcgtgacgtcactctgaggcgcctcatgcggcgccgcgcacgggtcccagagccttctgggtagcGGTTTAACCCCGC CTCTTGCGTCGGCGCCTTCCTTTTCCTCCCTGTCGCCACCGAGGTCGCACGCGTGAGACTTCTCC GCCGCCTCCGCCGCAGACGCCGCCGCgtgagtgtggtgaccgggcccggggtgccggctggggacgcggagtccgtggggatg cggggtgggcggcggggtattttgggacggaggcctacgggccgagccacgcgcggcctcgcccggcggcccccgggatgggccggcaggaggccg ccggggtaactccgccgtcgcgtcctctccgcccgcctcagG
GAPDH intron 1 with flanking exons	GGCTGGGACTGGCTGAGCCTGGCGGGAGGCGGGGTCCGAGTCACCGCCTGCCGCCGCGCCCCC GGTTTCTATAAATTGAGCCCGCAGCCTCCCGCTTCGCTCTCTGCTCCTCCTGTTCGACAGTCAGC CGCATCTTCTTTTGCGTCGCCAGgtgaagacgggcggagagaaacccgggaggctagggacggcctgaaggcggcaggggcgg gcgcaggccggatgtgttcgcgccgctgcggggtgggcccgggcggcctccgcattgcagggggcgggcggaggacgtgatgcggcgcgggctgggc atggaggcctggtgggggaggggaggggaggcgtgtgtgtcggccggggccactaggcgctcactgttctctccctccgcgcagCCGAGCCAC ATCGCTCAGACACC

Table B8, continued:

variant	sequence
EEF1A1e.308	GAGTAATTCATACAAAAGGACTCGCCCCTGCCTTGGGGAATCCCAGGGACCGTCGTTAAACTC CCACTAACGTAGAACCCAGAGATCGCTGCGTTCCCGCCCCCTCACCCGCCCGCTCTCGTCATCA CTGAGGTGGAGAAGAGCATGCGTGAGGCTCCGGTGCCCGTCAGTGGGCAGAGCGCACATCGCC CACAGTCCCCGAGAAGTTGGGGGGGAGGGGTCGGCAATTGAACCGGTGCCTAGAGAAGGTGGC GCGGGGTAAACTGGGAAAGTGATGTCGTGTACTGGCTCCGCCTTTTTCCCGAGGGT
EEF1A1e.156	GGCTCCGGTGCCCGTCAGTGGGCAGAGCGCACATCGCCCACAGTCCCCGAGAAGTTGGGGGGGA GGGGTCGGCAATTGAACCGGTGCCTAGAGAAGGTGGCGCGGGGTAAACTGGGAAAGTGATGT CGTGTACTGGCTCCGCCTTTTTCCCGAGGGT
EIF4A1e.450	CTGGGGAGGGGACCAACCAGGATTCATGATAGTACCCCAAAGCCCTTTACAGTTTTCTTCCATC CCTCCACCATCCAGCCAGGGGAATCCTCCCATCCCTACGATATCGCTGTTGATTCCTTCATCCCT GGCACACGTCCAGGCAGTGTCGAATCCATCTCTGCTACAGGGGAAAACAAATAACATTTGAGT CCAGTGGAGACCGGGAGCAGAAGTAAAGGGAAGTGATAACCCCCAGAGCCCGGAAGCCTCTG GAGGCTGAGACCTCGCCCCCCTTGCGTGATAGGGCCTACGGAGCCACATGACCAAGGCACTGT CGCCTCCGCACGTGTGAGAGTGCAGGGCCCCAAGATGGCTGCCAGGCCTCGAGGCCTGACTCT TCTATGTCACTTCCGTACCGGCGAGAAAGGCGGGCCCTCCAGCCAATGAGGCTGCGGGGCGGG CCTTCAC
ACTBe.182	AGGCGGCCAACGCCAAAACTCTCCCTCCTCCTCTTCCTCAATCTCGCTCTCGCTCTTTTTTTTTTT CGCAAAAGGAGGGGAGAGGGGGTAAAAAAATGCTGCACTGTGCGGCGAAGCCGGTGAGTGAG CGGCGCGGGGCCAATCAGCGTGCGCCGTTCCGAAAGTTGCCTTTTATGGCTCGA

Table B9: Additional regulatory regions from literature

source	sequence
LMO1 derived[281]: 553-bp	GTAGGGGTTGGAGTTCAGCCTGTTTCCCCTCCAATGTTGTTCCCCCCACATCCTGAGACTTAGG GGTGACCCTGGGTTGAGTGGACTGGTTTATTCTGCTGGGCCCAGCGCATGCATCTGAGTGTGTG CCCAGGCGTGCGTGTCGGCGCAAACATCATCCATTGTGAAATATCAGTGTTTTCATGGGTGAGT AGTAATTACTGGGTAATGCTTTAAAACCTTTCCTGAAGGAGCGCAAAGCCATTTTTTTCTAAAG TCAGGAGTACATTAAAAGGATTACCATGTAGATTTGATTTTTAGATAACACTAAAATGGATCCC AAATGGACTTCAGCAAAGGGATGCTATCTCCTTAATGGAAAGTGCATGGCCCGAGGCTCAGGT CCCAGAGCCAGGCTGgggaaggagggagggaagaggtgtctgcaggggggcaggctggcagattgggtgggggctaggtgggaatgg ggaaggcagagcaggagggagggCCTGGACCCTGTGGGGAGCTTATCCCTCCATCTGGGGAGCAGGAGAC TACAGAGCCCCT
MMse176 mutant[282]: 1742-bp	gagttaaaaagaggtgatttacagtgctatttgagaaggggtattaaggaatttgccagggcactgacgcgtgtcaggtgtaaacctcaggttgagagagagc taagtattttctgtccatgaGggtgataagcgagggcctgaagaaagagggactggggaggacactggcaccagaaataggaaagggcttGttgggggt gggaaggatgggtcaGggtgctatctaGaaagttgCctggcaAtggatgtaggatgtgggagtgagacatcaaacatgaagcagtcgcttaaagtctgaa gaaaCactagatatatGaacTgcaggagatagtagggaaactggaccggctcctcataaaacttcccgccttctatctccgggaggatcgcagggcatttc cgccaagacaggtgagactgcggttctgacctgcgggcctccgtgcatatgcgctagggcacctgggggccggcagagccgttcccctacgcaaagtaag cgtgttatgtctacaacccaaTggggacactgagagccccaaaggccctgctttcttcccagagaacAGCgcccatctgTGtaGtttctacctggctctat gaggtgagaacacaCtccccgctagcacagaaatcctacaaactcctgtgggggctgcgcttggaagcagaggctgtgtaagaggtgactTgggggta gggaaaaacacgaagattttcacacagggtgagaacccaagagactggagaccacggaccaatccctgcaaaaagcagccagggtagaaagggaagag ctgagcggacttcacgatagctaatttgtgttacaaagccgatacggctgatgctcgTtttttctcctatgGcatgcaggcgacatgttacttcctattcccataaa ccctccactgtaggattaacacctaagacaccaaccaagacaaaaaaagatatgacccttggTgtacagtctgtttttgaaactccagaaagtcaggggaaa gcgcgaacgcagtccccactaccacaaattatgcagtcgagtttcccacatttggggaaatcgcaggggtcagcacatccggagtgcaatggataagcctc gcTctgggaaaaccaccttcgtgatcatggtatctcccctgccaggtaagtatgaaacCttgtgcctctgccccgacacagcctcatacgcctcactctttaca cacacggtcacttgccccgcgcactcccgagccctttccagccctgacacacagctgggattctcacttccgatcagcggtcctgaacccgctcccagggca cgggaactcccttcgtggcgaagcagcaagtggcgaagcagcagcctctgcgctgcctcatctacatagaagtcgccctgtccgtgatgtcaccgacagtg ccttgcccagtccccgtctgccttctgccactcaaccgaccaatctgctgccagagCcgccaagGggaagtgacgtctgcctctccctttttccctcccgcc cctgcgtctgttctctcccaaagaagctggtccttagcctgtgttaaggagcaaCctttcggtggccagatggagccggggcatccttcttcaaataatggctttt aattcgcagactagaatgtttcggattacaaaagaaaccggttctcttcacatccttatccttgtgatgcagcattccgcttgcatttggaagccgtttaatatcaga gagaag

B.6. CHAPTER 6 SPECIFIC

Table B10: Base expression cassette for evaluating terminators

construct	sequence (transgene region only)
pEIF4A1.636-hrGFP	CTGGGGAGGGGACCAACCAGGATTCATGATAGTACCCCAAAGCCCTTTACAGTTTTCTTCCATC CCTCCACCATCCAGCCAGGGGAATCCTCCCATCCCTACGATATCGCTGTTGATTCCTTCATCCCT GGCACACGTCCAGGCAGTGTCGAATCCATCTCTGCTACAGGGGAAAACAAATAACATTTGAGT CCAGTGGAGACCGGGAGCAGAAGTAAAGGGAAGTGATAACCCCCAGAGCCCGGAAGCCTCTG GAGGCTGAGACCTCGCCCCCCTTGCGTGATAGGGCCTACGGAGCCACATGACCAAGGCACTGT CGCCTCCGCACGTGTGAGAGTGCAGGGCCCCAAGATGGCTGCCAGGCCTCGAGGCCTGACTCT TCTATGTCACTTCCGTACCGGCGAGAAAGGCGGGCCCTCCAGCCAATGAGGCTGCGGGGCGGG CCTTCACCTTGATAGGCACTCGAGTTATCCAATGGTGCCTGCGGGCCGGAGCGACTAGGAACT AACGTCATGCCGAGTTGCTGAGCGCCGGCAGGCGGGGCCGGGGCGGCCAAACCAATGCGATG GCCGGGGCGGAGTCGGGCGCTCTATAAGTTGTCGATAGGCGGGCACTCCGCCCTAGTTTCTAA GGATCGCTAGCATGGTGAGCAAGCAGATCCTGAAGAACACCGGCCTGCAGGAGATCATGAGCT TCAAGGTGAACCTGGAGGGCGTGGTGAACAACCACGTGTTCACCATGGAGGGCTGCGGCAAGG GCAACATCCTGTTCGGCAACCAGCTGGTGCAGATCCGCGTGACCAAGGGCGCCCCCCTGCCCTT CGCCTTCGACATCCTGAGCCCCGCCTTCCAGTACGGCAACCGCACCTTCACCAAGTACCCCGAG GACATCAGCGACTTCTTCATCCAGAGCTTCCCCGCCGGCTTCGTGTACGAGCGCACCCTGCGCT ACGAGGACGGCGGCCTGGTGGAGATCCGCAGCGACATCAACCTGATCGAGGAGATGTTCGTGT ACCGCGTGGAGTACAAGGGCCGCAACTTCCCCAACGACGGCCCCGTGATGAAGAAGACCATCA CCGGCCTGCAGCCCAGCTTCGAGGTGGTGTACATGAACGACGGCGTGCTGGTGGGCCAGGTGA TCCTGGTGTACCGCCTGAACAGCGGCAAGTTCTACAGCTGCCACATGCGCACCCTGATGAAGA GCAAGGGCGTGGTGAAGGACTTCCCCGAGTACCACTTCATCCAGCACCGCCTGGAGAAGACCT ACGTGGAGGACGGCGGCTTCGTGGAGCAGCACGAGACCGCCATCGCCCAGCTGACCAGCCTGG GCAAGCCCCTGGGCAGCCTGCACGAGTGGGTGTAAGGCGCGCC
pEIF4A1.636-SEAP	CTGGGGAGGGGACCAACCAGGATTCATGATAGTACCCCAAAGCCCTTTACAGTTTTCTTCCATC CCTCCACCATCCAGCCAGGGGAATCCTCCCATCCCTACGATATCGCTGTTGATTCCTTCATCCCT GGCACACGTCCAGGCAGTGTCGAATCCATCTCTGCTACAGGGGAAAACAAATAACATTTGAGT CCAGTGGAGACCGGGAGCAGAAGTAAAGGGAAGTGATAACCCCCAGAGCCCGGAAGCCTCTG GAGGCTGAGACCTCGCCCCCCTTGCGTGATAGGGCCTACGGAGCCACATGACCAAGGCACTGT CGCCTCCGCACGTGTGAGAGTGCAGGGCCCCAAGATGGCTGCCAGGCCTCGAGGCCTGACTCT TCTATGTCACTTCCGTACCGGCGAGAAAGGCGGGCCCTCCAGCCAATGAGGCTGCGGGGCGGG CCTTCACCTTGATAGGCACTCGAGTTATCCAATGGTGCCTGCGGGCCGGAGCGACTAGGAACT AACGTCATGCCGAGTTGCTGAGCGCCGGCAGGCGGGGCCGGGGCGGCCAAACCAATGCGATG GCCGGGGCGGAGTCGGGCGCTCTATAAGTTGTCGATAGGCGGGCACTCCGCCCTAGTTTCTAA GGATCGCTAGCATGCTGCTGCTGCTGCTGCTGCTGGGCCTGAGGCTACAGCTCTCCCTGGGCAT CATCCCAGTTGAGGAGGAGAACCCGGACTTCTGGAACCGCGAGGCAGCCGAGGCCCTGGGTGC CGCCAAGAAGCTGCAGCCTGCACAGACAGCCGCCAAGAACCTCATCATCTTCCTGGGCGATGG GATGGGGGTGTCTACGGTGACAGCTGCCAGGATCCTAAAAGGGCAGAAGAAGGACAAACTGG GGCCTGAGATACCCCTGGCCATGGACCGCTTCCCATATGTGGCTCTGTCCAAGACATACAATGT AGACAAACATGTGCCAGACAGTGGAGCCACAGCCACGGCCTACCTGTGCGGGGTCAAGGGCA ACTTCCAGACCATTGGCTTGAGTGCAGCCGCCCGCTTTAACCAGTGCAACACGACACGCGGCA ACGAGGTCATCTCCGTGATGAATCGGGCCAAGAAAGCAGGGAAGTCAGTGGGAGTGGTAACC ACCACACGAGTGCAGCACGCCTCGCCAGCCGGCACCTACGCCCACACGGTGAACCGCAACTGG TACTCGGACGCCGACGTGCCTGCCTCGGCCCGCCAGGAGGGGTGCCAGGACATCGCTACGCAG CTCATCTCCAACATGGACATTGACGTGATCCTAGGTGGAGGCCGAAAGTACATGTTTCGCATGG GAACCCCAGACCCTGAGTACCCAGATGACTACAGCCAAGGTGGGACCAGGCTGGACGGGAAG AATCTGGTGCAGGAATGGCTGGCGAAGCGCCAGGGTGCCCGGTATGTGTGGAACCGCACTGAG CTCATGCAGGCTTCCCTGGACCCGTCTGTGACCCATCTCATGGGTCTCTTTGAGCCTGGAGACA TGAAATACGAGATCCACCGAGACTCCACACTGGACCCCTCCCTGATGGAGATGACAGAGGCTG CCCTGCGCCTGCTGAGCAGGAACCCCCGCGGCTTCTTCCTCTTCGTGGAGGGTGGTCGCATCGA CCATGGTCATCATGAAAGCAGGGCTTACCGGGCACTGACTGAGACGATCATGTTCGACGACGC CATTGAGAGGGCGGGCCAGCTCACCAGCGAGGAGGACACGCTGAGCCTCGTCACTGCCGACCA CTCCCACGTCTTCTCCTTCGGAGGCTACCCCCTGCGAGGGAGCTCCATCTTCGGGCTGGCCCCT GGCAAGGCCCGGGACAGGAAGGCCTACACGGTCCTCCTATACGGAAACGGTCCAGGCTATGTG CTCAAGGACGGCGCCCGGCCGGATGTTACCGAGAGCGAGAGCGGGAGCCCCGAGTATCGGCA GCAGTCAGCAGTGCCCCTGGACGAAGAGACCCACGCAGGCGAGGACGTGGCGGTGTTCGCGC GCGGCCCGCAGGCGCACCTGGTTCACGGCGTGCAGGAGCAGACCTTCATAGCGCACGTCATGG CCTTCGCCGCCTGCCTGGAGCCCTACACCGCCTGCGACCTGGCGCCCCCCGCCGGCACCACCGA CGCCGCGCACCCGGGTTACTCTAGAGTCGGGGCGGCCGGCCGCTTCGAGCAGACATGA

Table B10, continued:

construct	sequence (transgene region only)
pEEF1A1.1356-hrGFP	GAGTAATTCATACAAAAGGACTCGCCCCTGCCTTGGGGAATCCCAGGGACCGTCGTTAAACTC CCACTAACGTAGAACCCAGAGATCGCTGCGTTCCCGCCCCCTCACCCGCCCGCTCTCGTCATCA CTGAGGTGGAGAAGAGCATGCGTGAGGCTCCGGTGCCCGTCAGTGGGCAGAGCGCACATCGCC CACAGTCCCCGAGAAGTTGGGGGGGAGGGGTCGGCAATTGAACCGGTGCCTAGAGAAGGTGGC GCGGGGTAAACTGGGAAAGTGATGTCGTGTACTGGCTCCGCCTTTTTCCCGAGGGTGGGGGAG AACCGTATATAAGTGCAGTAGTCGCCGTGAACGTTCTTTTTCGCAACGGGTTTGCCGCCAGAAC ACAGgtaagtgccgtgtgtggttcccgcgggcctggcctctttacgggttatggcccttgcgtgccttgaattacttccacgcccctggctgcagtacgtgat tcttgatcccgagcttcgggttggaagtgggtgggagagttcgaggccttgcgcttaaggagccccttcgcctcgtgcttgagttgaggcctggcttgggcgc tggggccgccgcgtgcgaatctggtggcaccttcgcgcctgtctcgctgctttcgataagtctctagccatttaaaatttttgatgacctgctgcgacgctttttttc tggcaagatagtcttgtaaatgcgggccaagatctgcacactggtatttcggtttttgggccgcgggcggcgacggggcccgtgcgtcccagcgcacatgt tcggcgaggcggggcctgcgagcgcggccaccgagaatcggacgggggtagtctcaagctggccggcctgctctggtgcctggcctcgcgccgccgtg tatcgccccgccctgggcggcaaggctggcccggtcggcaccagttgcgtgagcggaaagatggccgcttcccggccctgctgcagggagctcaaaatg gaggacgcggcgctcgggagagcgggcgggtgagtcacccacacaaaggaaaagggcctttccgtcctcagccgtcgcttcatgtgactccacggagta ccgggcgccgtccaggcacctcgattagttctcgagcttttggagtacgtcgtctttaggttgggggagggttttatgcgatggagtttccccacactgagtg ggtggagactgaagttaggccagcttggcacttgatgtaattctccttggaatttgccctttttgagtttggatcttggttcattctcaagcctcagacagtggttcaa agtttttttcttccatttcagGTGTCGTGAAAACTACCCCTAAAAGCCAAAGCTAGCATGGTGAGCAAGCAGA TCCTGAAGAACACCGGCCTGCAGGAGATCATGAGCTTCAAGGTGAACCTGGAGGGCGTGGTGA ACAACCACGTGTTCACCATGGAGGGCTGCGGCAAGGGCAACATCCTGTTCGGCAACCAGCTGG TGCAGATCCGCGTGACCAAGGGCGCCCCCCTGCCCTTCGCCTTCGACATCCTGAGCCCCGCCTT CCAGTACGGCAACCGCACCTTCACCAAGTACCCCGAGGACATCAGCGACTTCTTCATCCAGAG CTTCCCCGCCGGCTTCGTGTACGAGCGCACCCTGCGCTACGAGGACGGCGGCCTGGTGGAGAT CCGCAGCGACATCAACCTGATCGAGGAGATGTTCGTGTACCGCGTGGAGTACAAGGGCCGCAA CTTCCCCAACGACGGCCCCGTGATGAAGAAGACCATCACCGGCCTGCAGCCCAGCTTCGAGGT GGTGTACATGAACGACGGCGTGCTGGTGGGCCAGGTGATCCTGGTGTACCGCCTGAACAGCGG CAAGTTCTACAGCTGCCACATGCGCACCCTGATGAAGAGCAAGGGCGTGGTGAAGGACTTCCC CGAGTACCACTTCATCCAGCACCGCCTGGAGAAGACCTACGTGGAGGACGGCGGCTTCGTGGA GCAGCACGAGACCGCCATCGCCCAGCTGACCAGCCTGGGCAAGCCCCTGGGCAGCCTGCACGA GTGGGTGTAAGGCGCGC
fully viral construct: rpCMV.660-hrGFP-SV40pA	GACCGCCATGTTGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGT TCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTCGTGACCGC CCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGA CTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGT GTATCATATGCCAAGTCCGGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTA TGCCCAGTACATGACCTTACGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCT ATTACCATGGTGATGCGGTTTTGGCAGTACACCAATGGGCGTGGATAGCGGTTTGACTCACGG GGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGG ACTTTCCAAAATGTCGTAATAACCCCGCCCCGTTGACGCAAATGGGCGGTAGGCGTGTACGGT GGGAGGTCTATATAAGCAGAGCTCGTTTAGTGAACCGTCAGATCGCCTGGAGACGCCATCCAC GCTGTTTTGACCTCCATAGAAGACGCTAGCATGGTGAGCAAGCAGATCCTGAAGAACACCGGC CTGCAGGAGATCATGAGCTTCAAGGTGAACCTGGAGGGCGTGGTGAACAACCACGTGTTCACC ATGGAGGGCTGCGGCAAGGGCAACATCCTGTTCGGCAACCAGCTGGTGCAGATCCGCGTGACC AAGGGCGCCCCCCTGCCCTTCGCCTTCGACATCCTGAGCCCCGCCTTCCAGTACGGCAACCGCA CCTTCACCAAGTACCCCGAGGACATCAGCGACTTCTTCATCCAGAGCTTCCCCGCCGGCTTCGT GTACGAGCGCACCCTGCGCTACGAGGACGGCGGCCTGGTGGAGATCCGCAGCGACATCAACCT GATCGAGGAGATGTTCGTGTACCGCGTGGAGTACAAGGGCCGCAACTTCCCCAACGACGGCCC CGTGATGAAGAAGACCATCACCGGCCTGCAGCCCAGCTTCGAGGTGGTGTACATGAACGACGG CGTGCTGGTGGGCCAGGTGATCCTGGTGTACCGCCTGAACAGCGGCAAGTTCTACAGCTGCCA CATGCGCACCCTGATGAAGAGCAAGGGCGTGGTGAAGGACTTCCCCGAGTACCACTTCATCCA GCACCGCCTGGAGAAGACCTACGTGGAGGACGGCGGCTTCGTGGAGCAGCACGAGACCGCCA TCGCCCAGCTGACCAGCCTGGGCAAGCCCCTGGGCAGCCTGCACGAGTGGGTGTAACAGACAT GATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTAT TTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAAC AACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCA AGTAAAACCTCTACAAATGTGGTA

References

[1] Aggarwal, R. S. (2014) What's fueling the biotech engine-2012 to 2013, *Nat Biotechnol 32*, 32-39.

[2] (2014) New drug costs soar to [dollar]2.6 billion, *Nat Biotech 32*, 1176-1176.

[3] Tycko, J., Myer, V. E., and Hsu, P. D. (2016) Methods for Optimizing CRISPR-Cas9 Genome Editing Specificity, *Mol Cell 63*, 355-370.

[4] Wood, A. J., Lo, T.-W., Zeitler, B., Pickle, C. S., Ralston, E. J., Lee, A. H., Amora, R., Miller, J. C., Leung, E., Meng, X., Zhang, L., Rebar, E. J., Gregory, P. D., Urnov, F. D., and Meyer, B. J. (2011) Targeted genome editing across species using ZFNs and TALENs, *Science 333*, 307-307.

[5] Brown, A. J., Sweeney, B., Mainwaring, D. O., and James, D. C. (2014) Synthetic promoters for CHO cell engineering, *Biotechnology and bioengineering 111*, 1638-1647.

[6] Baik, J. Y., Gasimli, L., Yang, B., Datta, P., Zhang, F., Glass, C. a., Esko, J. D., Linhardt, R. J., and Sharfstein, S. T. (2012) Metabolic engineering of Chinese hamster ovary cells: towards a bioengineered heparin, *Metab Eng 14*, 81-90.

[7] Wurm, F. M. (2004) Production of recombinant protein therapeutics in cultivated mammalian cells, *Nat Biotechnol 22*, 1393-1398.

[8] Dietmair, S., Nielsen, L. K., and Timmins, N. E. (2011) Mammalian cells as biopharmaceutical production hosts in the age of omics, *Biotechnol J 7*, 75-89.

[9] Kantardjieff, A., Nissom, P. M., Chuah, S. H., Yusufi, F., Jacob, N. M., Mulukutla, B. C., Yap, M., and Hu, W. S. (2009) Developing genomic platforms for Chinese hamster ovary cells, *Biotechnol Adv 27*, 1028-1035.

[10] Keasling, J. D. (2010) Manufacturing molecules through metabolic engineering, *Science 330*, 1355-1358.

[11] Curran, K. A., and Alper, H. S. (2012) Expanding the chemical palate of cells by combining systems biology and metabolic engineering, *Metab Eng 14*, 289-297.

[12] Cost, G. J., Freyvert, Y., Vafiadis, A., Santiago, Y., Miller, J. C., Rebar, E., Collingwood, T. N., Snowden, A., and Gregory, P. D. (2009) BAK and BAX deletion using zinc-finger nucleases yields apoptosis-resistant CHO cells, *Biotechnology and bioengineering 105*, 330-340.

[13] Druz, A., Chu, C., Majors, B., Santuary, R., Betenbaugh, M., and Shiloach, J. (2011) A Novel MicroRNA mmu-miR-466h Affects Apoptosis Regulation in Mammalian Cells, *Biotechnology and bioengineering 108*, 1651-1661.

[14] Koh, T. C., Lee, Y. Y., Chang, S. Q., and Nissom, P. M. (2009) Identification and expression analysis of miRNAs during batch culture of HEK-293 cells, *Journal of biotechnology 140*, 149-155.

[15] Mulukutla, B. C., Khan, S., Lange, A., and Hu, W. S. (2010) Glucose metabolism in mammalian cell culture: new insights for tweaking vintage pathways, *Trends Biotechnol 28*, 476-484.

[16] Chen, K., Liu, Q., Xie, L., Sharp, P. A., and Wang, D. I. (2001) Engineering of a mammalian cell line for reduction of lactate formation and high monoclonal antibody production, *Biotechnology and bioengineering 72*, 55-61.

[17] Jeon, M. K., Yu da, Y., and Lee, G. M. (2011) Combinatorial engineering of ldh-a and bcl-2 for reducing lactate production and improving cell growth in dihydrofolate reductase-deficient Chinese hamster ovary cells, *Appl Microbiol Biotechnol 92*, 779-790.

[18] Dorai, H., Kyung, Y. S., Ellis, D., Kinney, C., Lin, C., Jan, D., Moore, G., and Betenbaugh, M. J. (2009) Expression of anti-apoptosis genes alters lactate metabolism of Chinese Hamster Ovary cells in culture, *Biotechnology and bioengineering 103*, 592-608.

[19] Fan, L., Kadura, I., Krebs, L. E., Hatfield, C. C., Shaw, M. M., and Frye, C. C. (2012) Improving the efficiency of CHO cell line generation using glutamine synthetase gene knockout cells, *Biotechnology and bioengineering 109*, 1007-1015.

[20] Dreesen, I. A., and Fussenegger, M. (2011) Ectopic expression of human mTOR increases viability, robustness, cell size, proliferation, and antibody production of chinese hamster ovary cells, *Biotechnology and bioengineering 108*, 853-866.

[21] Cornelissen, L. A., de Vries, R. P., de Boer-Luijtze, E. A., Rigter, A., Rottier, P. J., and de Haan, C. A. (2010) A single immunization with soluble recombinant trimeric hemagglutinin protects chickens against highly pathogenic avian influenza virus H5N1, *PLoS ONE 5*, e10645.

[22] Bosch, B. J., Bodewes, R., de Vries, R. P., Kreijtz, J. H., Bartelink, W., van Amerongen, G., Rimmelzwaan, G. F., de Haan, C. A., Osterhaus, A. D., and Rottier, P. J. (2010) Recombinant soluble, multimeric HA and NA exhibit distinctive types of protection against pandemic swine-origin 2009 A(H1N1) influenza virus infection in ferrets, *J Virol 84*, 10366-10374.

[23] Jacobs, P. P., and Callewaert, N. (2009) N-glycosylation engineering of biopharmaceutical expression systems, *Curr Mol Med 9*, 774-800.

[24] Dietmair, S., Nielsen, L. K., and Timmins, N. E. (2011) Engineering a mammalian super producer, *Journal of Chemical Technology & Biotechnology 86*, 905-914.

[25] Lanza, A. M., Kim do, S., and Alper, H. S. (2013) Evaluating the influence of selection markers on obtaining selected pools and stable cell lines in human cells, *Biotechnol J 8*, 811-821.

[26] Cheng, J. K., and Alper, H. S. (2014) The genome editing toolbox: a spectrum of approaches for targeted modification, *Curr Opin Biotechnol 30*, 87-94.

[27] Gersbach, C. A., Gaj, T., Gordley, R. M., Mercer, A. C., and Barbas, C. F., 3rd. (2011) Targeted plasmid integration into the human genome by an engineered zinc-finger recombinase, *Nucleic Acids Res 39*, 7868-7878.

[28] Sun, N., Abil, Z., and Zhao, H. (2012) Recent advances in targeted genome engineering in mammalian systems, *Biotechnol J*.

[29] Kramer, O., Klausing, S., and Noll, T. (2010) Methods in mammalian cell line engineering: from random mutagenesis to sequence-specific approaches, *Appl Microbiol Biotechnol 88*, 425-436.

[30] Silva, G., Poirot, L., Galetto, R., Smith, J., Montoya, G., Duchateau, P., and Paques, F. (2010) Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy, *Curr Gene Ther 11*, 11-27.

[31] Kennard, M. L., Goosney, D. L., Monteith, D., Zhang, L., Moffat, M., Fischer, D., and Mott, J. (2009) The generation of stable, high MAb expressing CHO cell lines based on the artificial chromosome expression (ACE) technology, *Biotechnology and bioengineering 104*, 540-553.

[32] Kim, J. Y., Kim, Y. G., and Lee, G. M. (2011) CHO cells in biotechnology for production of recombinant proteins: current state and further potential, *Appl Microbiol Biotechnol 93*, 917-930.

[33] Rita Costa, A., Elisa Rodrigues, M., Henriques, M., Azeredo, J., and Oliveira, R. (2009) Guidelines to cell engineering for monoclonal antibody production, *Eur J Pharm Biopharm 74*, 127-138.

[34] Ivics, Z., Li, M. A., Mates, L., Boeke, J. D., Nagy, A., Bradley, A., and Izsvak, Z. (2009) Transposon-mediated genome manipulation in vertebrates, *Nat Methods 6*, 415-422.

[35] Owens, J. B., Mauro, D., Stoytchev, I., Bhakta, M. S., Kim, M. S., Segal, D. J., and Moisyadi, S. (2013) Transcription activator like effector (TALE)-directed piggyBac transposition in human cells, *Nucleic Acids Res 41*, 9197-9207.

[36] Belcher, J. D., Vineyard, J. V., Bruzzone, C. M., Chen, C., Beckman, J. D., Nguyen, J., Steer, C. J., and Vercellotti, G. M. (2010) Heme oxygenase-1 gene delivery by Sleeping Beauty inhibits vascular stasis in a murine model of sickle cell disease, *J Mol Med (Berl) 88*, 665-675.

[37] Hackett, P. B., Largaespada, D. A., and Cooper, L. J. (2010) A transposon and transposase system for human application, *Mol Ther 18*, 674-683.

[38] Orban, T. I., Apati, A., Nemeth, A., Varga, N., Krizsik, V., Schamberger, A., Szebenyi, K., Erdei, Z., Varady, G., Karaszi, E., Homolya, L., Nemet, K., Gocza, E., Miskey, C., Mates, L., Ivics, Z., Izsvak, Z., and Sarkadi, B. (2009) Applying a "double-feature" promoter to identify cardiomyocytes differentiated from human embryonic stem cells following transposon-based gene delivery, *Stem Cells 27*, 1077-1087.

[39] Rad, R., Rad, L., Wang, W., Cadinanos, J., Vassiliou, G., Rice, S., Campos, L. S., Yusa, K., Banerjee, R., Li, M. A., de la Rosa, J., Strong, A., Lu, D., Ellis, P., Conte, N., Yang, F. T., Liu, P., and Bradley, A. (2010) PiggyBac transposon mutagenesis: a tool for cancer gene discovery in mice, *Science 330*, 1104-1107.

[40] Zhou, H., Liu, Z. G., Sun, Z. W., Huang, Y., and Yu, W. Y. (2010) 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*, 122-129.

[41] Lanza, A. M., Dyess, T. J., and Alper, H. S. (2012) Using the Cre/lox system for targeted integration into the human genome: loxFAS-loxP pairing and delayed introduction of Cre DNA improve gene swapping efficiency, *Biotechnol J 7*, 898-908.

[42] Patsch, C., Peitz, M., Otte, D. M., Kesseler, D., Jungverdorben, J., Wunderlich, F. T., Brustle, O., Zimmer, A., and Edenhofer, F. (2010) Engineering cell-permeant FLP recombinase for tightly controlled inducible and reversible overexpression in embryonic stem cells, *Stem Cells 28*, 894-902.

[43] Davis, M. W., Morton, J. J., Carroll, D., and Jorgensen, E. M. (2008) Gene activation using FLP recombinase in C. elegans, *PLoS Genet 4*, e1000028.

[44] Iida, Y., Kazuki, Y., Hayashi, M., Ueda, Y., Hasegawa, M., Kouprina, N., Larionov, V., and Oshimura, M. (2014) Bi-HAC Vector System toward Gene and Cell Therapy, *ACS Synthetic Biology*, 140109092736004.

[45] Enyeart, P. J., Chirieleison, S. M., Dao, M. N., Perutka, J., Quandt, E. M., Yao, J., Whitt, J. T., Keatinge-Clay, A. T., Lambowitz, A. M., and Ellington, A. D. (2013) Generalized bacterial genome editing using mobile group II introns and Cre-lox, *Mol Syst Biol 9*, 685.

[46] Mastroianni, M., Watanabe, K., White, T. B., Zhuang, F., Vernon, J., Matsuura, M., Wallingford, J., and Lambowitz, A. M. (2008) Group II intron-based gene targeting reactions in eukaryotes, *PLoS One 3*, e3121.

[47] Shibata, Y., Kumar, P., Layer, R., Willcox, S., Gagan, J. R., Griffith, J. D., and Dutta, A. (2012) Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues, *Science 336*, 82-86.

[48] Backliwal, G., Hildinger, M., Chenuet, S., Wulhfard, S., De Jesus, M., and Wurm, F. M. (2008) Rational vector design and multi-pathway modulation of HEK 293E cells yield recombinant antibody titers exceeding 1 g/l by transient transfection under serum-free conditions, *Nucleic Acids Res 36*, e96.

[49] Codamo, J., Hou, J. J. C., Hughes, B. S., Gray, P. P., and Munro, T. P. (2011) Efficient mAb production in CHO cells incorporating PEI-mediated transfection, mild hypothermia and the co-expression of XBP-1, *Journal of Chemical Technology & Biotechnology 86*, 923-934.

[50] Codamo, J., Munro, T. P., Hughes, B. S., Song, M., and Gray, P. P. (2011) Enhanced CHO cell-based transient gene expression with the epi-CHO expression system, *Mol Biotechnol 48*, 109-115.

[51] Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S., and Gregory, P. D. (2010) Genome editing with engineered zinc finger nucleases, *Nature Reviews Genetics 11*, 636-646.

[52] Liu, P.-Q., Chan, E. M., Cost, G. J., Zhang, L., Wang, J., Miller, J. C., Guschin, D. Y., Reik, A., Holmes, M. C., Mott, J. E., Collingwood, T. N., and Gregory, P. D. (2010) Generation of a triple-gene knockout mammalian cell line using engineered zinc-finger nucleases, *Biotechnology and bioengineering 106*, 97-105.

[53] Malphettes, L., Freyvert, Y., Chang, J., Liu, P. Q., Chan, E., Miller, J. C., Zhou, Z., Nguyen, T., Tsai, C., Snowden, A. W., Collingwood, T. N., Gregory, P. D., and Cost, G. J. (2010) Highly efficient deletion of FUT8 in CHO cell lines using zinc-finger nucleases yields cells that produce completely nonfucosylated antibodies, *Biotechnology and bioengineering 106*, 774-783.

[54] Kim, H. J., Lee, H. J., Kim, H., Cho, S. W., and Kim, J. S. (2009) Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly, *Genome Res 19*, 1279-1288.

[55] Sander, J. D., Dahlborg, E. J., Goodwin, M. J., Cade, L., Zhang, F., Cifuentes, D., Curtin, S. J., Blackburn, J. S., Thibodeau-Beganny, S., Qi, Y., Pierick, C. J., Hoffman, E., Maeder, M. L., Khayter, C., Reyon, D., Dobbs, D., Langenau, D. M., Stupar, R. M., Giraldez, A. J., Voytas, D. F., Peterson, R. T., Yeh, J. R., and Joung, J. K. (2011) Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA), *Nat Methods 8*, 67-69.

[56] Pattanayak, V., Ramirez, C. L., Joung, J. K., and Liu, D. R. (2011) Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection, *Nat Methods 8*, 765-770.

[57] Gabriel, R., Lombardo, A., Arens, A., Miller, J. C., Genovese, P., Kaeppel, C., Nowrouzi, A., Bartholomae, C. C., Wang, J., Friedman, G., Holmes, M. C., Gregory, P. D., Glimm, H., Schmidt, M., Naldini, L., and von Kalle, C. (2011) An unbiased genome-wide analysis of zinc-finger nuclease specificity, *Nat Biotechnol 29*, 816-823.

[58] Cradick, T. J., Fine, E. J., Antico, C. J., and Bao, G. (2013) CRISPR/Cas9 systems targeting beta-globin and CCR5 genes have substantial off-target activity, *Nucleic Acids Res 41*, 9584-9592.

[59] Mussolino, C., and Cathomen, T. (2012) TALE nucleases: tailored genome engineering made easy, *Curr Opin Biotechnol 23*, 644-650.

[60] Briggs, A. W., Rios, X., Chari, R., Yang, L., Zhang, F., Mali, P., and Church, G. M. (2012) Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers, *Nucleic Acids Res 40*, e117.

[61] Miller, J. C., Tan, S., Qiao, G., Barlow, K. A., Wang, J., Xia, D. F., Meng, X., Paschon, D. E., Leung, E., Hinkley, S. J., Dulay, G. P., Hua, K. L., Ankoudinova, I., Cost, G. J., Urnov, F. D., Zhang, H. S., Holmes, M. C., Zhang, L., Gregory, P. D., and Rebar, E. J. (2010) A TALE nuclease architecture for efficient genome editing, *Nat Biotechnol 29*, 143-148.

[62] Hockemeyer, D., Wang, H., Kiani, S., Lai, C. S., Gao, Q., Cassady, J. P., Cost, G. J., Zhang, L., Santiago, Y., Miller, J. C., Zeitler, B., Cherone, J. M., Meng, X., Hinkley, S. J., Rebar, E. J., Gregory, P. D., Urnov, F. D., and Jaenisch, R. (2011) Genetic engineering of human pluripotent cells using TALE nucleases, *Nat Biotechnol 29*, 731-734.

[63] Mussolino, C., Morbitzer, R., Lutge, F., Dannemann, N., Lahaye, T., and Cathomen, T. (2011) A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity, *Nucleic Acids Res 39*, 9283-9293.

[64] Hou, Z., Zhang, Y., Propson, N. E., Howden, S. E., Chu, L. F., Sontheimer, E. J., and Thomson, J. A. (2013) Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis, *Proc Natl Acad Sci U S A 110*, 15644-15649.

[65] Ma, N., Liao, B., Zhang, H., Wang, L., Shan, Y., Xue, Y., Huang, K., Chen, S., Zhou, X., Chen, Y., Pei, D., and Pan, G. (2013) Transcription activator-like effector nuclease (TALEN)-mediated gene correction in integration-free beta-thalassemia induced pluripotent stem cells, *The Journal of biological chemistry 288*, 34671-34679.

[66] Ding, Q., Lee, Y. K., Schaefer, E. A., Peters, D. T., Veres, A., Kim, K., Kuperwasser, N., Motola, D. L., Meissner, T. B., Hendriks, W. T., Trevisan, M., Gupta, R. M., Moisan, A., Banks, E., Friesen, M., Schinzel, R. T., Xia, F., Tang, A., Xia, Y., Figueroa, E., Wann, A., Ahfeldt, T., Daheron, L., Zhang, F., Rubin, L. L., Peng, L. F., Chung, R. T., Musunuru, K., and Cowan, C. A. (2013) A TALEN genome-editing system for generating human stem cell-based disease models, *Cell Stem Cell 12*, 238-251.

[67] Tesson, L., Usal, C., Menoret, S., Leung, E., Niles, B. J., Remy, S., Santiago, Y., Vincent, A. I., Meng, X., Zhang, L., Gregory, P. D., Anegon, I., and Cost, G. J. (2011) Knockout rats generated by embryo microinjection of TALENs, *Nat Biotechnol 29*, 695-696.

[68] Miller, J. C., Tan, S., Qiao, G., Barlow, K. A., Wang, J., Xia, D. F., Meng, X., Paschon, D. E., Leung, E., Hinkley, S. J., Dulay, G. P., Hua, K. L., Ankoudinova, I., Cost, G. J., Urnov, F. D., Zhang, H. S., Holmes, M. C., Zhang, L., Gregory, P. D., and Rebar, E. J. (2011) A TALE nuclease architecture for efficient genome editing, *Nat Biotechnol 29*, 143-148.

[69] Cermak, T., Doyle, E. L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J. A., Somia, N. V., Bogdanove, A. J., and Voytas, D. F. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting, *Nucleic Acids Res 39*, e82.

[70] Reyon, D., Tsai, S. Q., Khayter, C., Foden, J. A., Sander, J. D., and Joung, J. K. (2012) FLASH assembly of TALENs for high-throughput genome editing, *Nat Biotechnol 30*, 460-465.

[71] Heigwer, F., Kerr, G., Walther, N., Glaeser, K., Pelz, O., Breinig, M., and Boutros, M. (2013) E-TALEN: a web tool to design TALENs for genome engineering, *Nucleic Acids Res 41*, e190.

[72] Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., and Sander, J. D. (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells, *Nat Biotechnol 31*, 822-826.

[73] Fujii, W., Kawasaki, K., Sugiura, K., and Naito, K. (2013) Efficient generation of large-scale genome-modified mice using gRNA and CAS9 endonuclease, *Nucleic Acids Res 41*, e187.

[74] Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., and Zhang, F. (2013) Multiplex genome engineering using CRISPR/Cas systems, *Science 339*, 819-823.

[75] Esvelt, K. M., Mali, P., Braff, J. L., Moosburner, M., Yaung, S. J., and Church, G. M. (2013) Orthogonal Cas9 proteins for RNA-guided gene regulation and editing, *Nat Methods 10*, 1116-1121.

[76] Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., Lim, W. A., Weissman, J. S., and Qi, L. S. (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes, *Cell 154*, 442-451.

[77] Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S., and Kim, J. S. (2014) Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases, *Genome Res 24*, 132-141.

[78] Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G. W., Park, J., Blackburn, E. H., Weissman, J. S., Qi, L. S., and Huang, B. (2013) Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System, *Cell 155*, 1479-1491.

[79] Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., and Lim, W. A. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression, *Cell 152*, 1173-1183.

[80] Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., and Zhang, F. (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells, *Science 343*, 84-87.

[81] Maeder, M. L., Linder, S. J., Cascio, V. M., Fu, Y., Ho, Q. H., and Joung, J. K. (2013) CRISPR RNA-guided activation of endogenous human genes, *Nat Methods 10*, 977-979.

[82] Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G., and Zhang, F. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases, *Nat Biotechnol 31*, 827-832.

[83] Mali, P., Aach, J., Stranges, P. B., Esvelt, K. M., Moosburner, M., Kosuri, S., Yang, L., and Church, G. M. (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering, *Nat Biotechnol 31*, 833-838.

[84] Ran, F. A., Hsu, P. D., Lin, C. Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., Scott, D. A., Inoue, A., Matoba, S., Zhang, Y., and Zhang, F. (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity, *Cell 154*, 1380-1389.

[85] Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system, *Science 343*, 80-84.

[86] Konermann, S., Brigham, M. D., Trevino, A. E., Hsu, P. D., Heidenreich, M., Cong, L., Platt, R. J., Scott, D. A., Church, G. M., and Zhang, F. (2013) Optical control of mammalian endogenous transcription and epigenetic states, *Nature 500*, 472-476.

[87] Polstein, L. R., and Gersbach, C. A. (2012) Light-inducible spatiotemporal control of gene activation by customizable zinc finger transcription factors, *J Am Chem Soc 134*, 16480-16483.

[88] Kim, Y. K., Wee, G., Park, J., Kim, J., Baek, D., Kim, J. S., and Kim, V. N. (2013) TALEN-based knockout library for human microRNAs, *Nat Struct Mol Biol 20*, 1458-1464.
[89] Mendenhall, E. M., Williamson, K. E., Reyon, D., Zou, J. Y., Ram, O., Joung, J. K., and Bernstein, B. E. (2013) Locus-specific editing of histone modifications at endogenous enhancers, *Nat Biotechnol 31*, 1133-1136.
[90] Kearns, N. A., Pham, H., Tabak, B., Genga, R. M., Silverstein, N. J., Garber, M., and Maehr, R. (2015) Functional annotation of native enhancers with a Cas9-histone demethylase fusion, *Nat Meth advance on*.
[91] Certo, M. T., Ryu, B. Y., Annis, J. E., Garibov, M., Jarjour, J., Rawlings, D. J., and Scharenberg, A. M. (2011) Tracking genome engineering outcome at individual DNA breakpoints, *Nat Methods 8*, 671-676.
[92] Doyon, Y., Vo, T. D., Mendel, M. C., Greenberg, S. G., Wang, J., Xia, D. F., Miller, J. C., Urnov, F. D., Gregory, P. D., and Holmes, M. C. (2011) Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures, *Nat Methods 8*, 74-79.
[93] Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. A., Yan, W. X., and Zhang, F. (2015) Rationally engineered Cas9 nucleases with improved specificity, *Science 351*, 84-88.
[94] Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z., and Joung, J. K. (2016) High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects, *Nature 529*, 490-495.
[95] Davis, K. M., Pattanayak, V., Thompson, D. B., Zuris, J. A., and Liu, D. R. (2015) Small molecule–triggered Cas9 protein with improved genome-editing specificity, *Nat Chem Biol 11*, 316-318.
[96] Hagedorn, C., Antoniou, M. N., and Lipps, H. J. (2013) Genomic cis-acting Sequences Improve Expression and Establishment of a Nonviral Vector, *Mol Ther Nucleic Acids 2*, e118.
[97] Truong, D.-J. J., Kühner, K., Kühn, R., Werfel, S., Engelhardt, S., Wurst, W., and Ortiz, O. (2015) Development of an intein-mediated split–Cas9 system for gene therapy, *Nucleic Acids Res 43*, 6450-6458.
[98] Chew, W. L., Tabebordbar, M., Cheng, J. K. W., Mali, P., Wu, E. Y., Ng, A. H. M., Zhu, K., Wagers, A. J., and Church, G. M. (2016) A multifunctional AAV-CRISPR-Cas9 and its host response, *Nat Meth 13*, 868-874.
[99] Gwiazda, K. S., Grier, A. E., Sahni, J., Burleigh, S. M., Martin, U., Yang, J. G., Popp, N. A., Krutein, M. C., Khan, I. F., Jacoby, K., Jensen, M. C., Rawlings, D. J., and Scharenberg, A. M. (2016) High Efficiency CRISPR/Cas9-mediated Gene Editing in Primary Human T-cells Using Mutant Adenoviral E4orf6/E1b55k [ldquo]Helper[rdquo] Proteins, *Mol Ther 24*, 1570-1580.
[100] Cheong, T.-C., Compagno, M., and Chiarle, R. (2016) Editing of mouse and human immunoglobulin genes by CRISPR-Cas9 system, *Nat Commun 7*.
[101] Young, E., and Alper, H. (2010) Synthetic biology: tools to design, build, and optimize cellular processes, *J Biomed Biotechnol 2010*, 130781.

[102] Khalil, A. S., and Collins, J. J. (2010) Synthetic biology: applications come of age, *Nat Rev Genet 11*, 367-379.
[103] Leonard, E., Nielsen, D., Solomon, K., and Prather, K. J. (2008) Engineering microbes with synthetic biology frameworks, *Trends Biotechnol 26*, 674-681.
[104] Seo, S., Kim, S., and Jung, G. (2012) Synthetic regulatory tools for microbial engineering, *Biotechnology and Bioprocess Engineering 17*, 1-7.
[105] Siddiqui, M. S., Thodey, K., Trenchard, I., and Smolke, C. D. (2011) Advancing secondary metabolite biosynthesis in yeast with synthetic biology tools, *FEMS Yeast Res 12*, 144-170.
[106] Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Jr., Kinney, J. B., Kellis, M., Lander, E. S., and Mikkelsen, T. S. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay, *Nat Biotechnol 30*, 271-277.
[107] Schlabach, M. R., Hu, J. K., Li, M., and Elledge, S. J. (2010) Synthetic design of strong promoters, *Proc Natl Acad Sci U S A 107*, 2538-2543.
[108] Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., Lee, C., Andrie, J. M., Lee, S. I., Cooper, G. M., Ahituv, N., Pennacchio, L. A., and Shendure, J. (2012) Massively parallel functional dissection of mammalian enhancers in vivo, *Nat Biotechnol 30*, 265-270.
[109] Xu, X., Nagarajan, H., Lewis, N. E., Pan, S., Cai, Z., Liu, X., Chen, W., Xie, M., Wang, W., Hammond, S., Andersen, M. R., Neff, N., Passarelli, B., Koh, W., Fan, H. C., Wang, J., Gui, Y., Lee, K. H., Betenbaugh, M. J., Quake, S. R., Famili, I., and Palsson, B. O. (2011) The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line, *Nat Biotechnol 29*, 735-741.
[110] Alper, H., Fischer, C., Nevoigt, E., and Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering, *Proc Natl Acad Sci U S A 102*, 12678-12683.
[111] Blazeck, J., Liu, L., Redden, H., and Alper, H. (2012) Tuning Gene Expression in Yarrowia lipolytica by a Hybrid Promoter Approach, *Appl Environ Microbiol 77*, 7905-7914.
[112] Nevoigt, E., Kohnke, J., Fischer, C. R., Alper, H., Stahl, U., and Stephanopoulos, G. (2006) Engineering of promoter replacement cassettes for fine-tuning of gene expression in Saccharomyces cerevisiae, *Appl Environ Microbiol 72*, 5266-5273.
[113] Ferreira, J., Peacock, R., Lawhorn, I., and Wang, C. (2012) Modulating ectopic gene expression levels by using retroviral vectors equipped with synthetic promoters, *Systems and Synthetic Biology 5*, 131-138.
[114] Ferreira, J. P., Lawhorn, I. E., Peacock, R. W., and Wang, C. L. (2012) Quantitative assessment of Ras over-expression via shotgun deployment of vectors utilizing synthetic promoters, *Integr Biol (Camb) 4*, 108-114.
[115] Magnusson, T., Haase, R., Schleef, M., Wagner, E., and Ogris, M. (2011) Sustained, high transgene expression in liver with plasmid vectors using optimized promoter-enhancer combinations, *J Gene Med 13*, 382-391.

[116] Tigges, M., and Fussenegger, M. (2009) Recent advances in mammalian synthetic biology-design of synthetic transgene control networks, *Curr Opin Biotechnol 20*, 449-460.
[117] Mullick, A., Xu, Y., Warren, R., Koutroumanis, M., Guilbault, C., Broussau, S., Malenfant, F., Bourget, L., Lamoureux, L., Lo, R., Caron, A. W., Pilotte, A., and Massie, B. (2006) The cumate gene-switch: a system for regulated expression in mammalian cells, *BMC Biotechnol 6*, 43.
[118] Auslander, D., Wieland, M., Auslander, S., Tigges, M., and Fussenegger, M. (2011) Rational design of a small molecule-responsive intramer controlling transgene expression in mammalian cells, *Nucleic Acids Res 39*, e155.
[119] Wieland, M., Auslander, D., and Fussenegger, M. (2012) Engineering of ribozyme-based riboswitches for mammalian cells, *Methods*.
[120] Weber, W., Lienhart, C., Baba, M. D., and Fussenegger, M. (2009) A biotin-triggered genetic switch in mammalian cells and mice, *Metab Eng 11*, 117-124.
[121] Ho, S. C., Bardor, M., Feng, H., Mariati, Tong, Y. W., Song, Z., Yap, M. G., and Yang, Y. (2011) IRES-mediated Tricistronic vectors for enhancing generation of high monoclonal antibody expressing CHO cell lines, *Journal of biotechnology 157*, 130-139.
[122] Mariati, Ho, S. C., Yap, M. G., and Yang, Y. (2009) Evaluating post-transcriptional regulatory elements for enhancing transient gene expression levels in CHO K1 and HEK293 cells, *Protein Expr Purif 69*, 9-15.
[123] Sun, J., Li, D., Hao, Y., Zhang, Y., Fan, W., Fu, J., Hu, Y., Liu, Y., and Shao, Y. (2009) Posttranscriptional regulatory elements enhance antigen expression and DNA vaccine efficacy, *DNA Cell Biol 28*, 233-240.
[124] Zhang, F., Frost, A. R., Blundell, M. P., Bales, O., Antoniou, M. N., and Thrasher, A. J. (2010) A ubiquitous chromatin opening element (UCOE) confers resistance to DNA methylation-mediated silencing of lentiviral vectors, *Mol Ther 18*, 1640-1649.
[125] Boscolo, S., Mion, F., Licciulli, M., Macor, P., De Maso, L., Brce, M., Antoniou, M. N., Marzari, R., Santoro, C., and Sblattero, D. (2012) Simple scale-up of recombinant antibody production using an UCOE containing vector, *N Biotechnol*.
[126] Nair, A. R., Jinger, X., and Hermiston, T. W. (2011) Effect of different UCOE-promoter combinations in creation of engineered cell lines for the production of Factor VIII, *BMC Res Notes 4*, 178.
[127] Haase, R., Argyros, O., Wong, S. P., Harbottle, R. P., Lipps, H. J., Ogris, M., Magnusson, T., Vizoso Pinto, M. G., Haas, J., and Baiker, A. (2010) pEPito: a significantly improved non-viral episomal expression vector for mammalian cells, *BMC Biotechnol 10*, 20.
[128] Araki, Y., Hamafuji, T., Noguchi, C., and Shimizu, N. (2012) Efficient Recombinant Production in Mammalian Cells Using a Novel IR/MAR Gene Amplification Method, *PLoS ONE 7*, e41787.

[129] Lin, N., Davis, A., Bahr, S., Borgschulte, T., Achtien, K., and Kayser, K. (2011) Profiling highly conserved microrna expression in recombinant IgG-producing and parental chinese hamster ovary cells, *Biotechnol Prog*.
[130] Barron, N., Kumar, N., Sanchez, N., Doolan, P., Clarke, C., Meleady, P., O'Sullivan, F., and Clynes, M. (2011) Engineering CHO cell growth and recombinant protein productivity by overexpression of miR-7, *Journal of biotechnology 151*, 204-211.
[131] Johnson, K. C., Jacob, N. M., Nissom, P. M., Hackl, M., Lee, L. H., Yap, M., and Hu, W. S. (2011) Conserved microRNAs in Chinese hamster ovary cell lines, *Biotechnology and bioengineering 108*, 475-480.
[132] Daboussi, F., Zaslavskiy, M., Poirot, L., Loperfido, M., Gouble, A., Guyot, V., Leduc, S., Galetto, R., Grizot, S., Oficjalska, D., Perez, C., Delacote, F., Dupuy, A., Chion-Sotinel, I., Le Clerre, D., Lebuhotel, C., Danos, O., Lemaire, F., Oussedik, K., Cedrone, F., Epinat, J. C., Smith, J., Yanez-Munoz, R. J., Dickson, G., Popplewell, L., Koo, T., Vandendriessche, T., Chuah, M. K., Duclert, A., Duchateau, P., and Paques, F. (2012) Chromosomal context and epigenetic mechanisms control the efficacy of genome editing by rare-cutting designer endonucleases, *Nucleic Acids Res 40*, 6367-6379.
[133] Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L. A. (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems, *Nat Biotechnol 31*, 233-239.
[134] Sun, J., and Alper, H. S. (2015) Metabolic engineering of strains: from industrial-scale to lab-scale chemical production, *J Ind Microbiol Biotechnol 42*, 423-436.
[135] Swiech, K., Picanco-Castro, V., and Covas, D. T. (2012) Human cells: new platform for recombinant therapeutic protein production, *Protein Expr Purif 84*, 147-153.
[136] Ellis, T., Adie, T., and Baldwin, G. S. (2011) DNA assembly for synthetic biology: from parts to pathways and beyond, *Integr Biol (Camb) 3*, 109-118.
[137] Chusainow, J., Yang, Y. S., Yeo, J. H., Toh, P. C., Asvadi, P., Wong, N. S., and Yap, M. G. (2009) A study of monoclonal antibody-producing CHO cell lines: what makes a stable high producer?, *Biotechnology and bioengineering 102*, 1182-1196.
[138] Seth, G., Charaniya, S., Wlaschin, K. F., and Hu, W. S. (2007) In pursuit of a super producer-alternative paths to high producing recombinant mammalian cells, *Curr Opin Biotechnol 18*, 557-564.
[139] Porter, A. J., Racher, A. J., Preziosi, R., and Dickson, A. J. (2010) Strategies for selecting recombinant CHO cell lines for cGMP manufacturing: improving the efficiency of cell line generation, *Biotechnol Prog 26*, 1455-1464.
[140] Nakano, M., Odaka, K., Ishimura, M., Kondo, S., Tachikawa, N., Chiba, J., Kanegae, Y., and Saito, I. (2001) Efficient gene activation in cultured mammalian cells mediated by FLP recombinase-expressing recombinant adenovirus, *Nucleic Acids Res 29*, E40.
[141] Chalberg, T. W., Portlock, J. L., Olivares, E. C., Thyagarajan, B., Kirby, P. J., Hillman, R. T., Hoelters, J., and Calos, M. P. (2006) Integration specificity of phage phiC31 integrase in the human genome, *J Mol Biol 357*, 28-48.

[142] Thyagarajan, B., Olivares, E. C., Hollis, R. P., Ginsburg, D. S., and Calos, M. P. (2001) Site-specific genomic integration in mammalian cells mediated by phage phiC31 integrase, *Mol Cell Biol 21*, 3926-3934.
[143] Pruett-Miller, S. M., Connelly, J. P., Maeder, M. L., Joung, J. K., and Porteus, M. H. (2008) Comparison of zinc finger nucleases for use in gene targeting in mammalian cells, *Mol Ther 16*, 707-717.
[144] Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., and Church, G. M. (2013) RNA-guided human genome engineering via Cas9, *Science 339*, 823-826.
[145] Bandaranayake, A. D., and Almo, S. C. (2014) Recent advances in mammalian protein production, *FEBS Letters 588*, 253-260.
[146] Wang, J., Exline, C. M., DeClercq, J. J., Llewellyn, G. N., Hayward, S. B., Li, P. W.-L., Shivak, D. A., Surosky, R. T., Gregory, P. D., Holmes, M. C., and Cannon, P. M. (2015) Homology-driven genome editing in hematopoietic stem and progenitor cells using ZFN mRNA and AAV6 donors, *Nat Biotech 33*, 1256-1263.
[147] Wurm, F. M., and Petropoulos, C. J. (1994) Plasmid integration, amplification and cytogenetics in CHO cells: questions and comments, *Biologicals 22*, 95-102.
[148] Wurm, F. M. (1990) Integration, amplification and stability of plasmid sequences in CHO cell cultures, *Biologicals 18*, 159-164.
[149] Liang, Z., Breman, A. M., Grimes, B. R., and Rosen, E. D. (2008) Identifying and genotyping transgene integration loci, *Transgenic Res 17*, 979-983.
[150] Gierman, H. J., Indemans, M. H., Koster, J., Goetze, S., Seppen, J., Geerts, D., van Driel, R., and Versteeg, R. (2007) Domain-wide regulation of gene expression in the human genome, *Genome Res 17*, 1286-1295.
[151] Francia, M. V., and Garcia Lobo, J. M. (1996) Gene integration in the Escherichia coli chromosome mediated by Tn21 integrase (Int21), *J Bacteriol 178*, 894-898.
[152] Flagfeldt, D. B., Siewers, V., Huang, L., and Nielsen, J. (2009) Characterization of chromosomal integration sites for heterologous gene expression in Saccharomyces cerevisiae, *Yeast 26*, 545-551.
[153] Liu, W. Y., Wang, Y., Qin, Y., Wang, Y. P., and Zhu, Z. Y. (2007) Site-directed gene integration in transgenic zebrafish mediated by cre recombinase using a combination of mutant lox sites, *Mar Biotechnol (NY) 9*, 420-428.
[154] Sadelain, M., Papapetrou, E. P., and Bushman, F. D. (2012) Safe harbours for the integration of new DNA in the human genome, *Nat Rev Cancer 12*, 51-58.
[155] Eyquem, J., Poirot, L., Galetto, R., Scharenberg, A. M., and Smith, J. (2013) Characterization of three loci for homologous gene targeting and transgene expression, *Biotechnology and bioengineering 110*, 2225-2235.
[156] Ramachandra, C. J., Shahbazi, M., Kwang, T. W., Choudhury, Y., Bak, X. Y., Yang, J., and Wang, S. (2011) Efficient recombinase-mediated cassette exchange at the AAVS1 locus in human embryonic stem cells using baculoviral vectors, *Nucleic Acids Res 39*, e107.
[157] Sakurai, K., Shimoji, M., Tahimic, C. G., Aiba, K., Kawase, E., Hasegawa, K., Amagai, Y., Suemori, H., and Nakatsuji, N. (2010) Efficient integration of

transgenes into a defined locus in human embryonic stem cells, *Nucleic Acids Res 38*, e96.
[158] Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. (2008) A common sequence motif associated with recombination hot spots and genome instability in humans, *Nat Genet 40*, 1124-1129.
[159] Ikeda, R., Kokubu, C., Yusa, K., Keng, V. W., Horie, K., and Takeda, J. (2007) Sleeping beauty transposase has an affinity for heterochromatin conformation, *Mol Cell Biol 27*, 1665-1676.
[160] Esteller, M. (2007) Cancer epigenomics: DNA methylomes and histone-modification maps, *Nat Rev Genet 8*, 286-298.
[161] Mielke, C., Maass, K., Tummler, M., and Bode, J. (1996) Anatomy of highly expressing chromosomal sites targeted by retroviral vectors, *Biochemistry 35*, 2239-2252.
[162] Mielke, C., Maass, K., Tummler, M., and Bode, J. (1996) Anatomy of Highly Expressing Chromosomal Sites Targeted by Retroviral Vectors, *Biochemistry 35*, 2239-2252.
[163] Mitchell, R. S., Beitzel, B. F., Schroder, A. R., Shinn, P., Chen, H., Berry, C. C., Ecker, J. R., and Bushman, F. D. (2004) Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences, *Plos Biol 2*, E234.
[164] Schroder, A. R., Shinn, P., Chen, H., Berry, C., Ecker, J. R., and Bushman, F. (2002) HIV-1 integration in the human genome favors active genes and local hotspots, *Cell 110*, 521-529.
[165] Rosin, F. M., Watanabe, N., Cacas, J. L., Kato, N., Arroyo, J. M., Fang, Y., May, B., Vaughn, M., Simorowski, J., Ramu, U., McCombie, R. W., Spector, D. L., Martienssen, R. A., and Lam, E. (2008) Genome-wide transposon tagging reveals location-dependent effects on transcription and chromatin organization in Arabidopsis, *Plant J 55*, 514-525.
[166] Ozsolak, F., Song, J. S., Liu, X. S., and Fisher, D. E. (2007) High-throughput mapping of the chromatin structure of human promoters, *Nat Biotechnol 25*, 244-248.
[167] Feuk, L., Carson, A. R., and Scherer, S. W. (2006) Structural variation in the human genome, *Nat Rev Genet 7*, 85-97.
[168] Derse, D., Crise, B., Li, Y., Princler, G., Lum, N., Stewart, C., McGrath, C. F., Hughes, S. H., Munroe, D. J., and Wu, X. (2007) Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses, *J Virol 81*, 6731-6741.
[169] Lewinski, M. K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., Crawford, G., Collins, F., Shinn, P., Leipzig, J., Hannenhalli, S., Berry, C. C., Ecker, J. R., and Bushman, F. D. (2006) Retroviral DNA integration: viral and cellular determinants of target-site selection, *PLoS Pathog 2*, e60.
[170] Moalic, Y., Felix, H., Takeuchi, Y., Jestin, A., and Blanchard, Y. (2009) Genome areas with high gene density and CpG island neighborhood strongly attract porcine

endogenous retrovirus for integration and favor the formation of hot spots, *J Virol 83*, 1920-1929.

[171] Wu, X., Li, Y., Crise, B., and Burgess, S. M. (2003) Transcription start regions in the human genome are favored targets for MLV integration, *Science 300*, 1749-1751.

[172] Heuze-Vourc'h, N., Ainciburu, M., Planque, C., Brillard-Bourdet, M., Ott, C., Jolivet-Reynaud, C., and Courty, Y. (2006) Recombinant kallikrein expression: site-specific integration for hK6 production in human cells, *Biol Chem 387*, 687-695.

[173] Wiberg, F. C., Rasmussen, S. K., Frandsen, T. P., Rasmussen, L. K., Tengbjerg, K., Coljee, V. W., Sharon, J., Yang, C. Y., Bregenholt, S., Nielsen, L. S., Haurum, J. S., and Tolstrup, A. B. (2006) Production of target-specific recombinant human polyclonal antibodies in mammalian cells, *Biotechnology and bioengineering 94*, 396-405.

[174] Kawabe, Y., Makitsubo, H., Kameyama, Y., Huang, S., Ito, A., and Kamihira, M. (2012) Repeated integration of antibody genes into a pre-selected chromosomal locus of CHO cells using an accumulative site-specific gene integration system, *Cytotechnology 64*, 267-279.

[175] Cheng, J. K., Lewis, A. M., Kim, D. S., Dyess, T., and Alper, H. S. (2016) Identifying and retargeting transcriptional hot spots in the human genome, *Biotechnology Journal 11*, 1100-1109.

[176] Xu, X., Nagarajan, H., Lewis, N. E., Pan, S., Cai, Z., Liu, X., Chen, W., Xie, M., Wang, W., Hammond, S., Andersen, M. R., Neff, N., Passarelli, B., Koh, W., Fan, H. C., Wang, J., Gui, Y., Lee, K. H., Betenbaugh, M. J., Quake, S. R., Famili, I., Palsson, B. O., and Wang, J. (2011) The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line, *Nat Biotechnol 29*, 735-741.

[177] Wurm, F. M., and Hacker, D. (2011) First CHO genome, *Nat Biotechnol 29*, 718-720.

[178] Wurtele, H., Little, K. C., and Chartrand, P. (2003) Illegitimate DNA integration in mammalian cells, *Gene Ther 10*, 1791-1799.

[179] Lee, J. S., Kallehauge, T. B., Pedersen, L. E., and Kildegaard, H. F. (2015) Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway, *Sci Rep 5*, 8572.

[180] Bachu, R., Bergareche, I., and Chasin, L. A. (2015) CRISPR-Cas targeted plasmid integration into mammalian cells via non-homologous end joining, *Biotechnology and bioengineering 112*, 2154-2162.

[181] Liao, S., Tammaro, M., and Yan, H. (2015) Enriching CRISPR-Cas9 targeted cells by co-targeting the HPRT gene, *Nucleic Acids Res 43*, e134-e134.

[182] Lieber, M. R. (2010) The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway, *Annu Rev Biochem 79*, 181-211.

[183] Mao, Z., Bozzella, M., Seluanov, A., and Gorbunova, V. (2008) Comparison of nonhomologous end joining and homologous recombination in human cells, *DNA Repair 7*, 1765-1771.

[184] Taleei, R., and Nikjoo, H. (2013) Biochemical DSB-repair model for mammalian cells in G1 and early S phases of the cell cycle, *Mutation Research 756*, 206-212.
[185] Lai, T., Yang, Y., and Ng, S. K. (2013) Advances in Mammalian Cell Line Development Technologies for Recombinant Protein Production, *Pharmaceuticals 6*, 579-579.
[186] Huang, Y., Li, Y., Wang, Y. G., Gu, X., Wang, Y., and Shen, B. F. (2007) An efficient and targeted gene integration system for high-level antibody expression, *J Immunol Methods 322*, 28-39.
[187] Lieu, P. T., Machleidt, T., Thyagarajan, B., Fontes, A., Frey, E., Fuerstenau-Sharp, M., Thompson, D. V., Swamilingiah, G. M., Derebail, S. S., Piper, D., and Chesnut, J. D. (2009) Generation of Site-Specific Retargeting Platform Cell Lines for Drug Discovery Using phiC31 and R4 Integrases, *J Biomol Screen 14*, 1207-1215.
[188] Kawabe, Y., Shimomura, T., Huang, S., Imanishi, S., Ito, A., and Kamihira, M. (2016) Targeted transgene insertion into the CHO cell genome using Cre recombinase-incorporating integrase-defective retroviral vectors, *Biotechnology and bioengineering 113*, 1600-1610.
[189] Oberbek, A., Matasci, M., Hacker, D. L., and Wurm, F. M. (2011) Generation of stable, high-producing cho cell lines by lentiviral vector-mediated gene transfer in serum-free suspension culture, *Biotechnology and bioengineering 108*, 600-610.
[190] Howes, R., and Schofield, C. (2015) Genome Engineering Using Adeno-Associated Virus (AAV), pp 75-103.
[191] Bibikova, M., Beumer, K., Trautman, J. K., and Carroll, D. (2003) Enhancing Gene Targeting with Designed Zinc Finger Nucleases, *Science 300*, 764 LP-764.
[192] Orlando, S. J., Santiago, Y., DeKelver, R. C., Freyvert, Y., Boydston, E. A., Moehle, E. A., Choi, V. M., Gopalan, S. M., Lou, J. F., Li, J., Miller, J. C., Holmes, M. C., Gregory, P. D., Urnov, F. D., and Cost, G. J. (2010) Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology, *Nucleic Acids Res 38*, e152-e152.
[193] Lombardo, A., Cesana, D., Genovese, P., Di Stefano, B., Provasi, E., Colombo, D. F., Neri, M., Magnani, Z., Cantore, A., Lo Riso, P., Damo, M., Pello, O. M., Holmes, M. C., Gregory, P. D., Gritti, A., Broccoli, V., Bonini, C., and Naldini, L. (2011) Site-specific integration and tailoring of cassette design for sustainable gene transfer, *Nat Meth 8*, 861-869.
[194] Walsh, R. M., and Hochedlinger, K. (2013) A variant CRISPR-Cas9 system adds versatility to genome engineering, *Proc Natl Acad Sci 110*, 15514-15515.
[195] Cho, S. W., Kim, S., Kim, J. M., and Kim, J.-S. (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease, *Nat Biotechnol 31*, 230-232.
[196] Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S., and Kim, J.-S. (2013) Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases, *Genome Res 24*, 132-141.

[197] Pattanayak, V., Lin, S., Guilinger, J. P., Ma, E., Doudna, J. a., and Liu, D. R. (2013) High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity, *Nat Biotechnol 31*, 839-843.

[198] Kuscu, C., Arslan, S., Singh, R., Thorpe, J., and Adli, M. (2014) Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease, *Nat Biotechnol 32*, 677-683.

[199] Wright, A. V., Sternberg, S. H., Taylor, D. W., Staahl, B. T., Bardales, J. A., Kornfeld, J. E., and Doudna, J. A. (2015) Rational design of a split-Cas9 enzyme complex, *Proc Natl Acad Sci 112*, 2984-2989.

[200] Lee, C. M., Cradick, T. J., and Bao, G. (2016) The Neisseria meningitidis CRISPR-Cas9 System Enables Specific Genome Editing in Mammalian Cells, *Mol Ther 24*, 645-654.

[201] van Overbeek, M., Capurso, D., Carter, M. M., Thompson, M. S., Frias, E., Russ, C., Reece-Hoyes, J. S., Nye, C., Gradia, S., Vidal, B., Zheng, J., Hoffman, G. R., Fuller, C. K., and May, A. P. (2016) DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks, *Mol Cell 63*, 633-646.

[202] Chu, V. T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., and Kuhn, R. (2015) Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells, *Nat Biotechnol 33*, 543-548.

[203] Wang, J., DeClercq, J. J., Hayward, S. B., Li, P. W.-L., Shivak, D. A., Gregory, P. D., Lee, G., and Holmes, M. C. (2016) Highly efficient homology-driven genome editing in human T cells by combining zinc-finger nuclease mRNA and AAV6 donor delivery, *Nucleic Acids Res 44*, e30-e30.

[204] Blokland, H. J. M., Hoeksema, F., Siep, M., Otte, A. P., and Verhees, J. A. (2011) Methods to create a stringent selection system for mammalian cell lines, *Cytotechnology 63*, 371-384.

[205] Bochkov, Y., and Palmenberg, A. (2006) Translational efficiency of EMCV IRES in bicistronic vectors is dependent upon IRES sequence and gene location, *BioTechniques 41*, 283-292.

[206] Mullen, C. A., Kilstrup, M., and Blaese, R. M. (1992) Transfer of the bacterial gene for cytosine deaminase to mammalian cells confers lethal sensitivity to 5-fluorocytosine: a negative selection system, *Proc Natl Acad Sci 89*, 33-37.

[207] Kievit, E., Bershad, E., Ng, E., Sethna, P., Dev, I., Lawrence, T. S., and Rehemtulla, A. (1999) Superiority of Yeast over Bacterial Cytosine Deaminase for Enzyme/Prodrug Gene Therapy in Colon Cancer Xenografts, *Cancer Research 59*, 1417-1421.

[208] Li, J., Huang, S., Chen, J., Yang, Z., Fei, X., Zheng, M., Ji, C., Xie, Y., and Mao, Y. (2007) Identification and characterization of human uracil phosphoribosyltransferase (UPRTase), *J Hum Genet 52*, 415-422.

[209] O'Brien, T. A., Tuong, D. T., Basso, L. M., McIvor, R. S., and Orchard, P. J. (2006) Coexpression of the Uracil Phosphoribosyltransferase Gene with a Chimeric Human Nerve Growth Factor Receptor/Cytosine Deaminase Fusion Gene, Using a

Single Retroviral Vector, Augments Cytotoxicity of Transduced Human T Cells Exposed to 5-Fluorocytosine, *Hum Gene Ther 17*, 518-530.
[210] Mock, U., Hauber, I., and Fehse, B. (2016) Digital PCR to assess gene-editing frequencies (GEF-dPCR) mediated by designer nucleases, *Nat Protocols 11*, 598-615.
[211] Kelley, B. (2009) Industrialization of mAb production technology The bioprocessing industry at a crossroads, *Mabs 1*, 443-452.
[212] Pybus, L. P., Dean, G., West, N. R., Smith, A., Daramola, O., Field, R., Wilkinson, S. J., and James, D. C. (2014) Model-directed engineering of "difficult-to-express" monoclonal antibody production by Chinese hamster ovary cells, *Biotechnology and bioengineering 111*, 372-385.
[213] Le Fourn, V., Girod, P. A., Buceta, M., Regamey, A., and Mermod, N. (2014) CHO cell engineering to prevent polypeptide aggregation and improve therapeutic protein secretion, *Metab Eng 21*, 91-102.
[214] Le, H., Vishwanathan, N., Kantardjieff, A., Doo, I., Srienc, M., Zheng, X., Somia, N., and Hu, W. S. (2013) Dynamic gene expression for metabolic engineering of mammalian cells in culture, *Metab Eng 20*, 212-220.
[215] Tornoe, J., Kusk, P., Johansen, T. E., and Jensen, P. R. (2002) Generation of a synthetic mammalian promoter library by modification of sequences spacing transcription factor binding sites, *Gene 297*, 21-32.
[216] Ferreira, J. P., Peacock, R. W., Lawhorn, I. E., and Wang, C. L. (2011) Modulating ectopic gene expression levels by using retroviral vectors equipped with synthetic promoters, *Syst Synth Biol 5*, 131-138.
[217] Pasotti, L., Politi, N., Zucca, S., De Angelis, M. G. C., and Magni, P. (2012) Bottom-Up Engineering of Biological Systems through Standard Bricks: A Modularity Study on Basic Parts and Devices, *Plos One 7*, e39407-e39407.
[218] Fritz, B. R., Timmerman, L. E., Daringer, N. M., Leonard, J. N., and Jewett, M. C. (2010) Biology by design: from top to bottom and back, *J Biomed Biotechnol 2010*, 232016.
[219] Fussenegger, M., Moser, S., Mazur, X., and Bailey, J. E. (1997) Autoregulated multicistronic expression vectors provide one-step cloning of regulated product gene expression in mammalian cells, *Biotechnol Prog 13*, 733-740.
[220] Kramer, B. P., Viretta, A. U., Daoud-El-Baba, M., Aubel, D., Weber, W., and Fussenegger, M. (2004) An engineered epigenetic transgene switch in mammalian cells, *Nat Biotechnol 22*, 867-870.
[221] Tigges, M., Marquez-Lago, T. T., Stelling, J., and Fussenegger, M. (2009) A tunable synthetic mammalian oscillator, *Nature 457*, 309-312.
[222] Boshart, M., Weber, F., Jahn, G., Dorsch-Häsler, K., Fleckenstein, B., and Schaffner, W. (1985) A very strong enhancer is located upstream of an immediate early gene of human cytomegalovirus, *Cell 41*, 521-530.
[223] Gorman, C. M., Merlino, G. T., Willingham, M. C., Pastan, I., and Howard, B. H. (1982) The Rous sarcoma virus long terminal repeat is a strong promoter when

introduced into a variety of eukaryotic cells by DNA-mediated transfection, *Proc Natl Acad Sci 79*, 6777-6781.

[224] Qin, J. Y., Zhang, L., Clift, K. L., Hulur, I., Xiang, A. P., Ren, B. Z., and Lahn, B. T. (2010) Systematic comparison of constitutive promoters and the doxycycline-inducible promoter, *PLoS One 5*, e10611.

[225] Kim, M., O'Callaghan, P. M., Droms, K. A., and James, D. C. (2011) A mechanistic understanding of production instability in CHO cell lines expressing recombinant monoclonal antibodies, *Biotechnology and bioengineering 108*, 2434-2446.

[226] Williams, S., Mustoe, T., Mulcahy, T., Griffiths, M., Simpson, D., Antoniou, M., Irvine, A., Mountain, A., and Crombie, R. (2005) 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*, 17-17.

[227] Mariati, Yeo, J. H., Koh, E. Y., Ho, S. C., and Yang, Y. (2014) Insertion of core CpG island element into human CMV promoter for enhancing recombinant protein expression stability in CHO cells, *Biotechnol Prog 30*, 523-534.

[228] Moritz, B., Becker, P. B., and Gopfert, U. (2015) CMV promoter mutants with a reduced propensity to productivity loss in CHO cells, *Sci Rep 5*, 16952.

[229] Brooks, A. R., Harkins, R. N., Wang, P., Qian, H. S., Liu, P., and Rubanyi, G. M. (2004) Transcriptional silencing is associated with extensive methylation of the CMV promoter following adenoviral gene delivery to muscle, *The Journal of Gene Medicine 6*, 395-404.

[230] Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B. K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. A. (2012) The accessible chromatin landscape of the human genome, *Nature 489*, 75-82.

[231] Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M. Q., and Snyder, M. P. (2014) Genome-wide map of regulatory interactions in the human genome, *Genome Res 24*, 1905-1917.

[232] Bailey, S. D., Zhang, X., Desai, K., Aid, M., Corradin, O., Iari, R. C. S., Akhtar-Zaidi, B., Scacheri, P. C., Haibe-Kains, B., and Lupien, M. (2015) ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters, *Nat Commun 6*.

[233] Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C. Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou,

M., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles, *Nucleic Acids Res 42*, D142-147.

[234] Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res 34*, D108-110.

[235] Daily, K., Patel, V. R., Rigor, P., Xie, X., and Baldi, P. (2011) MotifMap: integrative genome-wide maps of regulatory motif sites for model species, *BMC Bioinformatics 12*, 495.

[236] Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., and Bulyk, M. L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions, *Nucleic Acids Res 43*, D117-122.

[237] Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., and Makeev, V. J. (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models, *Nucleic Acids Res 41*, D195-202.

[238] Bandaranayake, A. D., and Almo, S. C. (2014) Recent advances in mammalian protein production, *FEBS Lett 588*, 253-260.

[239] Ghosh, D., and Chinnaiyan, A. M. (2002) Mixture modelling of gene expression data from microarray experiments, *Bioinformatics 18*, 275-286.

[240] Cheng, J. K., and Alper, H. S. (2016) Transcriptomics-Guided Design of Synthetic Promoters for a Mammalian System, *ACS Synthetic Biology*.

[241] Schirm, S., Jiricny, J., and Schaffner, W. (1987) The Sv40 Enhancer Can Be Dissected into Multiple Segments, Each with a Different Cell Type Specificity, *Gene Dev 1*, 65-74.

[242] Running Deer, J., and Allison, D. S. (2004) High-level expression of proteins in mammalian cells using transcription regulatory sequences from the Chinese hamster EF-1alpha gene, *Biotechnol Prog 20*, 880-889.

[243] Ho, S. C., and Yang, Y. (2014) Identifying and engineering promoters for high level and sustainable therapeutic recombinant protein production in cultured mammalian cells, *Biotechnol Lett 36*, 1569-1579.

[244] Black, A. R., Black, J. D., and Azizkhan-Clifford, J. (2001) Sp1 and kruppel-like factor family of transcription factors in cell growth regulation and cancer, *J Cell Physiol 188*, 143-160.

[245] Van Der Velden, A., Kaminski, A., Jackson, R. J., and Belsham, G. J. (1995) Defective point mutants of the encephalomyocarditis virus internal ribosome entry site can be complemented in trans, *Virology 214*, 82-90.

[246] Hess, J., Angel, P., and Schorpp-Kistner, M. (2004) AP-1 subunits: quarrel and harmony among siblings, *J Cell Sci 117*, 5965-5973.

[247] Gunther, V., Strassen, T., Lindert, U., Dagani, P., Waldvogel, D., Georgiev, O., Schaffner, W., and Bethge, T. (2012) Simian virus 40 strains with novel properties generated by replacing the viral enhancer with synthetic oligonucleotides, *J Virol 86*, 3135-3142.
[248] Curran, K. A., Crook, N. C., Karim, A. S., Gupta, A., Wagman, A. M., and Alper, H. S. (2014) Design of synthetic yeast promoters via tuning of nucleosome architecture, *Nat Commun 5*, 4002.
[249] Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters, *Nat Biotechnol 30*, 521-530.
[250] Hannenhalli, S., and Levy, S. (2002) Predicting transcription factor synergism, *Nucleic Acids Res 30*, 4278-4284.
[251] Cheng, J. K., and Alper, H. S. (2014) The genome editing toolbox: a spectrum of approaches for targeted modification, *Current Opinion in Biotechnology 30*, 87-94.
[252] Lanza, A. M., Cheng, J. K., and Alper, H. S. (2012) Emerging synthetic biology tools for engineering mammalian cell systems and expediting cell line development, *Current Opinion in Chemical Engineering 1*, 403-410.
[253] Yin, H., Kanasty, R. L., Eltoukhy, A. A., Vegas, A. J., Dorkin, J. R., and Anderson, D. G. (2014) Non-viral vectors for gene-based therapy, *Nat Rev Genet 15*, 541-555.
[254] Redden, H., and Alper, H. S. (2015) The development and characterization of synthetic minimal yeast promoters, *Nat Commun 6*.
[255] Kosuri, S., and Church, G. M. (2014) Large-scale de novo DNA synthesis: technologies and applications, *Nat Methods 11*, 499-507.
[256] Kadonaga, J. T. (2004) Regulation of RNA Polymerase II Transcription by Sequence-Specific DNA Binding Factors, *Cell 116*, 247-257.
[257] Thomsen, D. R., Stenberg, R. M., Goins, W. F., and Stinski, M. F. (1984) Promoter-regulatory region of the major immediate early gene of human cytomegalovirus, *Proc Natl Acad Sci 81*, 659-663.
[258] Everett, R. D., Baty, D., and Chambon, P. (1983) The repeated GC-rich motifs upstream from the TATA box are important elements of the SV40 early promoter, *Nucleic Acids Res 1*, 28-30.
[259] Barrera-Saldana, H., Takahashi, K., Vigneron, M., Wildeman, a., Davidson, I., and Chambon, P. (1985) All six GC-motifs of the SV40 early upstream element contribute to promoter activity in vivo and in vitro, *EMBO J 4*, 3839-3849.
[260] Zenke, M., Grundstrom, T., Matthes, H., Wintzerith, M., Schatz, C., Wildeman, A., and Chambon, P. (1986) Multiple sequence motifs are involved in SV40 enhancer function, *EMBO J 5*, 387-397.
[261] Wakabayashi-Ito, N., and Nagata, S. (1994) Characterization of the regulatory elements in the promoter of the human elongation factor-1 alpha gene, *The Journal of biological chemistry 269*, 29831-29837.

[262] Yew, N. S., Przybylska, M., Ziegler, R. J., Liu, D., and Cheng, S. H. (2001) High and sustained transgene expression in vivo from plasmid vectors containing a hybrid ubiquitin promoter, *Mol Ther 4*, 75-82.
[263] Ferreira, J. P., Peacock, R. W. S., Lawhorn, I. E. B., and Wang, C. L. (2011) Modulating ectopic gene expression levels by using retroviral vectors equipped with synthetic promoters, *Syst Synth Biol 5*, 131-138.
[264] Li, M., Wang, J., Geng, Y., Li, Y., Wang, Q., Liang, Q., and Qi, Q. (2012) A strategy of gene overexpression based on tandem repetitive promoters in Escherichia coli, *Microb Cell Fact 11*, 19-19.
[265] Smale, S. T. (2001) Core promoters: active contributors to combinatorial gene regulation, *Genes Dev 15*, 2503-2508.
[266] Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W., and Kadonaga, J. T. (2008) The RNA polymerase II core promoter - the gateway to transcription, *Curr Opin Cell Biol 20*, 253-259.
[267] Juven-Gershon, T., Cheng, S., and Kadonaga, J. T. (2006) Rational design of a super core promoter that enhances gene expression, *Nat Meth 3*, 917-922.
[268] Nott, A., Shlomo, M. H., and Moore, M. J. (2003) A quantitative analysis of intron effects on mammalian gene expression, *RNA 9*, 607-617.
[269] Bianchi, M., Crinelli, R., Giacomini, E., Carloni, E., and Magnani, M. (2009) A potent enhancer element in the 5'-UTR intron is crucial for transcriptional regulation of the human ubiquitin C gene, *Gene 448*, 88-101.
[270] Ng, S.-Y., Gunning, P., Liu, S.-H., Leavitt, J., and Kedes, L. (1989) Regulation of the human β-actin promoter by upstream and intron domains, *Nucleic Acids Res 17*, 601-615.
[271] Xia, W., Bringmann, P., McClary, J., Jones, P. P., Manzana, W., Zhu, Y., Wang, S., Liu, Y., Harvey, S., Madlansacay, M. R., McLean, K., Rosser, M. P., MacRobbie, J., Olsen, C. L., and Cobb, R. R. (2006) High levels of protein expression using different mammalian CMV promoters in several cell lines, *Protein Express Purif 45*, 115-124.
[272] Bornstein, P., McKay, J., Morishima, J. K., Devarayalu, S., and Gelinas, R. E. (1987) Regulatory elements in the first intron contribute to transcriptional control of the human alpha 1(I) collagen gene, *Proc Natl Acad Sci 84*, 8869-8873.
[273] Quilici, L. S., Silva-Pereira, I., Andrade, a. C., Albuquerque, F. C., Brigido, M. M., and Maranhão, a. Q. (2013) A minimal cytomegalovirus intron A variant can improve transgene expression in different mammalian cell lines, *Biotechnol Lett 35*, 21-27.
[274] Tokusumi, Y., Ma, Y., Song, X., Jacobson, R. H., and Takada, S. (2007) The New Core Promoter Element XCPE1 (X Core Promoter Element 1) Directs Activator-, Mediator-, and TATA-Binding Protein-Dependent but TFIID-Independent RNA Polymerase II Transcription from TATA-Less Promoters, *Mol Cell Biol 27*, 1844-1858.

[275] Anish, R., Hossain, M. B., Jacobson, R. H., and Takada, S. (2009) Characterization of transcription from TATA-less promoters: identification of a new core promoter element XCPE2 and analysis of factor requirements, *PLOS ONE 4*, e5103-e5103.
[276] Marbach-Bar, N., Bahat, A., Ashkenazi, S., Golan-Mashiach, M., Haimov, O., Wu, S.-Y., Chiang, C.-M., Puzio-Kuter, A., Hirshfield, K. M., Levine, A. J., and Dikstein, R. (2016) DTIE, a novel core promoter element that directs start site selection in TATA-less genes, *Nucleic Acids Res 44*, 1080-1094.
[277] Juven-Gershon, T., and Kadonaga, J. T. (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery, *Dev Biol 339*, 225-229.
[278] Deng, W., and Roberts, S. G. E. (2005) A core promoter element downstream of the TATA box that is recognized by TFIIB, *Genes Dev 19*, 2418-2423.
[279] Lee, N., Iyer, S. S., Mu, J., Weissman, J. D., Ohali, A., Howcroft, T. K., Lewis, B. a., and Singer, D. S. (2010) Three novel downstream promoter elements regulate MHC class I promoter activity in mammalian cells, *PLOS ONE 5*, e15278-e15278.
[280] Sawicki, J. A., Morris, R. J., Monks, B., Sakai, K., and Miyazaki, J.-i. (1998) A Composite CMV-IE Enhancer/β-Actin Promoter Is Ubiquitously Expressed in Mouse Cutaneous Epithelium, *Exp Cell Res 244*, 367-369.
[281] Oldridge, D. A., Wood, A. C., Weichert-Leahey, N., Crimmins, I., Sussman, R., Winter, C., McDaniel, L. D., Diamond, M., Hart, L. S., Zhu, S., Durbin, A. D., Abraham, B. J., Anders, L., Tian, L., Zhang, S., Wei, J. S., Khan, J., Bramlett, K., Rahman, N., Capasso, M., Iolascon, A., Gerhard, D. S., Guidry Auvil, J. M., Young, R. A., Hakonarson, H., Diskin, S. J., Thomas Look, A., and Maris, J. M. (2015) Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism, *Nature 528*, 418-421.
[282] Lovén, J., Hoke, H. a., Lin, C. Y., Lau, A., Orlando, D. a., Vakoc, C. R., Bradner, J. E., Lee, T. I., and Young, R. a. (2013) Selective inhibition of tumor oncogenes by disruption of super-enhancers, *Cell 153*, 320-334.
[283] Le Hir, H., Nott, A., and Moore, M. J. (2003) How introns influence and enhance eukaryotic gene expression, *Trends Biochem Sci 28*, 215-220.
[284] Chapman, B. S., Thayer, R. M., Vincent, K. a., and Haigwood, N. L. (1991) Effect of intron A from human cytomegalovirus (Towne) immediate-early gene on heterologous expression in mammalian cells, *Nucleic Acids Res 19*, 3979-3986.
[285] Kwek, K. Y., Murphy, S., Furger, A., Thomas, B., O'Gorman, W., Kimura, H., Proudfoot, N. J., and Akoulitchev, A. (2002) U1 snRNA associates with TFIIH and regulates transcriptional initiation, *Nat Struct Mol Biol 9*, 800-805.
[286] Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles, *Nucleic Acids Res 44*, D110-115.
[287] Sasaki-Haraguchi, N., Shimada, M. K., Taniguchi, I., Ohno, M., and Mayeda, A. (2012) Mechanistic insights into human pre-mRNA splicing of human ultra-short

introns: Potential unusual mechanism identifies G-rich introns, *Biochem Biophys Res Commun 423*, 289-294.
[288] Piovesan, A., Caracausi, M., Ricci, M., Strippoli, P., Vitale, L., and Pelleri, M. C. (2015) Identification of minimal eukaryotic introns through GeneBase, a user-friendly tool for parsing the NCBI Gene databank, *DNA Research 22*, 495-503.
[289] Mariati, Ng, Y. K., Chao, S.-H., Yap, M. G. S., and Yang, Y. (2010) Evaluating regulatory elements of human cytomegalovirus major immediate early gene for enhancing transgene expression levels in CHO K1 and HEK293 cells, *Journal of biotechnology 147*, 160-163.
[290] Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., Herman, B., Happe, S., Higgs, A., LeProust, E., Follows, G. A., Fraser, P., Luscombe, N. M., and Osborne, C. S. (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C, *Nat Genet 47*, 598-606.
[291] Proudfoot, N. J., Furger, A., and Dye, M. J. (2002) Integrating mRNA Processing with Transcription, *Cell 108*, 501-512.
[292] West, S., and Proudfoot, N. J. (2009) Transcriptional Termination Enhances Protein Expression in Human Cells, *Mol Cell 33*, 354-364.
[293] Antoniou, M., Geraghty, F., Hurst, J., and Grosveld, F. (1998) Efficient 3′-end formation of human β-globin mRNA in vivo requires sequences within the last intron but occurs independently of the splicing reaction, *Nucleic Acids Res 26*, 721-729.
[294] Kaufman, R. (2000) Overview of vector design for mammalian gene expression, *Mol Biotechnol 16*, 151-160.
[295] Baek, K. H., Sato, K., Ito, R., and Agarwal, K. (1986) RNA polymerase II transcription terminates at a specific DNA sequence in a HeLa cell-free reaction, *Proc Natl Acad Sci 83*, 7623-7627.
[296] Sato, K., Ito, R., Baek, K. H., and Agarwal, K. (1986) A specific DNA sequence controls termination of transcription in the gastrin gene, *Mol Cell Biol 6*, 1032-1043.
[297] Connelly, S., and Manley, J. L. (1988) A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II, *Genes Dev 2*, 440-452.
[298] Carswell, S., and Alwine, J. C. (1989) Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences, *Mol Cell Biol 9*, 4248-4258.
[299] Schek, N., Cooke, C., and Alwine, J. C. (1992) Definition of the upstream efficiency element of the simian virus 40 late polyadenylation signal by using in vitro analyses, *Mol Cell Biol 12*, 5386-5393.
[300] Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J.-M., and Gautheret, D. (2000) Patterns of Variant Polyadenylation Signal Usage in Human Genes, *Genome Res 10*, 1001-1010.

[301] Zarudnaya, M. I., Kolomiets, I. M., Potyahaylo, A. L., and Hovorun, D. M. (2003) Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures, *Nucleic Acids Res 31*, 1375-1386.
[302] Proudfoot, N. J. (2011) Ending the message: poly(A) signals then and now, *Genes Dev 25*, 1770-1782.
[303] Danckwardt, S., Kaufmann, I., Gentzel, M., Foerstner, K. U., Gantzert, A. S., Gehring, N. H., Neu-Yilik, G., Bork, P., Keller, W., Wilm, M., Hentze, M. W., and Kulozik, A. E. (2007) Splicing factors stimulate polyadenylation via USEs at non-canonical 3′ end formation signals, *EMBO J 26*, 2658-2669.
[304] McLauchlan, J., Gaffney, D., Whitton, J. L., and Clements, J. B. (1985) The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3′ termini, *Nucleic Acids Res 13*, 1347-1368.
[305] Curran, K. A., Morse, N. J., Markham, K. A., Wagman, A. M., Gupta, A., and Alper, H. S. (2015) Short Synthetic Terminators for Improved Heterologous Gene Expression in Yeast, *ACS Synthetic Biology 4*, 824-832.
[306] Kim, D., Kim, J. D., Baek, K., Yoon, Y., and Yoon, J. (2003) Improved mammalian expression systems by manipulating transcriptional termination regions, *Biotechnol Progr 19*, 1620-1622.
[307] Curran, K. A., Karim, A. S., Gupta, A., and Alper, H. S. (2013) Use of expression-enhancing terminators in Saccharomyces cerevisiae to increase mRNA half-life and improve gene expression control for metabolic engineering applications, *Metab Eng 19*, 88-97.
[308] West, S., Proudfoot, N. J., and Dye, M. J. (2008) Molecular Dissection of Mammalian RNA Polymerase II Transcriptional Termination, *Mol Cell 29*, 600-610.
[309] Proudfoot, N. (1991) Poly(A) signals, *Cell 64*, 671-674.
[310] Gil, A., and Proudfoot, N. J. (1984) A sequence downstream of AAUAAA is required for rabbit β-globin mRNA 3′-end formation, *Nature 312*, 473-474.
[311] Salisbury, J., Hutchison, K. W., and Graber, J. H. (2006) A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif, *BMC Genomics 7*, 55-55.
[312] Nunes, N. M., Li, W., Tian, B., and Furger, A. (2010) A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence, *EMBO J 29*, 1523-1536.
[313] Brown, K. M., and Gilmartin, G. M. (2003) A Mechanism for the Regulation of Pre-mRNA 3′ Processing by Human Cleavage Factor Im, *Mol Cell 12*, 1467-1476.
[314] Venkataraman, K., Brown, K. M., and Gilmartin, G. M. (2005) Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition, *Genes Dev 19*, 1315-1327.
[315] Levitt, N., Briggs, D., Gil, a., and Proudfoot, N. J. (1989) Definition of an efficient synthetic poly(A) site, *Genes Dev 3*, 1019-1025.

[316] Sharova, L. V., Sharov, A. a., Nedorezov, T., Piao, Y., Shaik, N., and Ko, M. S. H. (2009) Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells, *DNA research 16*, 45-58.
[317] Pérez Cañadillas, J. M., and Varani, G. (2003) Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein, *EMBO J 22*, 2821-2830.
[318] Legendre, M., and Gautheret, D. (2003) Sequence determinants in human polyadenylation site selection, *BMC Genomics 4*, 7-7.
[319] Antoniou, M., Harland, L., Mustoe, T., Williams, S., Holdstock, J., Yague, E., Mulcahy, T., Griffiths, M., Edwards, S., Ioannou, P. a., Mountain, A., and Crombie, R. (2003) Transgenes encompassing dual-promoter CpG islands from the human TBP and HNRPA2B1 loci are resistant to heterochromatin-mediated silencing, *Genomics 82*, 269-279.
[320] Guo, Z., and Sherman, F. (1996) Signals sufficient for 3'-end formation of yeast mRNA, *Mol Cell Biol 16*, 2772-2776.
[321] Singer, M. F. (1982) SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes, *Cell 28*, 433-434.
[322] Smeenk, G., and van Attikum, H. (2013) The Chromatin Response to DNA Breaks: Leaving a Mark on Genome Integrity, *Annu Rev Biochem 82*, 55-80.
[323] Ferguson, D. O., and Alt, F. W. (2001) DNA double strand break repair and chromosomal translocation: lessons from animal models, *Oncogene 20*, 5572-5579.
[324] Liang, X., Potter, J., Kumar, S., Zou, Y., Quintanilla, R., Sridharan, M., Carte, J., Chen, W., Roark, N., Ranganathan, S., Ravinder, N., and Chesnut, J. D. (2015) Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection, *Journal of biotechnology 208*, 44-53.
[325] Aymard, F., Bugler, B., Schmidt, C. K., Guillou, E., Caron, P., Briois, S., Iacovoni, J. S., Daburon, V., Miller, K. M., Jackson, S. P., and Legube, G. (2014) Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks, *Nat Struct Mol Biol 21*, 366-374.
[326] Richard, P., and Manley, J. L. (2009) Transcription termination by nuclear RNA polymerases, *Genes Dev 23*, 1247-1269.
[327] Liu, Y. G., and Huang, N. (1998) Efficient amplification of insert end sequences from bacterial artificial chromosome clones by thermal asymmetric interlaced PCR, *Plant Mol Biol Rep 16*, 175-181.
[328] Liu, Y. G., and Whittier, R. F. (1995) Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking, *Genomics 25*, 674-681.
[329] Rahmutula, D., Nakayama, T., Soma, M., Kosuge, K., Aoi, N., Izumi, Y., Kanmatsuse, K., and Ozawa, Y. (2002) Structure and polymorphisms of the human natriuretic peptide receptor C gene, *Endocrine 17*, 85-90.

[330] Bryda, E. C., Pearson, M., Agca, Y., and Bauer, B. A. (2006) Method for detection and identification of multiple chromosomal integration sites in transgenic animals created with lentivirus, *Biotechniques 41*, 715-719.
[331] Ochman, H., Gerber, A. S., and Hartl, D. L. (1988) Genetic applications of an inverse polymerase chain reaction, *Genetics 120*, 621-623.
[332] Do, C. B., and Batzoglou, S. (2008) What is the expectation maximization algorithm?, *Nat Biotech 26*, 897-899.