# Bioinformatics design of *cis*-regulatory elements controlling human gene expression

by

Rachelle Farkas

B.Comp., Queen's University at Kingston, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

**Master of Science**

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Bioinformatics)

The University of British Columbia
(Vancouver)

October 2017

# Abstract

Gene therapy has the potential to not only treat, but cure individuals suffering from inherited diseases. Advances in understanding the human genome and the discovery of causal genes underlying diseases has heightened the need to solve the gene therapy challenge. Viral vectors are often used as a delivery tool for therapeutics, but their safety and efficacy are still being studied. To contribute to this goal, we have created 49 small viral promoters by bioinformatically annotating *cis*-regulatory regions from which a subset are concatenated with the goal of driving cell-specific expression of a reporter gene. We have tested a subset of these in mice *in vivo*. Regulatory region analysis can take a trained designer multiple weeks. To resolve this issue, we have created a semi-automated approach to regulatory region identification, named OnTarget. The OnTarget database accumulates thousands of cell and tissue-specific experiments in order to identify regions informative of regulatory properties. OnTarget is able to identify regulatory regions consistent with those identified by designers. In this capacity, we expect OnTarget to lead to better and faster identification of *cis*-regulatory regions for the design of promoters targeting specific sets of cells.

# Lay Summary

Millions of people currently live with incurable genetic diseases. Although many treatments exist to ease symptoms of these diseases, they are often expensive and invasive, resulting in both financial and emotional burdens on patients, their families, and healthcare systems. Gene therapy has the potential to not only treat, but potentially cure genetic diseases. The concept is simple: replace a malfunctioning gene with a working version. Current gene therapies often do not discriminate in the delivery of these genes, which can lead to healthy cells receiving these unnecessary genes potentially causing unwanted side effects. In order to address this issue, we have designed a method to limit the replacement gene to be active in the right types of cells. We have created software to make this process available to other researchers.

# Preface

This thesis contains original work as well as extensions to the MiniPromoter project led by the laboratory of Dr. Elizabeth M. Simpson (UBC). All work was performed at the UBC Centre for Molecular Medicine and Therapeutics at the BC Childrens Hospital Research Institute under the supervision of Dr. Wyeth Wasserman. No text is taken from previously published material.

The MiniPromoter design protocol was defined by myself and Dr. Oriol Fornes, building from reported approaches of past members of the lab. I established the On-Target analysis steps, creating pseudocode and flowcharts, which David Arenillas programmed. All data was downloaded for free academic use from the FANTOM5 consortium, ENCODE project, Roadmap Project, the UCSC Genome Browser, and GEO archives. With the exception of the KRT12 Ple326 and Ple334 constructs (which was analyzed by myself within the Simpson laboratory), the mouse work was performed by our collaborators in the Simpson laboratory, including Jack Hickmott, Andrea Korecki, and Siu Ling Lam, and was covered under the UBC Animal Ethics Certificate A14-0295 and BioSafety Certificate B14-0131.

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

**AAV** Adeno-associated virus

**ATAC-SEQ** Assay for Transposable Accessible Chromatin Sequencing

**BAC** Bacterial artificial chromosome

**CAGE** Cap Analysis of Gene Expression

**CDS** Coding start site

**CHIP-SEQ** Chromatin immunoprecipitation Sequencing

**CNS** Central nervous system

**CTCF** CCTC-binding factor

**DHS** DNase I hypersensitive sites

**DNASE-SEQ** DNase I hypersensitive sites Sequencing

**EMGFP** Emerald green fluorescence protein

**ENCODE** The Encyclopedia of DNA Elements

**ERNA** Enhancer RNA

**FAIRE-SEQ** Formaldehyde-Assisted Isolation of Regulatory Elements Sequencing

**FANTOM** Functional Annotation of the Mammalian Genome

**FANTOM5** FANTOM consortium fifth project

**GENSAT** Gene Expression Nervous System Atlas

**GRO-SEQ** Global Run-On Sequencing

**HI-C** High-resolution chromosome conformation capture

**LINE** Long interspersed nuclear element

**LTR** Long terminal repeat

**MRNA** Messenger RNA

**RNAPII** RNA polymeraseII

**RNA-SEQ** RNA Sequencing

**RR** *cis*-regulatory region

**SINE** Short interspersed nuclear element

**SMCBA** Small chicken beta actin

**TAD** Topologically Associating Domain

**TF** Transcription factor

**TSS** Transcription start site

**UCSC** University of California, Santa Cruz

# Acknowledgments

I would like to thank my supervisor Dr. Wyeth Wasserman for not only giving me the chance to work on an exciting project, but for all the support, encouragement, and guidance throughout my studies. An extended thank you to everyone in the Wasserman lab, including: Dr. Oriol Fornes for the endless hours of support and collaboration, Dora Pak for managing all schedules and overall support, David Arenillas for all things programming and computational discussion, and Phillip Richmond, Allen Zhang, Cynthia Ye, and Dr. Robin van der Lee for many helpful discussions. Additionally, thanks to members of the Simpson lab (Dr. Elizabeth Simpson, Andrea Korecki, Jack Hickmott, Siu Ling Lam, Zeinab Mohanna) for taking care of me throughout our collaborations and my directed studies work. I would also like to thank Shams Bhuiyan and Louie Dinh for helpful discussion and providing me with food throughout this time. A special thanks to the members of my committee, Dr. Paul Pavlidis, Dr. Pamela Hoodless, and Dr. Cristina Conati, for their helpful suggestions and comments throughout my studies and this thesis.

# Dedication

To my parents, sisters, partner, and cat for always giving me so much love and support.

# Chapter 1

# Introduction

As healthcare costs continue to rise, it is imperative to further not only the understanding of human diseases but to provide new and effective treatments[63]. Unlike diseases contracted by foreign agents, a large portion of inherited diseases are currently incurable. Treatments exists to alleviate symptoms, but often times provide no cure. Certain individuals suffering from inherited diseases must continue treatment for life, which produces both financial and emotional burdens on patients, their families, and the healthcare system as a whole.

At the conceptual level, the problem of genetic disorders seems simple; a malfunctioning gene can be replaced by a functioning version. At the implementation level, however, there are challenges in the identification of the gene(s) involved, in the delivery of the restorative gene to the appropriate cells in the body, in the maintenance of expression of the replacement gene, and in the prevention of unintended effects[54]. Efforts spanning more than 25 years[25] to produce gene therapies have confronted these issues, with mixed success[54, 81].

Substantial advances in the understanding of the human genome and the discovery of causal genes underlying diseases has heightened the need to solve the gene therapy challenge. Improvements in the delivery of nucleic acids[54, 75] have allowed for a new era of new gene therapies, with hundreds of new clinical trials underway worldwide[25]. To realize the full potential of gene therapy, additional advances will be required, including improving delivery of therapeutic DNA to relevant cells and tissues[37]. In particular, the most popular method of *in vivo*

delivery is through viral vectors, engineered to remove pathogenic properties[75].

One of the identified challenges in the field is the establishment of 'promoter' sequences capable of directing therapeutic gene expression in a targeted manner (the formal meaning of 'promoter' will be fully discussed below). Most existing vectors incorporate ubiquitous promoters, but calls have been made to find promoters capable of directing gene expression in the correct subset of cells which also affects gene therapy safety and efficacy[80]. Promoter design and selection is a challenge, as the DNA sequence must be capable of utilizing a host cell's transcriptional machinery[75]. It is possible to form promoter sequences from piecing together endogenous DNA known to promote gene expression in a desired pattern[22, 23, 40, 60]. Many of these sequences are part of the non-coding regions of the genome, which accounts for 98% of all genetic material in the human genome[27]. Recognizing these sequences is therefore an important goal, and several methodologies and technologies have been developed to aid in their identification.

Designing promoters for use in viral vectors is a key step to a future in which gene therapies are widely used to treat and, in the best cases, cure genetic disorders.

## 1.1 Gene Therapy

While the knowledge of gene transfer dates back to the 1947 discovery of bacterial conjugation, the launch of the modern field of gene therapy is marked by the first clinical trial in 1989[81]. By the mid-1990s, trials for the treatment of diverse disorders, such as adenosine deaminase deficiency[11] and cystic fibrosis[14], caused a boom within the field. While results were often mixed, the growth of the field continued rapidly until 1999, when the death of Jesse Gelsinger occurred during a trial to correct the effects of ornithine transcarbamylase deficiency. The cause of his death was multi-organ failure due to a large immune response over an administered high dose of the adenosine virus vector.[42].

Renewed hope arose in 2000 when researchers cured patients with X-linked severe combined immunodeficiency-X1 through the use of a retrovirus[18]. However, a couple years after publication, two of the patients treated for the disease developed a leukemia-like disease, due to the retroviral gene inserted near the LMO2

oncogene[38] (the insertions were presumed to be activating).

The gravity of these failures weighed on the community and prompted reviews of the gene therapy field as a whole[69]. Emphasis was placed on the identification/creation of new viral vectors capable of safe delivery of therapeutic genes[75]. Advances in vector technologies have led to success in animal models, and since the early 2010s, the number of clinical trials for gene therapies has increased dramatically once again[54].

### 1.1.1 Gene Therapy via Viral Vectors

Viruses are currently the dominant tool for delivery of therapeutic DNA[25]. As viruses have the ability to transmit their genetic material into cells, they are highly relevant to gene therapy. The failures of gene therapy at the turn of the century highlighted a need for deeper understanding of viruses, and how they could be selected/modified to circumvent the known problems[43]. The subsequent research revealed advantages and drawbacks for specific viral vectors. Proper vector selection for the scope of each trial therapy increased both safety and efficacy[54]. Individual vectors differ in the types of cells they are able to transduce, the length of DNA (or in some cases RNA) they can deliver, how long therapeutic expression will persist, and, in some aspects, expected host immune response post-injection[1] (see Figure 1).

The Journal of Gene Medicine maintains a database of gene therapies of the past (since 1989), and current clinical trials[25]. Approximately 70% (69.5%) of all trials used viral vectors (1989-2016). In 2016, viral vectors made up 84% of newly approved gene therapy trials. While diverse viruses have been used for past gene therapy trials, four prominent vectors are used in therapies today: adenovirus, retrovirus, lentivirus, and adeno-associated (AAV) virus.

The adenovirus is historically the most used viral vector for gene therapies, accounting for 30% of all viral vector clinical trials[25]. As a medium-sized virus, the adenovirus genome contains $\sim$36 kb of double stranded DNA, although only about 8 kb can be used to package desired DNA with the remaining space occupied by genes that are important for transcription and virus integrity[43]. Amongst the popular viral vectors, adenovirus carries the largest payload of DNA[75]. It

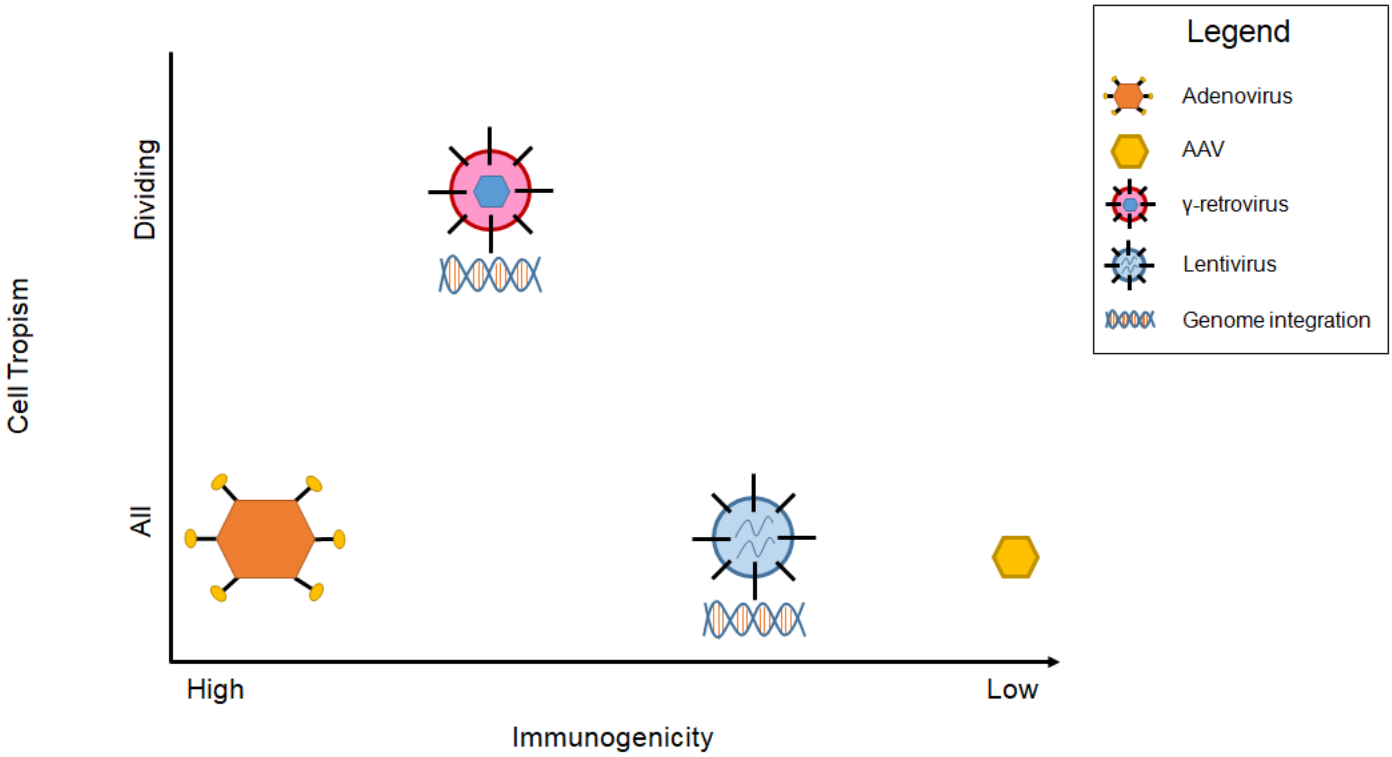



**Figure 1: Overview of Popular Viral Vectors.** Vector choice depends on a variety of factors including immunogenic host response, entry into cells, genome integration, and packaging size. Larger icons indicate larger packaging capability. The adenovirus can package just over 8 kb of non-endogenous DNA and transduce all cells, however its expression is transient and it produces a large host immune response. The $\gamma$-retrovirus can package around 8 kb of non-endogenous RNA and will achieve stable expression due to genome integration, however it will produce a host immune response and can only transduce actively dividing cells. The lentivirus can package around 8 kb of non-endogenous RNA, transduces all cells, and usually does not promote a host immune response. Its expression is stable, but it inserts its viral genome into the host genome at random loci. The AAV generally does not promote a host immune response and readily transduces most cells. It is the smallest of the popular viral vectors, packaging under 5 kb of non-endogenous DNA, and expression is transient in quickly dividing cells.

can transduce both actively dividing and non-dividing (quiescent) cells[1]. The adenovirus does not integrate into its host genome, existing in the cell as a non-replicating episome, and therefore the expression of its therapeutic is transient in dividing cells (as the episome is diluted)[8]. The biggest disadvantage to the adenovirus is its highly immunogenic nature[75]. Adenoviral vectors have largely been replaced by other, less immunogenic vectors, however it is currently popular in cancer clinical trials[19, 26].

Retroviral vectors have accounted for 27% of all viral vector clinical trials[25]. The widely used $\gamma$-retrovirus vectors can package $\sim$8 kb of RNA. Retroviruses integrate into host genomes, therefore enabling stable expression of a transgene[10]. The insertion location is random, however, which may lead to oncogene activation[37]. Additionally, $\gamma$-retroviruses can only transduce actively dividing cells, which limits their utility for targeting cells or tissues that do not replicate often[43].

The lentivirus, whose vectors are often based on the HIV-1 virus[75], has become an increasingly popular in current trials. While only about 9% of all historic viral gene therapies used a lentiviral vector, in 2016 it comprised of 24% of all recorded trials[25]. A variety of sources differ on the payload capacity of the vector[70, 72], but a consensus is that robust packaging tends to occur when RNA is be less than 8 kb[2]. Although a subclass of retrovirus, lentiviral vectors can transduce both dividing and non-dividing cells[1]. Furthermore, these vectors do not produce a large immune response[75]. As lentivirus contents are inserted into the host genome, their stable expression is at the expense of the risk for oncogene activation[70, 75].

Lastly, the AAV has also increased in usage as a vector over the years. While accounting for 10% of all historic viral gene therapies, AAV vectors were used in 23% of all 2016 trials[25]. The AAV vector, the focus of the research in this thesis, is further described in the following section.

### 1.1.2 Adeno-associated Virus (AAV) Vectors

AAVs have become an increasingly popular choice of vector in viral gene therapies, as some of its most desirable features include its low human pathogenicity, its ability to transduce both dividing and non-dividing cells, and non-replicative

nature[32]. Additionally, another appeal is that engineered vectors have ensured that the AAV will not integrate into the host genome, due to its removal of the viral *rep* genes[43]. There are two main drawbacks for AAVs. First, the AAV has a small payload capacity. At less than 5 kb per virus, the AAV is the smallest of all highly used viral vectors for gene therapies[70]. Second, AAV episomes are lost over cell divisions[1].

There are nine main serotypes of the AAV that can infect human cells[85], and each enter a subset set of cells with greater specificity than others due to differences in capsid structure[82]. While the AAV2 serotype has been the most widely studied, it transduces cells slower and is less efficient than most other serotypes[85]. More recently, hybrid systems, usually made by combining viral capsid proteins to create mosaic capsids, allow for a greater range of specific cell types to be targeted[5]. Proper AAV serotype selection is important for the design of therapies. For example, AAV9 is efficient at targeting neurons in cells of the central nervous system (CNS)[67], while AAV2 is still the vector of choice for targeting cells in the kidney[82].

In order to deliver a therapeutic of interest, the vector offers little room for the inclusion of other genomic elements. An AAV must include inverted terminal repeats (ITRs) at the 5' and 3' end of their genomes, followed by a promoter, a transgene, and a polyadenylation sequence (e.g. simian virus 40 late)[32]. As AAV serotypes are similar in their packaging capacity, to allow larger transgenes most studies utilize small, ubiquitous promoters, such as the approximately 500 base pair sized CMV[33] and CAG[56] promoters. Thus AAV vectors will express in off-target cells, which may not be appropriate for all therapeutics.

To achieve the highest and most specific therapeutic effects, the designers of new therapies must therefore consider carefully both the capsid (serotype) and promoter properties. By optimizing the capsid properties of viruses, one can bias the uptake of the therapy to certain cell types, and much research is currently addressing this mechanism[5, 85]. However, there have been calls to incorporate more selective regulatory sequences controlling the transcription of the therapeutic gene. This thesis focuses on this opportunity to improve the delivery of gene therapy by designing these promoter sequences.

## 1.2 Regulatory Elements

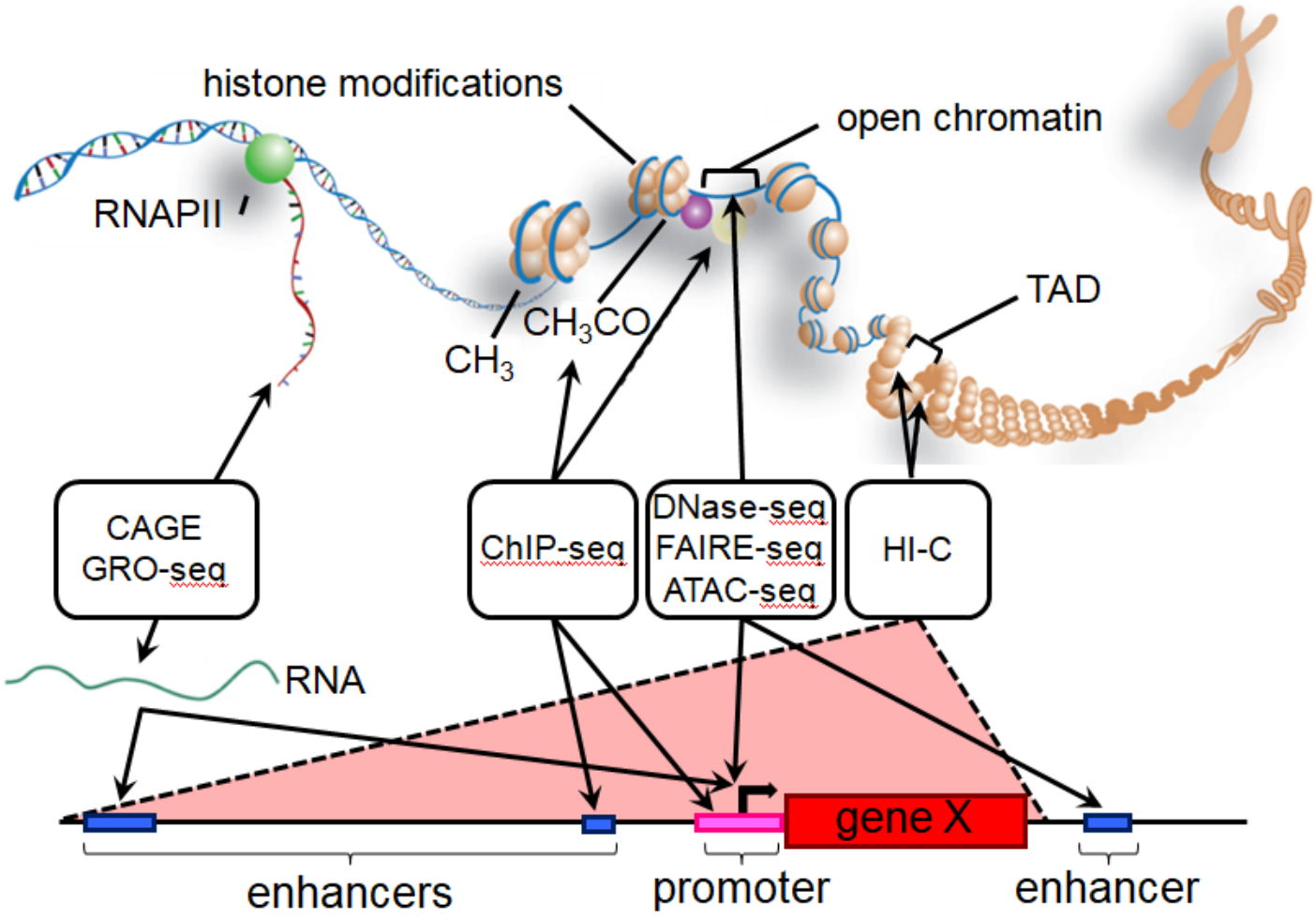
Great progress has been made in understanding the mechanisms which regulate mammalian gene transcription. As a basic model, the RNA Polymerase II complex (RNAPII), which is required to transcribe gene DNA into messenger RNA (mRNA) must assemble on DNA before a gene. This region overlapping transcription start site(s) (TSSs) is called a promoter region. Other elements that affect the rate of transcription enable recruitment of other factors necessary for the formation (or obstruction) of RNAPII. Such regions have been labeled 'enhancers' (or 'silencers'). Both promoters and enhancers contain short elements to which DNA binding proteins, called transcription factors (TFs) can bind in a sequence specific manner. Characteristics of these regulatory features are further described below. For clarity, TFs are a broad category of proteins, of which only a subset exhibit sequence-specific DNA binding, but within this thesis TFs will refer specifically to this subset. An overview of regulatory elements and profiling methods is shown in Figure 2.

### 1.2.1 Promoters

In eukaryotes, promoters are regulatory DNA sequences proximal to the 5' end of genes and are important in the initiation of transcription from DNA to RNA. All gene promoters include one or more TSSs, where the DNA first starts to be transcribed by a RNA Polymerase complex[34]. Promoters contain TF binding sites necessary for the recruitment/assembly of RNA polymerase complexes. Certain genes are regulated by multiple promoters and TSSs, often in cell-type or developmental-type contexts[31].

A subset of promoters (24%) include a TATA-box feature[83], to which a component of the RNAPII can bind. Many mammalian promoters ($\sim$70%) overlap CpG islands[83] (regions in which CpG dinucleotides have been retained over evolution at levels consistent with C and G mononucleotide frequencies, reflecting a lack of methylation of CpGs in promoter regions that promotes CpG elimination over evolution). Some promoters combine both TATA-box and CpG islands, while others have neither[74].

Over the past decade extensive profiling of the locations and activities of pro-




Adapted from The ENCODE Project Consortium website; Fornes O.

**Figure 2: Overview of Transcriptional Regulation Data.** Within this thesis, diverse types of experimental data are used to assist in the selection of *cis*-regulatory regions involved in the transcriptional regulation of gene expression. The figure highlights promoters (form which RNA production initiates) and enhancers (regions which modulate the activity of promoters). Types of experimental techniques used to collect data about the locations of *cis*-regulatory regions and the regions within which regulatory regions act are depicted.

moters has been performed. While original definitions of promoters highlighted a directionality to them, recent studies have shown that many promoters direct bidirectional transcription production (albeit most (90%) of these are still preferentially expressed in one direction)[76]. These bidirectional promoters are usually overlapped with CpG islands, and are depleted of TATA-boxes[76]. Bidirectional promoters that do not produce functional mRNA products in both directions generally produce promoter upstream transcripts, away from the 5' end of the gene[57]. These transcripts are short, and are generally sensitive to exosome-mediated decay[61].

The fact that promoters can produce bidirectional transcripts contributes to an emerging viewpoint in which promoters and enhancers (discussed below) are recognized as two ends of a continuous spectrum rather than as completely discrete categories.

### 1.2.2 Enhancers

Enhancers are DNA sequences that act upon promoters to modulate the pattern and magnitude of transcript production. Enhancer regions are composed of a mixture of TF binding sites[50]. Some of the bound TFs help recruit RNAPII proteins, or maintain chromatin (the material of which chromosomes are made, consisting mostly of DNA, RNA, and proteins) characteristics that are favorable or unfavorable for RNAPII recruitment, which ultimately influences the rate of transcriptional initiation[3]. Enhancer sequences can be found upstream, downstream, or within exons and introns[59]. Often, enhancers affect multiple genes, and most genes are affected by multiple enhancers[59]. Enhancers are often implicated in cell-specific transcription, although ubiquitous enhancers can be extensive[84].

Until recently, enhancers were distinguished from promoters in two ways - first, promoters were locations at which RNA transcripts were initiated, and second, promoters were directionally dependent and enhancers were not. With further study, the distinction between enhancers and promoters has become increasingly blurry[3]. Although conceptually different, enhancers share many properties with promoters. They are capable of being transcribed by RNAPII, producing short enhancer RNA (eRNA) transcripts[59]. This transcription is performed in a bidi-

rectional manner. Much like the promoter-upstream transcripts, eRNAs are short-lived, highly sensitive to exosome-mediated decay[3]. To further support the view that promoters and enhancers are ends of a continuum, recent studies have shown that at least a subset of promoters can function as enhancers in enhancer activity assays[3].

In the context of this work, we classify regulatory elements as either promoters or enhancers, despite the emerging biochemical data. Here, promoters contain TSS(s) for a gene of interest. Enhancers are defined as cis-regulatory regions (identified based on specific properties discussed below) that modulate the rate of transcription initiation from promoters.

## 1.3 Profiling Methods for Annotating Regulatory Properties

Since the completion of the human reference genome, current research attention has focused on its annotation. As up to 98% of the genome appears to be primarily involved in the control of gene expression, the annotation of regulatory sequences (i.e. promoters and enhancers) and chromatin modification properties (discussed below) has been given particular attention. Innovative high-throughput profiling technologies and new computational methods have proliferated, each providing insights into aspects of regulation.

### 1.3.1 TSS and Enhancer Identification

Promoter and enhancer localization is one of the main objectives of genome annotation efforts. With the completion of the human genome project and the reference genomes, locations of protein coding and non-coding RNA genes have been mapped, largely due to RNA sequencing (RNA-seq). However, the exact locations of transcript starts have long been ambiguous, as RNA-seq preferentially captures mature mRNAs. New technologies, such as Cap Analysis of Gene Expression[71] (CAGE) and Global Run-On Sequencing[21] (GRO-seq), have been developed in order to capture the capped 5' ends of RNA transcripts. These capped RNAs relate not only to mRNAs, but also to eRNA products. The newer GRO-seq technique, although more sensitive to easily degraded transcripts such as many eRNAs, is an

expensive and time-consuming procedure[66]. As only a small number of datasets are available in few cell lines, we focus on CAGE as the primary source of capped transcript identification.

First introduced in 2003 by Shiraki *et al.*[71], CAGE technology captures the 5' end of mRNA transcripts (that is–the capped portion of the mRNA) at a given time-point. These trapped ends, called tags, are sequenced and mapped back to a reference genome, delineating the specific TSS from which each mRNA transcript was produced. Efforts largely through the Functional Annotation of the Mammalian Genome (FANTOM) consortium (http://fantom.gsc.riken.jp/) have been able to collect large amounts of CAGE data across every major human organ. In this capacity, it is possible to obtain a quantitative snapshot of the human transcriptome in cell and tissue-specific contexts. At the time of publication of the consortium's fifth project (FANTOM5)[45], samples from 573 primary human cells, 152 human post-mortem tissues, and 250 cancer cell lines have been used to generate CAGE data and describe gene TSSs and their strengths[31]. The FANTOM5 CAGE data provides TSS locations and relative strength for 91% of protein coding genes (or 94% using a more permissive threshold).

Furthermore, due to the nature of the CAGE protocol, it can also be used to capture eRNAs, as many are capped at their 5' ends. The FANTOM5 project identified over 43,000 enhancers from 808 samples based on eRNA positions[4]. Many of these CAGE-identified enhancers showed expression in a cell type-specific manner, and a small portion expressed in a ubiquitous fashion.

### 1.3.2 Transcription Factor Binding

TFs are DNA-binding proteins that are involved in regulation, either by promoting or repressing transcription of genes to RNA. Activator TFs are able to recruit the RNA polymerase complex (usually with the help of other coactivator proteins or other TFs), while repressor TFs work to block RNA polymerase from initiating transcription[35]. TFs bind to both promoter and enhancer regions. Some TFs are present in all cells and are required for basic transcription. These TFs are often present in promoter regions at ubiquitous enhancers. The TATA-binding protein TF, for example, binds to TATA-box-like sequences on DNA, located upstream of

gene TSSs in about a quarter of human genes[55]. Other TFs are only present in specific types of cells or are active only at certain developmental timepoints. The GATA binding protein 2 (GATA2), for example, plays a key role in regulating hematopoietic stem and progenitor cells[65], whereas the SRY-box 2 (SOX2) is essential for maintaining stem cells in the CNS[6].

The Encyclopedia of DNA Elements (ENCODE) project[28, 29] is a public repository amassing data informative of regulation. A large part of the ENCODE project holds information on hundreds of TFs and where they bind to DNA in a variety of primary cells, tissue samples, and immortalized cell lines. Almost all of this data comes from 'ChIP-seq' experiments. Chromatin immunoprecipitation (ChIP) has become a standard technique to locate DNA-binding proteins within a cell of interest. As described by Mundade *et al.*[53], protein-DNA interactions are subjected to crosslinking; DNA is sheared and immunoprecipitation is performed with antibodies targeting TFs or other DNA-bound proteins. The recovered DNA can be sequenced to identify where in the genome the protein of interest preferentially binds; high-throughput DNA sequencing-based approaches are referred to as ChIP-seq[68]. As recovered DNA fragments are enriched at specific loci, peak-calling algorithms determine the general area in which the original protein was bound. Once a large set of DNA sequences bound by a TF are determined, computational models can be generated to detect the specific DNA sequence patterns to which the TF preferentially binds. Databases such as JASPAR[52] contain collections of these predictive TF binding models.

### 1.3.3 Histone Modifications

Histones are proteins around which DNA can be coiled in order to package large genomes into cell nuclei. A nucleosome is the core unit of chromatin, which contains 8 histone proteins and is looped twice by DNA[3]. Individual histones are subject to diverse post-translational modifications. The covalent attachment of different molecular groups to specific amino acids on specific histones can alter the structure of chromatin in the nucleus. These modifications ultimately lead to the remodelling of chromatin, where chromatin that becomes more loosely packed becomes more accessible to DNA-binding proteins and ultimately favours gene tran-

scription.

While histones may undergo numerous types of modifications (such as phosphorylation and ubiquitination), arguably histone methylation and acetylation have been the most extensively studied[7]. Similarly, while multiple amino acids present on the histones may be modified, lysine (K) residues have been the most informative of gene regulation[29]. The addition of one or more methyl groups can be a sign of transcriptional activation or repression. For example, the tri-methylation (Me3) at lysine 9 (K9) on histone H3 (together, labeled as H3K9Me3) is associated with repetitive elements and the formation of heterochromatin, while H3K4Me3 marks regions proximal to TSSs[7]. Acetylation (Ac) of lysine residues traditionally indicative of active transcription. The H3K27Ac modification marks active (as opposed to poised) regulatory regions[29]. These patterns or trends of histone modifications are observed in certain functional regions, although functional regions can be found lacking such marks, and conversely such marks can be found in other regions of the genome.

Histone modifications can be detected by ChIP-seq[53]. Such experiments have been conducted in various cell lines and primary tissues, and are available in repositories from large projects such as ENCODE[28, 29] and Roadmap[64].

### 1.3.4 Chromatin Accessibility

In general, the more tightly chromatin is packed, the more likely DNA is not being actively transcribed[35]. Chromatin remodelling proteins can unwind sections of DNA from the nucleosome complexes allowing for other DNA-binding proteins to access these regions[47]. Often, the presence of TFs in a so-called 'open-chromatin' regions is indicative of regulatory activity.

There are generally three common laboratory methods in use for detecting open chromatin regions: DNase I hypersensitive sites Sequencing (DNase-seq), Formaldehyde-Assisted Isolation of Regulatory Elements Sequencing (FAIRE-seq), and Assay for Transposable Accessible Chromatin Sequencing (ATAC-seq). First described in 2008[15], DNase-seq leverages the nuclease DNase I, which cuts double-stranded DNA. The existence of DNase I hypersensitive sites (DHSs), nucleosome-free regions of DNA, allows the DNA to be cut by the nuclease. These fragments

can be amplified, sequenced, and mapped back to a reference genome. FAIRE-seq[36] uses formaldehyde to crosslink proteins to DNA, and then DNA is sheared via sonication. Fragmented DNA is then suspended in a phenol-chloroform solution, which separates into an aqueous layer sitting atop an organic layer. DNA linked to proteins will sink to the organic layer, where nucleosome-free regions float into the aqueous layer. The sequencing step is similar to the DNase-seq method. ATAC-seq[16], developed to require less cells and significantly reduce experiment preparation time, uses a modified transposase to introduce adaptor elements into nucleosome-free regions of DNA. Tagged DNA fragments can be mapped back to a reference genome and and indicative of the transposase cut sites, which have a preference for open-chromatin regions. All three methods produce similar open-chromatin peak signals.

As the oldest method, DNase-seq data is the most represented form of chromatin accessibility data in ENCODE, however a large portion of the data were produced from immortalized cell line samples (as opposed to primary tissue). There are far fewer ATAC-seq datasets, although all of these data have been created between June 2016 and March 2017, and are highly biased towards human tissues. FAIRE-seq datasets are limited within ENCODE, but contain a mix of immortalized cell line samples and primary cell samples (https://www.encodeproject.org/matrix/?type=Experiment).

### 1.3.5 Topologically Associating Domains

Another important consideration in regulation is the 3D structure of chromatin. Intuitively, DNA regions in close proximity will be more likely to interact with one another. In 2012, Dixon *et al.*[24] coined the term topological domains (and later changed to topologically associating domains (TADs)), which are generally megabase-sized genomic regions of highly interacting regulatory elements. TADs have been found in both mice and humans covering similar genomic regions, and are therefore thought to be conserved among mammals. Similarly, the analysis of several tissues, primary cells, and cell lines show that most TADs overlap the same genomic regions, indicating that TADs tend to be consistent across tissues. Studies have shown that abnormalities in TAD boundaries or the rearrangement

of genes within them plays a role in several disease phenotypes[46], potentially indicating a disruption of interactions between regulatory regions and the intended target genes. It has therefore been proposed that most regulatory regions and their target promoter(s) will be co-localized within the same TAD.

TAD discovery is mostly achieved using a technique called high-resolution chromosome conformation capture (Hi-C)[77]. Cell DNA is crosslinked forming bonds between proximal chromatin regions. These linked regions are then ligated together and then sheared, resulting in fragments of DNA that were originally linked. DNA 'reads' are sequenced and mapped, allowing for the determination of which genomic regions have been interacting. Boundaries of TAD regions are detected at positions where the number of interactions drops[24]. Analysis of the TAD boundaries have shown to be enriched in binding sites for the CCTC-binding factor (CTCF) which is known for being largely involved in chromatin looping[24].

### 1.3.6 Computational Predictions of Regulatory Elements

Although the previous types of data are indicative of regulatory regions, it is both time-consuming and expensive to conduct experiments across all cells and tissues of interest, and to confirm the functional roles of DNA segments. However, due to the vast amounts of data now compiled, computational methods have been developed for more comprehensive labelling of regulatory regions. Two of the most popular methods, ChromHMM and Segway, can now be used independently or in conjunction to provide unsupervised machine-learned genome-wide predictions informative of regulatory potential.

First published in 2012, ChromHMM[30] uses histone modifications and CTCF-bound regions from ENCODE as primary input, as these are known to be associated with different forms of regulation. Segmentation analysis is performed using a multivariate Hidden Markov Model, assigning each segment of the genome into one of ten states (including active and inactive promoter regions, enhancer regions, insulator regions, transcribed regions, and repressed regions). Segway, similarly published in 2012[41], uses a dynamic Bayesian network to generate genome states (also ten). It factors in chromatin accessibility, certain TF binding peaks, as well as histone modifications into its predictions.

Both ChromHMM and Segway can be used to predict regulatory regions using any supplied genomes containing data from various histone modifications, chromatin accessibility and TF ChIP-seq peaks. Pre-computed predictions are available from the University of California, Santa Cruz (UCSC) Genome Browser[44] for six cell lines profiled extensively in the ENCODE project.

## 1.4 Preceding Work: Compact Promoters for Gene Delivery

It has long been known that external DNA can be introduced into cells in a specific manner. Transgenic mice, for example, can be generated with non-endogenous DNA by injecting the DNA of interest into embryos. This research has played a large part in human disease discovery and therapeutics. One such endeavour was the Gene Expression Nervous System Atlas (GENSAT) project[39], where researchers studied thousands of genes across the CNS through the insertion of bacterial artificial chromosomes (BACs). These BACs, often containing 100-200 kb of mouse DNA, can reproduce endogenous gene expression. However, within the gene of interest on the BAC, a reporter gene was placed after the target gene's coding start (ATG) sequence. By visualizing the co-expression of the reporter protein and endogenous protein, these BACs ensured the gene and all of its necessary regulatory regions could recapitulate the expected expression pattern.

The transgenic approach with long DNA sequences allows recapitulation of endogenous gene expression patterns, but the use of such long sequences is not therapeutically relevant because delivery is not feasible. Delivery by a small particle, such as a virus, restricts the amount of DNA that can be included[37]. There has been success with using small, ubiquitous promoters in viral gene therapy. While there have been numerous transgenic studies in which shorter DNA segments drive specific patterns of gene transcription, the use of compact selective promoters in gene therapy is just starting[49]. Notably, a trial to treat individuals with Leber congenital amaurosis-2, a childhood eye disorder that leads to blindness, used a 1,400 bp sequence from the RPE65 gene within a AAV2 vector to drive selective expression of the RPE65 protein[20].

Our lab has been pursuing the development of sets of compact promoters suit-

able for selective patterns of gene delivery. The systematic design of MiniPromoters (human regulatory sequences of ∼4 kb or less, which promote cell type and tissue-specific expression) was first described by Portales-Casamar *et al.* in 2010[60] and in a follow-up study by de Leeuw *et al.* in 2014[22]. The goal was to identify human cis-regulatory regions (RRs) in genes targeting the CNS. This approach was based primarily on the identification of non-coding, highly conserved genomic regions closeby a gene of interest. TF binding site predictions were generated across these conserved regions, and used to suggest functional roles for TFs relevant to CNS regions of interest. These regions of interest were fused with promoter regions of the same gene, and the resulting MiniPromoters were assessed in transgenic knockin mice using a procedure that placed the MiniPromoters and a reporter gene at a specific location on the X-chromosome.

This primary work was the basis for the 2016 paper by de Leeuw *et al.*[23], who began to use MiniPromoters packaged in recombinant AAV2/9 hybrid vectors. Many tested MiniPromoters were those found to express in the previous transgenic mouse initiatives. Newly designed MiniPromoters defined RRs similarly to methods described above, but also included the use of DHSs, TF ChIP-seq peaks, regions of specific histone modifications, regions of high conservation. Hickmott *et al.*[40] used the newer MiniPromoter design strategy to find RRs for the PAX6 gene. This paper introduced the use of TADs to constrain the search space for RRs. It also introduced the use of CAGE data for identifying TSSs, which resolved ambiguity of choosing an appropriate promoter RR in the case of genes having multiple transcripts.

## 1.5 Hypothesis

Compact *cis*-regulatory sequences can be computationally designed based on annotated properties of the genome that overlap designs generated by human experts in a painstaking and time consuming process. Further, the use of annotated properties relevant to the tissue of desired expression will improve automated design success. Based on these hypotheses, I have taken the following approaches in this thesis to establish and assess a semi-automated bioinformatics procedure for the design of compact promoters for use in AAV-based viral vectors. I manually de-

signed a set of MiniPromoters based on sets of regulatory features, with the goal of defining a reference set of designs against which a semi-automated procedure could be assessed. A subset of these were tested through *in vivo* experiments in mice. I then designed a semi-automated approach inspired by the manual design process, amalgamating thousands of experiments informative of genome regulation in cell-specific contexts in order to predict key RR consistent with the qualitative assessments of a trained designer. From here, I validated the capacity of the semi-automated procedure to reproduce the bespoke designs.

# Chapter 2

# Methods

## 2.1 Data

All datasets used in the manual design MiniPromoters as well as the automatic regulatory region identification were previously published and are available publicly. The CAGE datasets (TSS annotations and enhancer annotations) and ontology were obtained from the FANTOM5 consortium[4, 45]. Hi-C, TF ChIP-seq, histone modifications, DNase-seq, and FAIRE-seq data were obtained from the ENCODE project[28, 29]. Additional TF histone modification and DNase-seq data were obtained from RoadMap[64]. Additional Hi-C data was obtained from GEO (accession number: GSE87112). Gene annotations from RefSeq[58, 62], ChromHMM[30] and Segway[41] chromatin states, repeat regions identified with RepeatMasker[73], and PhastCons and PhyloP (http://compgen.cshl.edu/phast/) scores were obtained from the UCSC Genome Browser[44] tables. All data were based on the hg19 reference human genome.

Eight bespoke MiniPromoters (Ple360, Ple366, Ple367, Ple368, Ple370, Ple371, Ple372, Ple373) were designed to include sequences contained within previously published reporter gene-containing BAC constructs (RP24-269I17, RP23-234I17, RP23-440L10, RP24-98L14, RP23-281A14, RP24-260F14, RP23-305H12, and RP24-285B17, respectively). This mouse BAC data was obtained from GENSAT[39] and Mouse Genome Informatics (MGI)[12]. The published reporter gene activity indicated that the BAC region contained sufficient and proper *cis*-regulatory ele-

ments to drive endogenous gene expression.

## 2.2 Bespoke MiniPromoter Construct Design

The bioinformatics for MiniPromoter and RR design has been described[23, 40]. Briefly, this process involves the selection of a gene of interest, specification of a promoter region, and in a subset of cases the selection of one or more enhancer regions. All genes selected for MiniPromoter design must include a TSS that is supported by experimental evidence indicating gene expression in a relevant cell or tissue. TSS identification is based on CAGE data. CAGE reads were extracted for each TSS of each gene using the Zenbu browser (for visual comparison, http://fantom.gsc.riken.jp/zenbu/) or the SSTAR view (for numerical comparison, http://fantom.gsc.riken.jp/5/sstar/). TADs were used to delineate boundaries within which searches for RRs were constrained. A consensus TAD region was determined visually by taking the overlap between TAD data from a H1 human embryonic stem cell line and the IMR90 (lung fibroblast) cell line therefore creating a consensus TAD. In certain cases, relevant mouse BACs were used to narrow this search space, if the reporter gene co-expressed with the endogenous protein in published studies. The BAC coordinates were then converted from mouse into human coordinates using the UCSC Genome Browser LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

RR identification was based on visual assessment of data (see above) displayed within the UCSC genome browser, including the following tracks: RefSeq genes, FANTOM5-identified enhancers,TF ChIP-seq peaks, DNaseI hypersensitive clusters, histone modification marks, computational predictions from ChromHMM/Segway, multi-species conservation, and RepeatMasker. RefSeq genomic annotations were reviewed to ensure all RRs excluded known open reading frames or splice sites. A set of 32,693 FANTOM5 enhancers included within SlideBase (from the original 65,423 FANTOM5 set; http://slidebase.binf.ku.dk) were included. ChIP-seq experiments provided by ENCODE were limited to a set of 161 TFs that included Factorbook motifs[78, 79]. While the DHS experiments were performed on 125 cell lines (ENCODE V3), the data was used to predict which areas of the genome would be more likely to be open regardless of cellular context as well as

areas open in only specific types of cells. H3K4Me1 and H3K27Ac were used to identify both active and poised RRs. Combined ChromHMM/Segway predictions across six common cell lines were used to identify insulated (CTCF) regions of the genome, in order to constrain the RR search region. Two types of conservation tracks were used: the 100-vertebrate base pair-conservation track by PhyloP score (to identify non-exonic genomic regions indicative of important genomic elements without introducing a large bias based on more closely-related primate species) and the Multiz Alignments[13] of the rhesus and mouse genomes (under a hypothesis that conserved sequence would increase the likelihood that designs using human sequence would be functional in subsequent in vivo analyses in mouse and rhesus). The RepeatMasker track was used to remove RRs that contained short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), and long terminal repeats (LTRs).

RR boundaries were chosen qualitatively based on amount and types of overall evidence present in the search space. It was determined that regions which overlapped large amounts of TFs, DHS clusters, and had high H3K27Ac activity were marks of general, ubiquitous enhancers. Many of the self-identified cell-specific enhancers were enriched in specific TFs known to be present in the cell-type of interest, were regions of high conservation, contained a FANTOM5 enhancer that was linked to a TSS present in the promoter RR, or a combination of these features. Boundaries were chosen conservatively, constraining RRs to contain the most overlap of chosen features, in order to minimize the size of each region.

### 2.2.1 RR Selection

Identified candidate regions were presented to a team of scientists for consideration. Each presented RR had to contain one or more forms of evidence mentioned in the previous section. These regions were then ranked based on their perceived likeliness to be an enhancer–either ubiquitous or in a cell specific manner. Other factors contributing to RR selection included region size, past description in published literature, or similarity of features to the selected promoter sequence. In total, the selected promoter and any additional RRs could not be more than 2.7 kb in size due to the AAV payload restriction.

### 2.2.2 MiniPromoter Assembly

MiniPromoter RRs were assembled in the 5' to 3' direction. If the RR was located endogenously on the antisense strand, the reverse complement of the RR sequence was used. Promoter RRs were always placed at the most 3' end of the MiniPromoter designs. Enhancer RRs were added where the more distal upstream RRs (from the endogenous promoter RR) were placed closest to the 5' end of the construct. Additionally, all RRs located endogenously downstream of the promoter were placed at the 5' end, regardless if the region was more proximal than an upstream RR. Finally, the addition of two restriction enzyme sites were added to the 5' (Fse recognition sequence) and 3' (AscI recognition sequence) ends of the construct in order to properly clone the MiniPromoter sequence into a vector plasmid.

## 2.3 Experimental Validation of MiniPromoters

Virus production, injections into mice, mouse harvesting, immunostaining, and imaging methods have been previously described by de Leeuw *et al.*, with the following amendments: Only wild type mice were used for testing MiniPromoters, with the virus injected into the superficial temporal vein of mice at two time points (either postnatal day 0 or postnatal day 4). These dates were chosen based on optimal injection time point studies by Byrne *et al.* Control mice were injected with $3.3 \times 10^{12}$ vg/mL (viral genomes per milliliter). Mice were harvested 4 weeks post-injection (at time points P28 or P32). In addition to retaining the brain, eyes, spinal cord, and heart for image analysis, the liver and pancreas were also studied. All other methods (virus preparation, animal injections, and fluorescent imaging processes) were performed using the procedure outlined for emerald green fluorescent protein (EmGFP) constructs in de Leeuw *et al.*

For the study of two viruses containing MiniPromoters targeting the corneal epithelium (based on the KRT12 gene), a different protocol was followed. Both viruses were of AAV9 serotype, and each contained one of three different promoters (outlined below) to express the EmGFP transgene. All injections were performed intrastromally on adult mice (ages ranged between 2-4 months). Injections were all $2\mu$L, and contained $5 \times 10^{12}$ vg/mL and a 1:20 dilution of stock of FluoSpheres. Left eyes of nine mice were injected in this study; three eyes

were injected for each type of virus created for this experiment. All uninjected (contralateral) eyes were used as negative controls.

Three eyes were injected intrastromally for each construct, including the ubiquitous small chicken beta actin (smCBA) promoter, Ple326 and Ple334. Tissues were harvested 6 days post-injection. All eyes (injected left eyes and uninjected right eyes) were embedded in Tissue-Tek O.C.T. compound and sectioned at 20$\mu$m on a Microm HM550 cryostat. A subset of these section (generally 2-4 sections per experiment) were pressed, and rinsed in 0.1M phosphate buffer saline (PBS) twice, for five minutes each. After rinsing the sections for five minutes in 0.1M PBST (PBS + Triton X-100), they were blocked for 30 minutes before being incubated overnight at room temperature in a primary EmGFP antibody stain (at a 1:500 dilution). The following day, the sections were rinsed again in 0.1M PBST three times for ten minutes. Sections were then additionally incubated and stained with a secondary antibody (Alexa448 conjugated antibody at 1:1000 dilution) and co-stained with Hoechst33342 dye (at 1:1000 dilution) for one hour at room temperature. Finally, all sections were washed in a 0.1M phosphate buffer (PB) three times for ten minutes and in 0.01M PB for ten minutes, and were mounted with ProLong Gold Antifade Mountant. All stained sections were then imaged at three different colour channels (DAPI–blue, for Hoechst; TXRED–red, for FluoSpheres; FITC–green, for EmGFP) on an Olympus BX61 fluorescence microscope through the software cellSens at either 10x or 20x magnification. Each image was further processed into composite and single-colour TIFF images using the freeware program ImageJ and its Bio-Formats plugin.

## 2.4 Semi-automated RR Selection

### 2.4.1 Promoter Selection

Importantly, each defined promoter RR must contain at least one TSS. After receiving a valid HGNC gene name, OnTarget retrieves each identified FANTOM5 TSS stored in its underlying database (where each TSS is required to have at least one tag in at least one sample, out of 1,829 possible samples). Each TSS is then extended in both the upstream and downstream direction in order to achieve a min-

imal promoter length.

Downstream, each sequence is extended until one of the following conditions are met:

1. If a TSS is located before the annotated gene start and
   (a) if the annotated coding start site (CDS) is not in the first exon, the TSS will be extended through until the end of the first exon, minus a splice site offset (default of 10 bp);
   (b) if the annotated CDS is in the first exon, the TSS will be extended through until the CDS, minus a KOZAK sequence offset (default of 10 bp);
2. If a TSS is located within an exon before the annotated CDS and
   (a) the CDS is not located in the same exon, the TSS will be extended through until the end of the exon, minus the splice site offset;
   (b) the annotated CDS is in the same exon, the TSS will be extended through until the CDS, minus a KOZAK sequence offset;
3. If a TSS is located within a gene intron before the annotated CDS and
   (a) the annotated CDS is not in the following exon, the TSS will be extended through until the end of the following exon, minus the splice site offset;
   (b) the annotated CDS is in the following exon, the TSS will be extended to the CDS, minus the KOZAK sequence offset;
4. If the TSS is in an intron downstream of the annotated CDS, the TSS will be extended through until the end of the following exon, minus the splice site offset;
5. If the TSS is in a coding exon, the TSS will be extended until the end of the exon, minus the splice site offset.

This downstream expansion is then tested for unwanted elements, such as ATG sequences (which could create possible ORFs) or other FANTOM5 annotated (unextended) TSSs. The extensions are trimmed to no longer include any of these unwanted elements. In the case of ATG sites, these are only trimmed if they fall within the annotated gene area.

Upstream, each TSS is extended until one of the following conditions are met:

1. If the TSS is located before the annotated gene or within the gene but before the annotated CDS in the first exon, and
   (a) there is another annotated FANTOM5 TSS (unextended) further upstream, the TSS will be extended until 1 bp before the closest upstream TSS;
   (b) there is no other FANTOM5 TSS further upstream, the TSS will be extended until the Phastcons conservation score falls below a threshold (default 60%);
2. If the TSS is located within an exon and
   (a) there is another annotated FANTOM5 TSS (unextended) further upstream within the same exon, the TSS will be extended until 1 bp before the closest upstream TSS;
   (b) there is no other TSS within the same exon, the TSS will be extended up until the start of the exon, excluding nucleotides within the splice site offset;
3. If the TSS is located within an intron and
   (a) there is another annotated FANTOM5 TSS (unextended) further upstream within the same intron, the TSS will be extended until 1 bp before the closest upstream TSS;
   (b) there is no other TSS within the same intron, the TSS will be extended up until the start of the intron, excluding nucleotides within the splice site offset

Each minimal promoter is returned to the user, where it is possible to combine multiple promoters into an extended promoter RR. This can be done as long as each minimal promoter neighbours another desired minimal promoter without overlapping splice and KOZAK sites.

### 2.4.2 Enhancer Selection

An enhancer RR is described as a region lacking any annotated FANTOM5 TSSs. For a given search space, RRs are selected based on the overlap of regulatory features. An underlying feature matrix and weighting vector defines boundaries and provide each RR with a score. The higher scoring regions contain the most informative data indicative of regulation. The procedure with default settings (all of which can be adjusted) is described below.

The underlying Data Repository of OnTarget stores cell or tissue type-specific data for the following features:

1. Hi-C TAD datasets from 33 cell lines and tissue samples: As this is our smallest data-set, often all TADs are taken into consideration, and the consensus TAD is chosen for delineating a search space.

2. 1,284 TF ChIP-seq experiments that cover 145 primary cell/tissue types and cell lines: When creating ubiquitous RR profiles, each TF track is condensed into a single vector based on presence or absence, and can be cell type specific or agnostic. These vectors are then summed into one consensus TF vector.

3. Chromatin accessibility data based off DNase-seq and FAIRE-seq in 301 primary cell/tissue types and cell lines: Should a specific cell type be unavailable, accessibility data across all datasets are used as a consensus tracks.

4. Histone modification data for 33 histone modification signatures in 197 primary cell/tissue types and cell lines: When a specific cell type is unavailable, data across all datasets for a specific histone and modification are used as a consensus tracks. We primarily focus on H3K27Ac for regions indicative of active enhancer elements. H3K4Me2 marks are also considered, however

this repressor mark is negatively associated to expression and is therefore negatively affects RR assignment.

5. FANTOM5 experimental enhancers: certain enhancers show activity in a variety of cell and tissue types, while others have not been shown to associate to any particular location. These non-specific enhancers appear regardless if a certain cell or tissue type is selected.

Additionally, we include the following cell-type agnostic datasets into the RR identification pipeline:

1. Per-base conservation data across 100 vertebrates with PhastCons scores: Unlike the other RR features where data is stored in BED files, PhastCons nucleotide scores are implemented as a Wiggle track. This method is taken from the UCSC Genome Browser method, which has used the the PHAST package described online (http://compgen.cshl.edu/phast/).

2. Repeated elements from RepeatMasker: We consider 3 out of the possible 10 assignments from this data source. RRs overlapping SINE, LINE, and LTR elements are excluded from further analysis.

Once appropriate cell type data is selected, the RR method begins by defining a search space based on the TAD track. The chosen TAD is the one in which a gene of interest is located. While TAD boundaries are the default search space for each iteration of RR identification, this can be changed to a defined chromosomal range or the intergenic region between the gene of interest and its closest up and downstream annotated RefSeq genes.

A search space consists of a defined number of nucleotides $n$ and default features $f$. A $[n \times f]$ feature matrix $M$ is created and initiated with zeros in all cells. As shown in Figure 3a, each cell of the matrix represents the presence (1) or absence (0) of a feature. M is then multiplied by a weight vector, of size $[1 \times f]$, where each feature is assigned a default weight corresponding to its overall importance in RR identification. Selection of values for the weight vector is discussed in the Results section. The columns of the new M matrix are summed, creating the sum vector of size $[n \times 1]$. Each position in the sum vector is then multiplied by

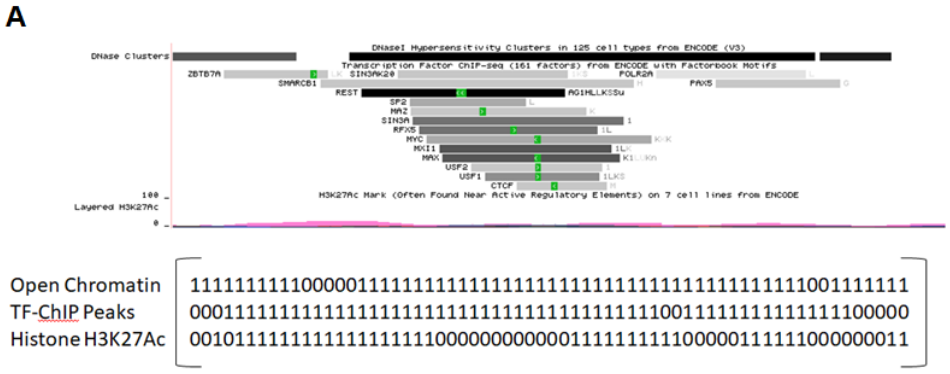
individual mask vectors (also of size $[n \times 1]$). Mask vectors act as absolute features that must be present or absent in each identified RR. Two default mask vectors are the coding exon mask and the RepeatMasker mask. By default, coding exons are excluded from RR identification, and are represented by 0s in the mask. Similarly, SINE, LINE, and LTR elements receive a representation of 0 in this mask, in order to exclude these regions from RR identification. By multiplying the mask vectors to the sum vector results in the score vector $S$, which contains the final score of each nucleotide in the search space (Figure 3b).

Segments of qualifying positions are reported when 10 or more contiguous nucleotide scores pass the threshold (Figure 3c). This threshold is calculated from the distribution of scores from each $S$ vector. At each run, the score at the $99^{th}$ percentile is chosen. Regions scoring equal or above this threshold are reported as RRs.

## 2.5 Validation of Semi-automated Design Performance

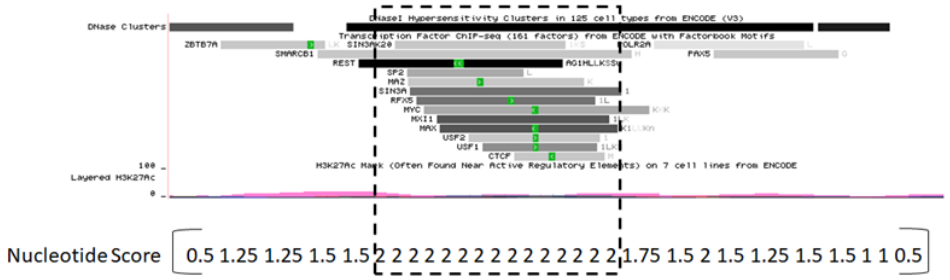
We evaluate OnTarget based on two expectations. First, we expect RRs identified by using an accumulation of all data to be different than those identified by using cell or tissue-specific data. Second, OnTarget should detect RRs from successful bespoke MiniPromoter designs.

In my first experiment, I decided to use liver and hepatocyte datasets, as these are the most abundant datasets among ChIP-seq (for TFs and histone marks) and DHS experiments. No cell line data (*i.e.* HepG2 cells) were used in the cell-specific analysis. I then chose two different TADs. One TAD (hg19:chr7:87,000,001-87,802,064; an $\sim$800 kb region) contained the gene ABCB4, known to express in the liver and hepatocytes. This gene was chosen by searching the Human Protein Atlas (http://www.proteinatlas.org/) for all genes that almost exclusively expressed both RNA and protein in the liver, across datasets from the Protein Atlas, the Genotype-Tissue Expression project (GTEx: http://www.gtexportal.org/), and FANTOM5. The gene GYS2 was originally identified for analysis as the first gene to fit the criteria, although it was discarded because this was the only gene located within its TAD. A similar selection process was chosen for a separate TAD. The gene NOS1 was chosen using the same procedure, appearing in the set of genes




**(a)** The OnTarget 0-1 matrix.


**(b)** Each nucleotide receives a score based on its features.




**(c)** Contiguous high-scoring nucleotides are reported back as regulatory regions.

**Figure 3:** Visual Representation of the OnTarget Selection Module. The top UCSC data tracks allow for the visualisation of each feature corresponding to one nucleotide. **A.** At each nucleotide position, a feature is either present (represented by a 1), or absent (represented by a 1). **B.** After being multiplied by the importance weighting of each vector, all features are summed, resulting in a final score for each individual nucleotide. **C.** Contiguous high-scoring regions are reported as a potential regulatory region.

listed as 'not expressed' in all liver samples across the same three datasets. This TAD (hg19:chr12:117,640,001-118,475,617; an ∼835 kb region) covers mostly brain-expressing and housekeeping genes.

My second experiment used a subset of eye, brain, and neuron datasets, as our bespoke MiniPromoters were used to target cells within the eye and brain. I tested OnTarget for its capacity to predict the component regions of three MiniPromoter constructs: one successful design, one unsuccessful design, and one design awaiting testing.

# Chapter 3

# Results

## 3.1 Bespoke Designs

Bioinformatics analysis procedures were established for the delineation of RRs in human genes, which are described in the Methods. A total of 49 MiniPromoters were designed based on detailed analyses of 35 genes. Approximately 50 additional genes were partially analyzed but discontinued due to endogenous expression pattern, lack of a homologous gene pair between human and mouse, or because a AAV-suitable design was already reported in the literature. Genes were determined by expression data (predominantly CAGE and literature-derived data such as Drop-seq[48]) within a target cell or tissue of interest. Table 1 shows a list of all designed MiniPromoters. Most designs fit within the 2.7 kb limit, with the exception of Ple346 (NEFM gene base) which spanned 2,711 bp. All other MiniPromoters ranged in size from 331 bp to 2,700 bp. The average design size was 1.71 kb. A subset of 40 designs incorporated at least one enhancer RR in addition to a promoter RR. Two of the designs (Ple326 and Ple334) used a total of six RRs (promoter inclusive), which was the most RRs included in any design. As the project progressed, and the importance of extremely compact MiniPromoters emerged, there was a trend to shorter designs.

**Table 1:** List of all designed MiniPromoters between January 2016 and July 2017. All tested designs are described in Table 2

| Design number | Gene | Target cell/tissue | MiniPromoter size (bp) | Tested | Number of regulatory regions |
|---|---|---|---|---|---|
| Ple326 | KRT12 | corneal epithelium | 2,313 | Y | 6 |
| Ple328 | PAX6 | amacrine, horizontal, Müller glia, ganglion cells | 2,148 | Y | 3 |
| Ple329 | PAX6 | amacrine, horizontal, Müller glia, ganglion cells | 2,513 | Y | 3 |
| Ple330 | PAX6 | amacrine, horizontal, Müller glia, ganglion cells | 1,982 | Y | 2 |
| Ple331 | PAX6 | amacrine, horizontal, Müller glia, ganglion cells | 1,982 | Y | 2 |
| Ple332 | KCNJ8 | pericytes | 2,100 | Y | 3 |
| Ple333 | ABCC9 | pericytes | 2,332 | Y | 3 |
| Ple334 | KRT12 | corneal epithelium | 2,326 | Y | 6 |
| Ple338 | CLDN5 | endothelial cells | 2,567 | Y | 4 |
| Ple339 | CLDN5 | endothelial cells | 1,973 | Y | 2 |
| Ple340 | CLDN5 | endothelial cells | 2,700 | Y | 4 |
| Ple341 | PCP2 | bipolar cells | 784 | Y | 2 |
| Ple342 | TUBB3 | retinal ganglion cells | 1,992 | Y | 2 |
| Ple343 | TUBB3 | retinal ganglion cells | 2,669 | Y | 3 |
| Ple344 | TUBB3 | retinal ganglion cells | 801 | Y | 2 |
| Ple345 | NEFL | retinal ganglion cells | 2,693 | Y | 5 |
| Ple346 | NEFM | retinal ganglion cells | 2,711 | Y | 5 |
| Ple347 | GNGT2 | cones | 1,197 | Y | 2 |
| Ple348 | PDE6H | cones | 2,025 | Y | 3 |
| Ple349 | PDE6H | cones | 2,005 | Y | 4 |
| Ple350 | AQP4 | Müller glia | 1,802 | N | 2 |

*Continued on next page*

| Design number | Gene | Target cell/tissue | MiniPromoter size (bp) | Tested | Number of regulatory regions |
|---|---|---|---|---|---|
| Ple351 | GPR37 | Müller glia | 1,890 | N | 2 |
| Ple352 | TACR3 | bipolar OFF subtypes BC1A, BC1B, BC2 | 2,643 | N | 2 |
| Ple353 | GRIK1 | bipolar OFF subtypes BC2, BC3A, BC3B, BC4 | 2,367 | N | 3 |
| Ple354 | GRIK1 | bipolar OFF subtypes BC2, BC3A, BC3B, BC4 | 2,646 | N | 3 |
| Ple355 | ADORA2A | striatum | 2,666 | N | 3 |
| Ple356 | DBH | locus coeruleus | 2,479 | N | 4 |
| Ple357 | DRD1 | striatum | 2,200 | N | 4 |
| Ple358 | DRD2 | striatum | 1,659 | N | 2 |
| Ple359 | DRD2 | striatum | 2,680 | N | 4 |
| Ple360 | SLC6A3 | substantia nigra | 2,322 | N | 2 |
| Ple361 | PTPN3 | thalamus | 2,092 | N | 3 |
| Ple362 | RGS16 | thalamus | 2,027 | N | 5 |
| Ple363 | PDGFRB | pericytes | 846 | N | 1 |
| Ple364 | PDGFRB | pericytes | 1,396 | N | 2 |
| Ple365 | PDGFRB | pericytes | 730 | N | 1 |
| Ple366 | CCK | GABAergic neurons | 1,469 | N | 1 |
| Ple367 | DLX1 | GABAergic neurons | 970 | N | 1 |
| Ple368 | GAD2 | GABAergic neurons | 1,091 | N | 2 |
| Ple369 | SST | GABAergic neurons | 681 | N | 2 |
| Ple370 | CORT | GABAergic neurons | 399 | N | 2 |
| Ple371 | DLX5 | GABAergic neurons | 595 | N | 2 |
| Ple372 | PVALB | GABAergic neurons | 832 | N | 2 |
| Ple373 | CX3CR1 | microglia | 372 | N | 1 |
| Ple374 | P2RY12 | microglia | 505 | N | 1 |

*Continued on next page*

| Design number | Gene | Target cell/tissue | MiniPromoter size (bp) | Tested | Number of regulatory regions |
|---|---|---|---|---|---|
| Ple375 | P2RY12 | microglia | 943 | N | 1 |
| Ple376 | TMEM119 | microglia | 651 | N | 1 |
| Ple377 | TREM2 | microglia | 717 | N | 2 |
| Ple378 | TYROBP | microglia | 331 | N | 1 |

## 3.2 Experimental Validation of Bespoke Designs

Twenty MiniPromoters have been tested in young mice at P0 and P4. Two of the 20 MiniPromoters were additionally tested in adult mice. Table 2 summarizes the results of the tested MiniPromoters. While the brain, eyes, spinal cord, heart, liver, and pancreas were all analyzed, only expression in targeted tissues will be discussed.

**Table 2:** MiniPromoters tested in mice *in vivo*. * Off-target expression observed along with expected expression. † Expected expression not observed, experiments still ongoing. ‡ A subset of expected expression observed. U Unknown. A positive control could not be established for the target cell type, and therefore success or failure could not be accurately determined.

| Design number | Gene | Target cell/tissue | Actual expression | Success |
|---|---|---|---|---|
| Ple326 | KRT12 | corneal epithelium | corneal stroma | U |
| Ple328 | PAX6 | amacrine, horizontal, Müller glia, ganglion cells | amacrine, horizontal, ganglion cells | Y‡ |
| Ple329 | PAX6 | amacrine, horizontal, Müller glia, ganglion cells | amacrine, horizontal, ganglion cells | Y‡ |
| Ple330 | PAX6 | amacrine, horizontal, Müller glia, ganglion cells | amacrine, horizontal, Müller glia, ganglion cells, ganglion cells | Y |

*Continued on next page*

| Design number | Gene | Target cell/tissue | Actual expression | Success |
|---|---|---|---|---|
| Ple331 | PAX6 | amacrine, horizontal, Müller glia, ganglion cells | amacrine, horizontal, Müller glia, ganglion cells, ganglion cells | Y |
| Ple332 | KCNJ8 | ocular pericytes | N/A | N |
| Ple333 | ABCC9 | ocular pericytes | N/A | N |
| Ple334 | KRT12 | corneal epithelium | N/A | U |
| Ple338 | CLDN5 | endothelial cells | endothelial cells, horizontal cells | Y* |
| Ple339 | CLDN5 | endothelial cells | endothelial cells, horizontal cells | Y* |
| Ple340 | CLDN5 | endothelial cells | endothelial cells, amacrine cells | Y* |
| Ple341 | PCP2 | bipolar cells | bipolar cells | Y |
| Ple342 | TUBB3 | retinal ganglion cells | retinal ganglion cells, amacrine cells | Y* |
| Ple343 | TUBB3 | retinal ganglion cells | retinal ganglion cells, amacrine cells | Y* |
| Ple344 | TUBB3 | retinal ganglion cells | retinal ganglion cells | Y |
| Ple345 | NEFL | retinal ganglion cells | retinal ganglion cells | Y |
| Ple346 | NEFM | retinal ganglion cells | retinal ganglion cells | Y |
| Ple347 | GNGT2 | cones | cones (including cone bipolar cells) | Y |
| Ple348 | PDE6H | cones | retinal ganglion cells, amacrine cells | N$^{\dagger}$ |
| Ple349 | PDE6H | cones | retinal ganglion cells, amacrine cells | N$^{\dagger}$ |

Ple326 and Ple334, based off the KRT12 gene, were tested for their capacity to direct reporter gene expression from AAV preparations by temporal vein injection of AAV and injection into the corneal stroma in adult mice. Both MiniPromoters contained the same five enhancer RRs, while their promoter RRs were based off

two distinct FANTOM5 TSSs (see Figure 4). Corneas injected with Ple326 showed expression in the corneal stroma, at levels below that directed by the smCBA-EmGFP control virus. Mice injected with Ple334 showed no apparent EmGFP expression throughout the corneal stroma. As displayed in Figure 5, the bespoke MiniPromoters and positive control could not direct observable expression in the epithelial layer (where expression was anticipated for Ple326 and Ple334). This result, therefore, does not indicate a success or failure of Ple326 or Ple334, as we were unable to determine a baseline expression pattern with which to compare. These MiniPromoters are the only ones of the design set to contain undetermined results. Expectedly, P0 and P4 mice showed no expression after harvest, as the corneal epithelium is not fully formed until P12-14, when mice first open their eyes[17].

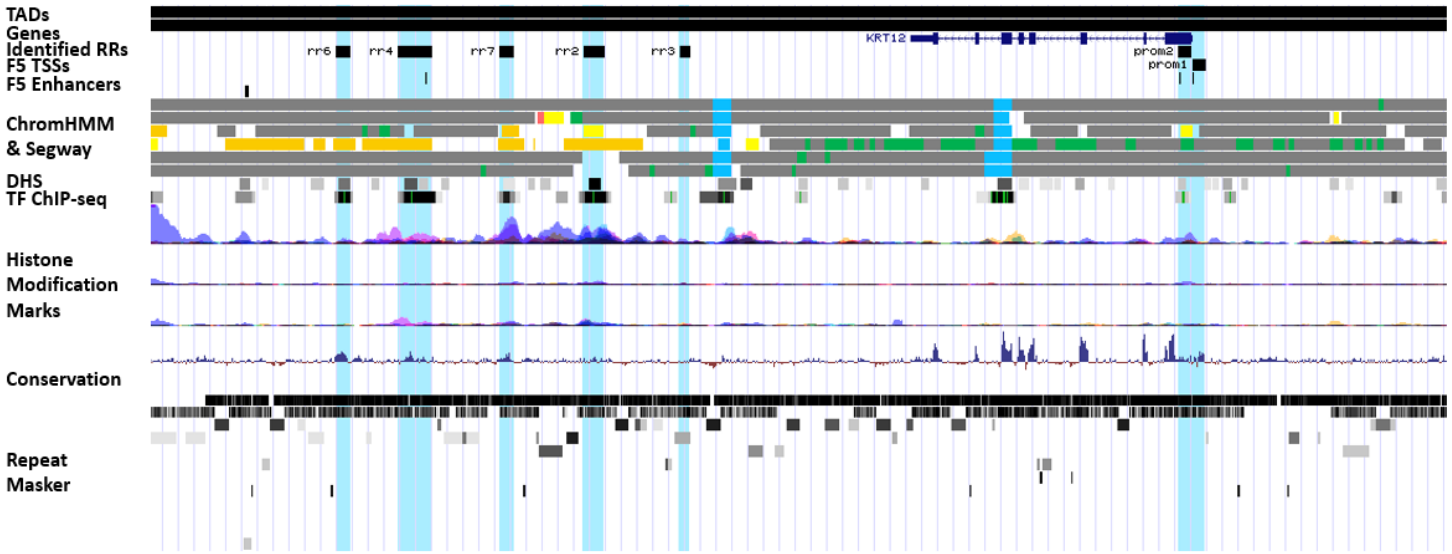



**Figure 4: Manual Bioinformatics Design of Ple326 and Ple334 Novel MiniPromoters from the KRT12 Gene.** The blue highlights indicate the RRs selected for use in both Ple326 and Ple334. Each MiniPromoter uses the same RRs, but different promoters. Promoter regions were based off two distinct FANTOM5-identified TSSs. Additional RRs were chosen based on the overlap of DHS, TF ChIP-seq, histone mark, and conservation data. One identified RR overlaps another FANTOM5 TSS, however its expression was considered negligible upon further inspection.

Ple328, Ple329, Ple330 and Ple331 were based upon the PAX6 gene. It was predicted that the PAX6 promoter RR and additional enhancer RRs could restrict expression to four specific cell types in the retina (see Table 2): amacrine cells, horizontal cells, ganglion cells, and Müller glia. Ple328 and Ple329 contained 2 out of 3 of the same RRs (one enhancer RR and the promoter RR). Ple330 and Ple331 contained 2 RRs each, and were exactly the same construct, except for a 8 bp change in a PAX6 TF binding site in the enhancer RR. Previous PAX6 MiniPromoters could only achieve expression in combinations of three out of four cell types. Ple330 and Ple331 were able to achieve expression in all four expected cell types, although expression levels of EmGFP were stronger in the latter. In Ple328, there was no obvious expression of Mller glia. Some injections of Ple329 covered all four cell types, although it was not as clear as the expression seen in Ple331.

Ple332 and Ple333 were based off KCNJ8 and ABCC9 respectively, and were designed to target eye pericytes. Both genes are located adjacently in both human and mouse genomes, and encode components of the same potassium channel. Due to this reasoning, both MiniPromoters used the same two RRs, and promoter RRs were designed to incorporate the TSS of each gene. FANTOM5 expression levels of each gene suggested that both constructs would be very lowly expressed within the eye. After imaging of mouse eyes at both timepoints, no clear expression was found for these two MiniPromoters.

Ple338, Ple339, and Ple340 were based off the CLDN5 gene, and were partial re-designs of an old MiniPromoter design (Ple32, data not shown) to target CNS endothelial cells. All three MiniPromoters contained the same promoter RR, which was a cut-down version of Ple32, which contained only a promoter RR. Ple338 additionally was packaged with three other enhancer RRs, Ple339 was packaged with one additional enhancer RR, and Ple340 also contained three different enhancer RRs. Each enhancer RR was different, and enhancer RRs were grouped by predicted linkage to the CLDN5 TSS, the inclusion of a FANTOM5-derived enhancer, and by regions not found (conserved) in the mouse genome, respectively. All MiniPromoters drove expression in endothelial cells. Ple340 was found to also express in off target locations including amacrine and bipolar cells in the eye. Ultimately, the original MiniPromoter (Ple32) had the strongest expression with the

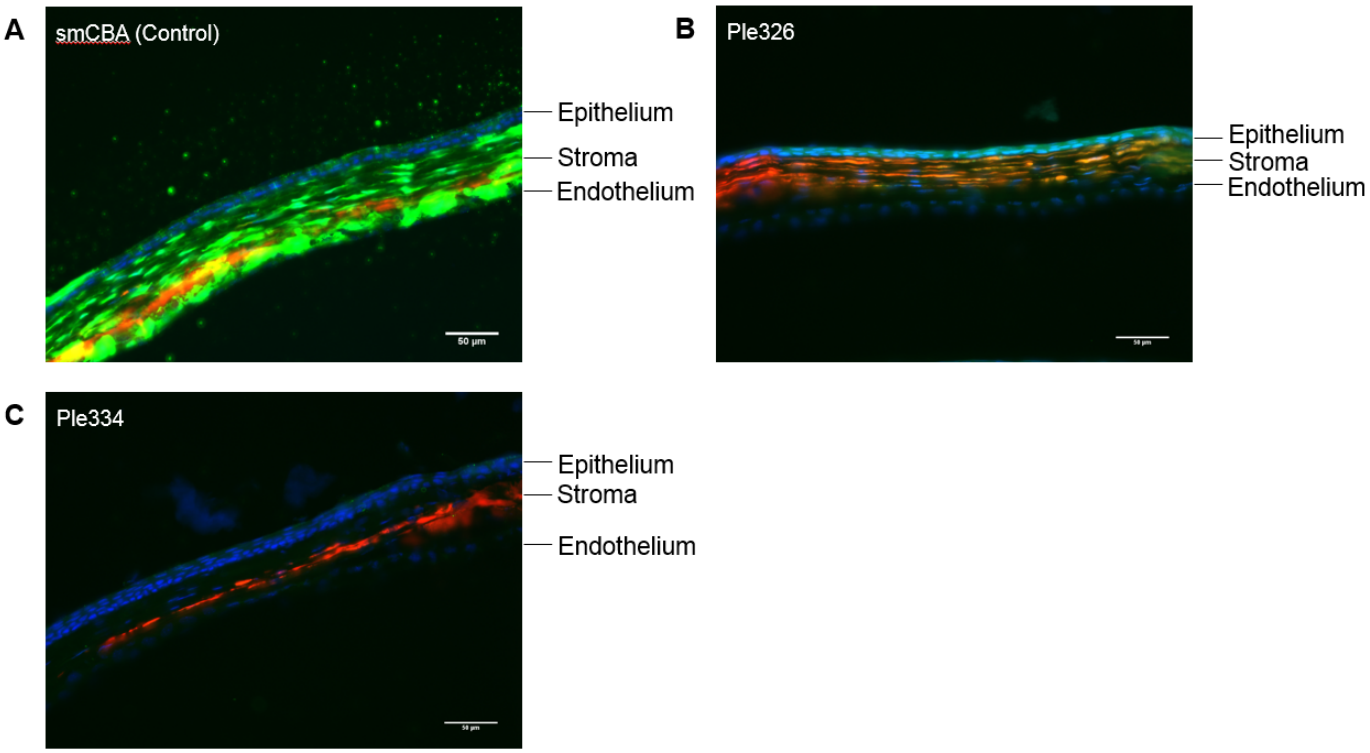



**Figure 5: MiniPromoters Ple326 and Ple334 from the KRT12 Gene Drive Gene Expression in Layers of the Cornea.** The smCBA (A) promoter was tested against Ple326 (B) and Ple334 (C). All intrastromal injections included the EmGFP reporter protein and FluoSpheres, injected into adult mice. All eyes were harvested six days post-injection. Cell nuclei are visible in blue with Hoechst33342. EmGFP antibodies are visible in green. FluoSphere locations are visible in red. **A.** EmGFP expression is seen strongly in the stroma and endothelium layers of the cornea. One line of antibody stain can be seen overlapping the epithelial layers, however it was undetermined if this was true EmGFP or an artifact, due to not seeing this pattern anywhere else along the cornea surface over three replicates. **B.** There is some overlap in the stroma with EmGFP, although it is much weaker than the smCBA promoter. No obvious EmGFP expression seen in the epithelium or endothelium layers. **C.** No apparent expression of EmGFP in any layer of the cornea.

least amount of off-target expression, indicating that the promoter RR is enough to reproduce endothelial expression.

Ple341, based on the PCP2 gene (see Figure 6), was a cut-down version of an older MiniPromoter design (Ple265) to target bipolar ON cells. While Ple265 was composed of only one RR, Ple341 contained a smaller promoter RR, accompanied by a small enhancer RR, which was contained in the original Ple265 promoter. Ple341 (Figure 7) produced comparable expression to Ple265 using less DNA (784 bp compared to 986 bp). It is still undetermined if the smaller promoter RR used in Ple341 is sufficient to reproduce expression in bipolar cells, or the additional enhancer element is required.

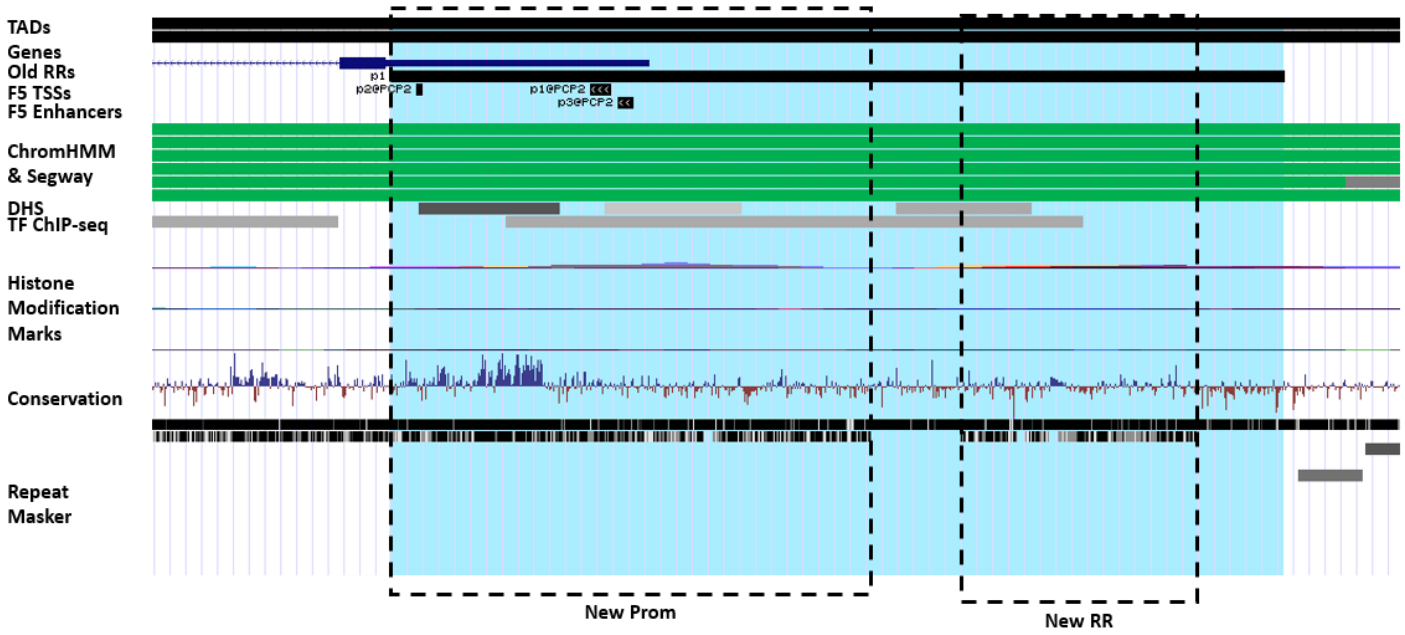



**Figure 6: Manual Bioinformatics Design of Cutting Down the Original Promoter of Ple265 to Form the Ple341 MiniPromoter.** The blue highlight indicates the original promoter sequence of Ple265, from which Ple341 was based. The new promoter region included the main FANTOM5-identified TSS until the loss of conservation between the human and mouse DNA sequences. The new RR was based on the remaining conserved sequence from the original Ple265 design.

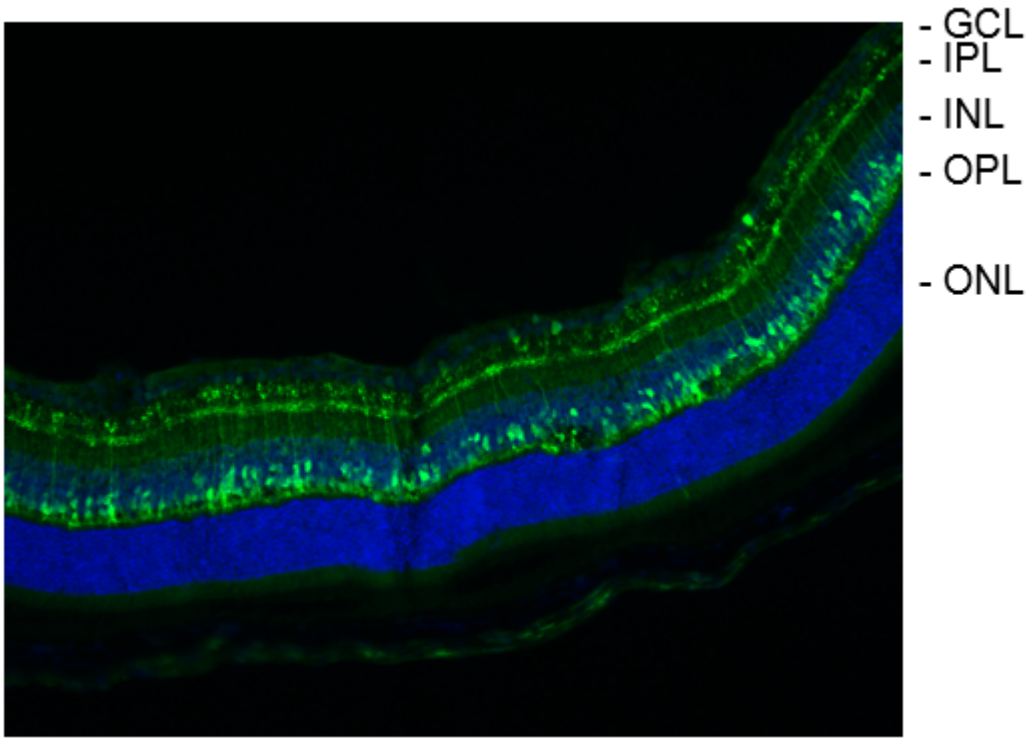



**Figure 7: MiniPromoter Ple341 from the PCP2 Gene Drives Gene Expression in Retinal Bipolar Cells.** Ple341 (PCP2 - 784 bp): The construct contains the Ple341 promoter driving the EmGFP reporter gene, injected into P4 mice and harvested after 28 days. Cell nuclei are visible in blue with Hoechst33342. EmGFP antibodies are visible in green. GCL –ganglion cell layer, IPL –inner plexiform layer, INL –inner nuclear layer, OPL. –outer plexiform layer, ONL –outer nuclear layer. *Image by Andrea Korecki*.

Ple342, Ple343, and Ple344 were based off the TUBB3 gene and were partial re-designs of an old MiniPromoter (Ple321) designed to target retinal ganglion cells. Ple342 and Ple343 contained a newly identified enhancer RR, chosen for its likeliness to be a ubiquitous enhancer to increase the expression of the original MiniPromoter. Ple342 contained only this new enhancer RR and the original promoter RR from Ple321. Ple343 contained the new RR with the old enhancer RR of Ple321 and the original promoter RR. Ple344 contained a compact version (310 bp)

of the original promoter (2,669 bp). Additionally, an enhancer-type RR was identified in the original promoter, and was subsequently identified as a new RR. This new enhancer RR was 491 bp, resulting in a compact MiniPromoter of 801 bp (see Figure 8). All three newly designed MiniPromoters produced expression in retinal ganglion cells and basal ganglia in the brain. Both Ple342 and Ple343 produced off-target expression in some amacrine cells, however, which was not observed in the original Ple321 analysis, suggesting that amacrine expression came from the addition of the ubiquitous enhancer RR. Ple344 (Figure 9) produced expression comparable to Ple321 using over 1 kb less space.

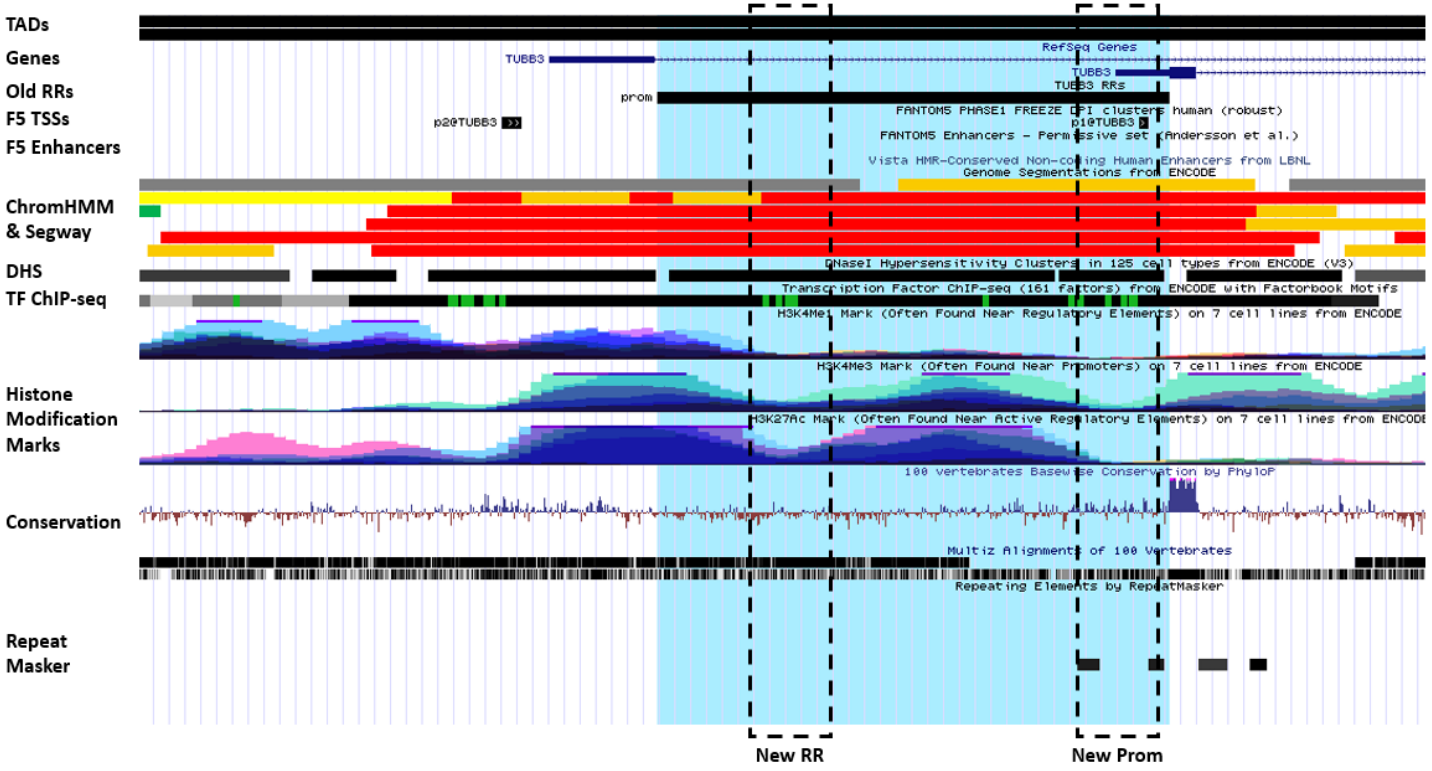



**Figure 8: Manual Bioinformatics Design of Cutting Down the Original Promoter of Ple321 to Form the Ple344 MiniPromoter.** The blue highlight indicates the original promoter sequence of Ple321, from which Ple344 was based. A new RR was chosen based upon both TF ChIP-seq data and histone mark data. The new promoter region included the main FANTOM5-identified TSS along with a large amount of overlapped TF ChIP-seq data.

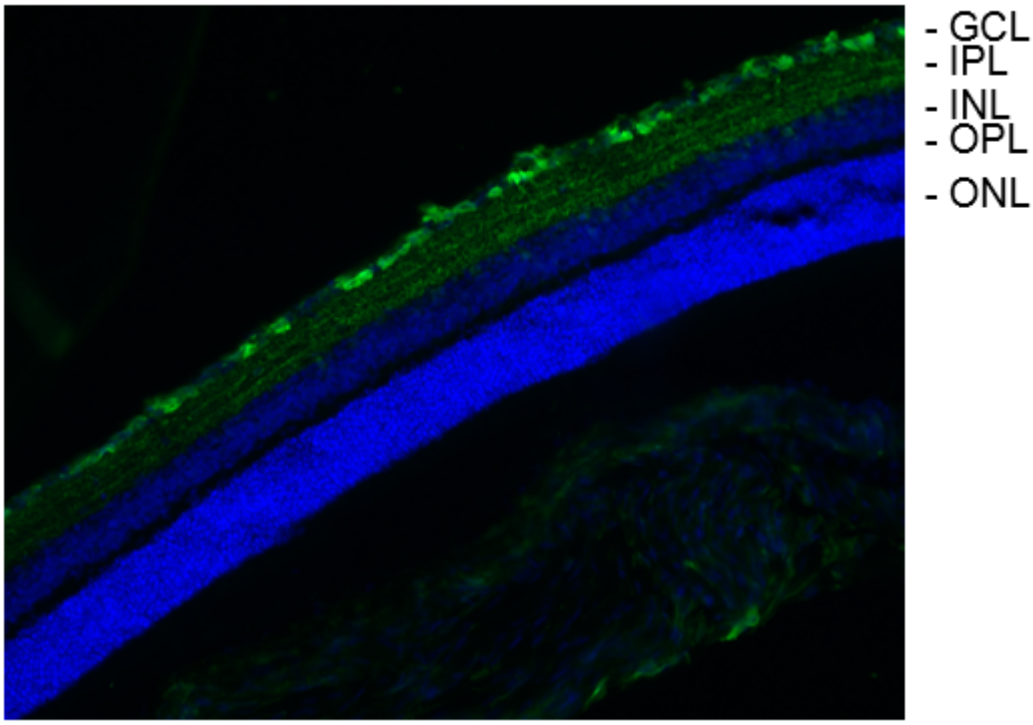



**Figure 9: MiniPromoter Ple344 from the TUBB3 Gene Drives Gene Expression in Retinal Ganglion Cells.** Ple344 (TUBB3 - 801 bp): The construct contains the Ple344 promoter driving the EmGFP reporter gene, injected into P4 mice and harvested after 28 days. Cell nuclei are visible in blue with Hoechst33342. EmGFP antibodies are visible in green. GCL –ganglion cell layer, IPL –inner plexiform layer, INL –inner nuclear layer, OPL –outer plexiform layer, ONL –outer nuclear layer. *Image by Andrea Korecki*.

Ple345 and Ple346 were based off the genes NEFL and NEFM respectively, and were designed to target retinal ganglion cells (see Figure 10). Both genes are located adjacently in both human and mouse genomes. Four common enhancer RRs were used in both MiniPromoters along with a separate promoter RR for each gene. Both MiniPromoters showed high expression levels of the reporter gene in the retinal ganglion cells, as well as very high expression in the basal ganglia in the brain. No significant off-target expression was observed. Ple345 (Figure 11)

and Ple346 showed a much higher level of reporter expression than that seen in Ple344, at the expense of being a much larger MiniPromoter (2,693 bp and 2,711 bp compared to 801 bp).

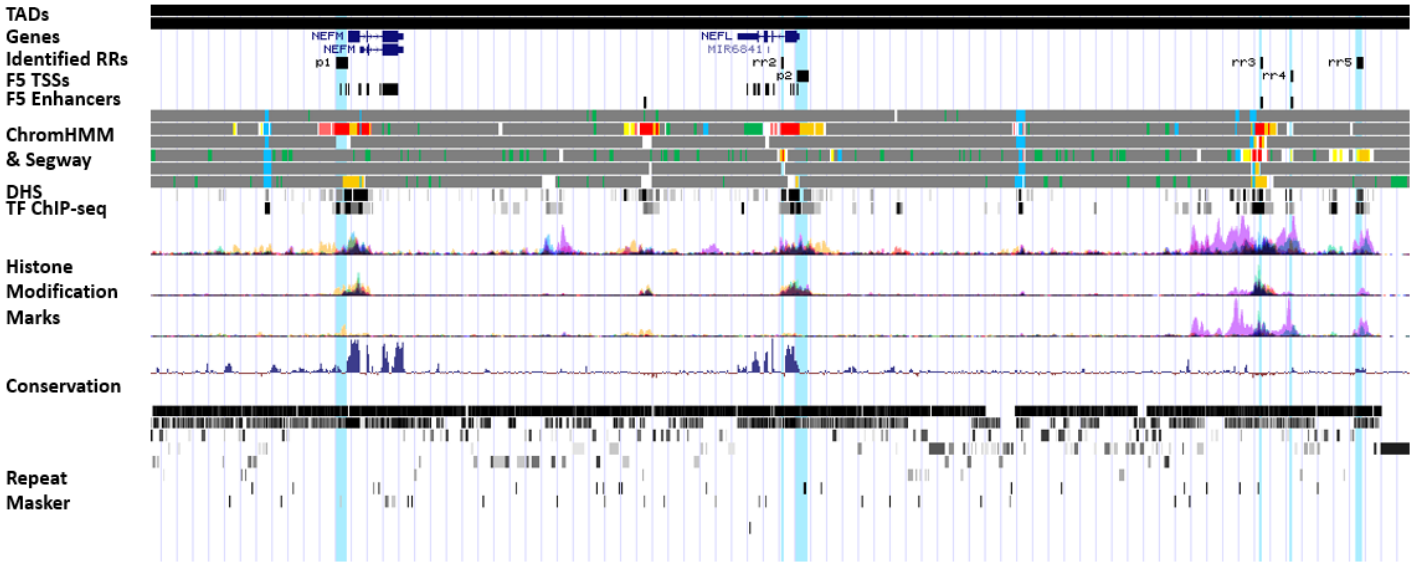



**Figure 10: Manual Bioinformatics Design of Ple345 and Ple346 Novel MiniPromoters from the NEFM and NEFL Genes.** The blue highlights indicate the RRs selected for use in both Ple345 and Ple346. Each MiniPromoter uses the same RRs, but different promoters. Promoter regions were based off FANTOM5-identified TSSs. Additional RRs were chosen based on the overlap of DHS, TF ChIP-seq, histone mark, and conservation data. Additionally, two RRs overlap FANTOM5-identified enhancers.

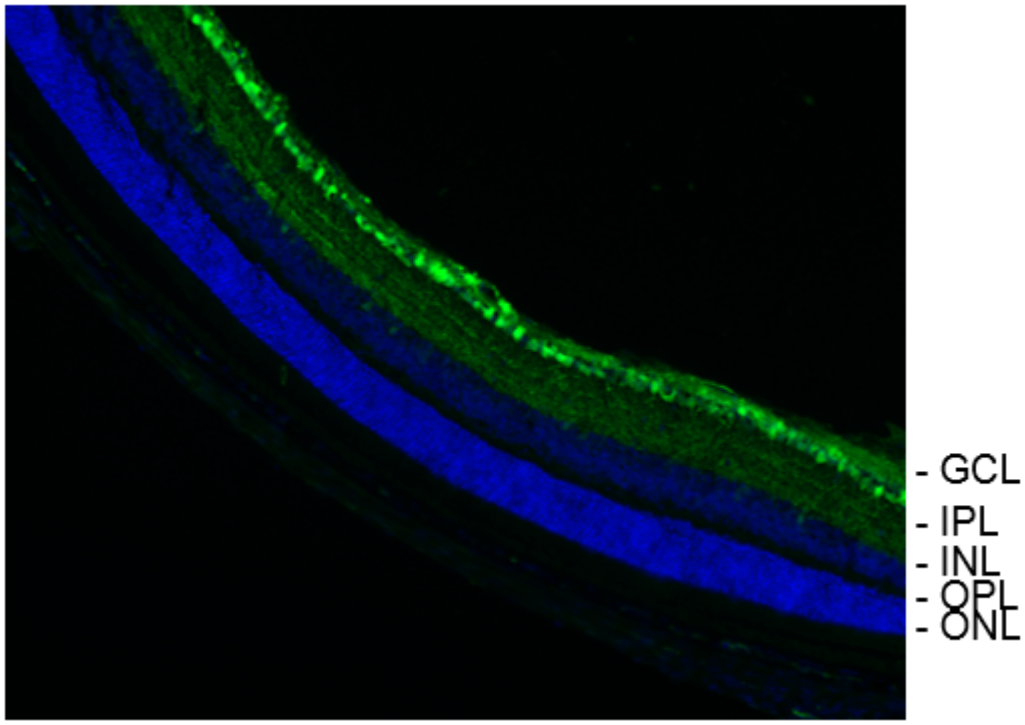



**Figure 11: MiniPromoter Ple345 from the NEFM Gene Drives Gene Expression in Retinal Ganglion Cells.** Ple345 (NEFM - 2,711 bp): The construct contains the Ple345 promoter driving the EmGFP reporter gene, injected into P4 mice and harvested after 28 days. Cell nuclei are visible in blue with Hoechst33342. EmGFP antibodies are visible in green. GCL –ganglion cell layer, IPL –inner plexiform layer, INL –inner nuclear layer, OPL –outer plexiform layer, ONL –outer nuclear layer. *Image by Andrea Korecki*.

Ple347, based off the gene GNGT2, and Ple348 and Ple349, both based off PDE6H, were designed to target cone photoreceptors. Ple347 (Figure 12) contained one enhancer RR and one promoter RR. While Ple348 and Ple349 both contained the same two enhancer RRs, they differed in the length of the promoter (Ple348 contained 1088 bp compared to 418 bp for Ple349) and 650bp of the deleted sequence was included as an additional enhancer RR. Ple347 expressed highly in cones. Surprisingly, more reporter activity was observed in cones trans-

duced with Ple347 than with the ubiquitous smCBA promoter. Expression was also detected in cone bipolar cells (see Figure 13). While not originally the target, cone photoreceptors and cone bipolar cells share similar properties, and therefore the observation is not unexpected. Ple348 and Ple349 did not show any cone photoreceptor expression, although off-target cone bipolar and amacrine cells seemed to be transduced. It should be noted, however, that at the time of writing, Ple347, Ple348 and Ple349 were only analysed in P4 mice. Cone transduction would be the strongest in earlier stages (P0)[17], and therefore we cannot determine the true strength of Ple347 or if Ple348 and Ple349 are truly negative.

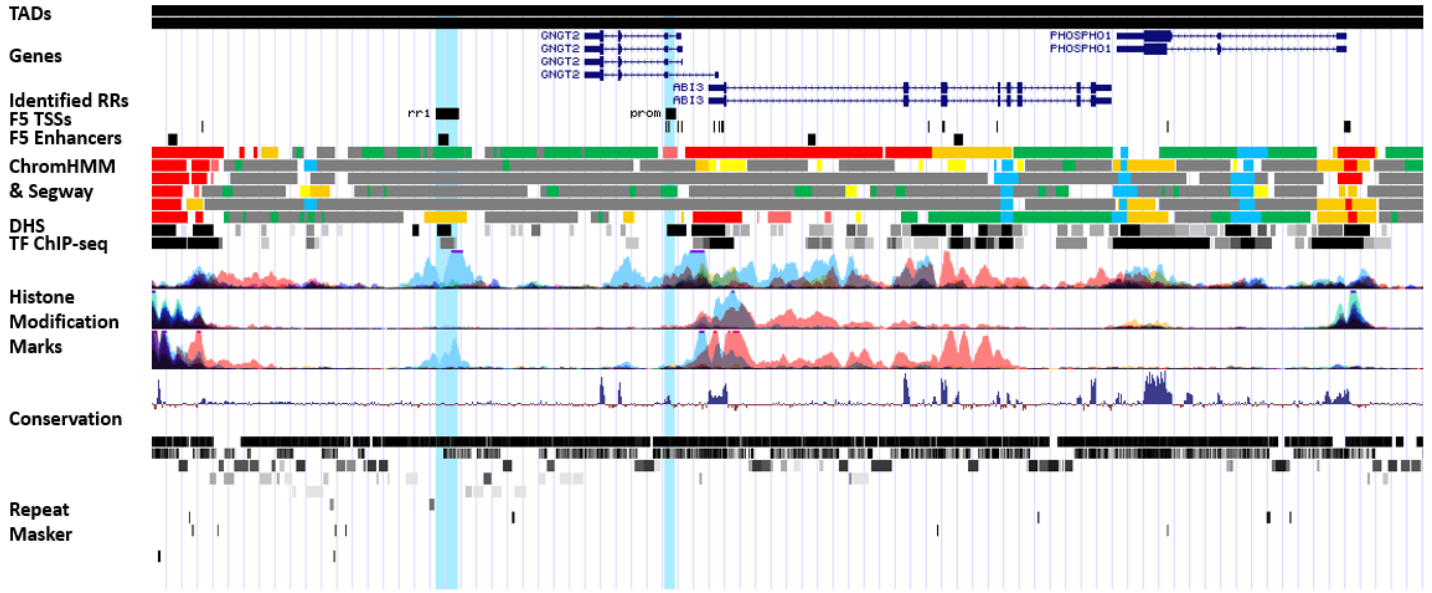



**Figure 12: Manual Bioinformatics Design of the Ple347 Novel MiniPromoter based off the GNGT2 Gene.** The blue highlights indicate the RRs selected for use in both Ple347. The GNGT2 promoter encompasses two of the six FANTOM5-identified TSSs. Other TSSs were not included due to their off-target potential, and overlap of TSS of the nearby ABI13 gene. The additional RR was chosen based on proximity to a FANTOM5-identified enhancer and a CRX TF binding site, a known photoreceptor-specific TF.

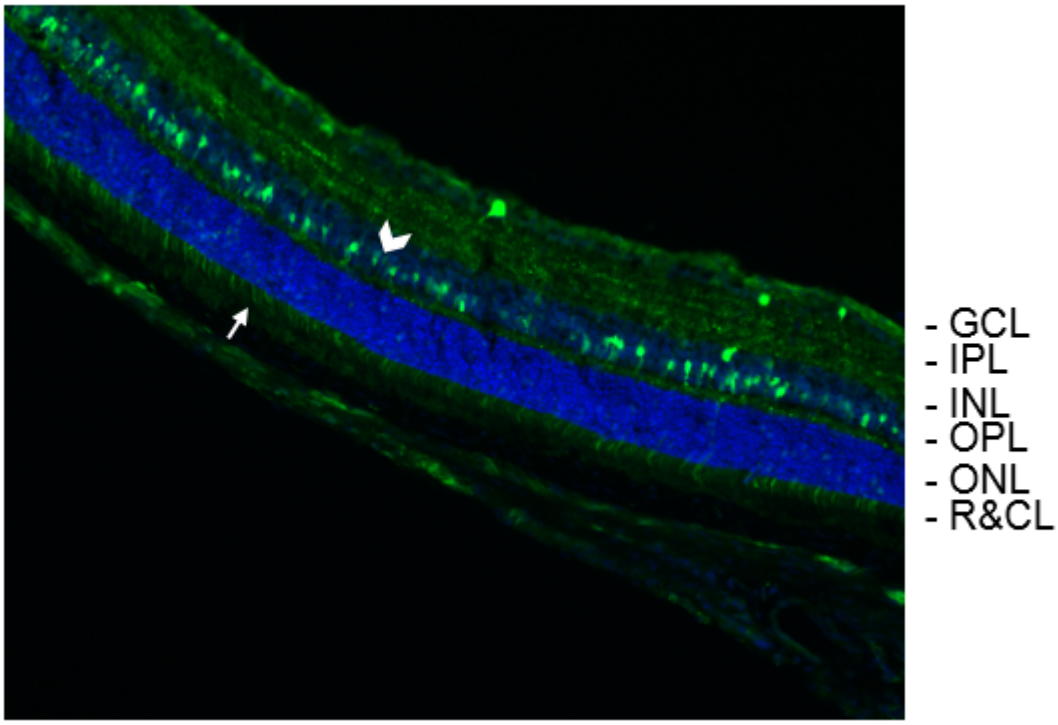



**Figure 13: MiniPromoter Ple347 from the GNGT2 gene drives gene expression in cone cells**. Ple347 (GNGT2 - 1,197 bp): The construct contains the Ple347 promoter driving the EmGFP reporter gene, injected into P4 mice and harvested after 28 days. Cell nuclei are visible in blue with Hoechst33342. EmGFP antibodies are visible in green. Expression observed in cone photoreceptors (white arrow) and in cone bipolar ON cells (white chevron). GCL –ganglion cell layer, IPL –inner plexiform layer, INL –inner nuclear layer, OPL –outer plexiform layer, ONL –outer nuclear layer, R&CL –rod and cone photoreceptor cell layer. *Image by Andrea Korecki*.

## 3.3 Automated System Creation of OnTarget

Based on the bespoke designs a semi-automated procedure was implemented for compact promoter design. Distinct modules were created for the selection of promoter and enhancer regions. Minimal promoters were identified for every FAN-

TOM5 TSS, resulting in 201,802 sequences. Enhancer RRs were calculated 'on-the-fly', as changes to default settings give rise to a different number and set of sequences. Three specific examples are described below.

We have based our initial predictions from the following features, each with a corresponding weight between 0 and 1: FANTOM5 enhancers (1), TF-ChIP peaks (0.75), chromatin accessibility (0.5), H3K27Ac (0.25), and human-mouse conservation (0.5). These weights were motivated by the feature priority used qualitatively when creating bespoke MiniPromoters. Importantly, the weighting of each feature is the driving force of our Enhancer Identification step. A higher weight represents a stronger importance placed on a feature. While these weights can be changed by the user, we empirically selected a default set which we have used for our assessment of OnTarget. As described below, using these weights, we were able to reproduce most of our chosen RRs in constructs that produced positive results.

As described in the Methods section, after feature weighting, each nucleotide in the search space is given a score. Contiguous highly-scoring regions are then returned as being potentially involved in regulation. In order to restrict the number of potential RRs, a threshold score is calculated based on the overall distribution of scores observed within each search space. The cumulative distribution charts for 2 example spaces are shown in Figure 14. Based on our analyses, and constrained by the limited number of known regulatory regions, we could not determine a universal threshold for RR identification, instead opting to only include contiguous regions scoring above the $99^{th}$ percentile of nucleotide scores. Using this threshold, we were able to reproduce 12 out of 20 regions identified by MiniPromoter designers across six different TADs.

## 3.4 Assessing the Performance of OnTarget on Experimental Data

To test our proof-of-concept, RR identification was compared by using all available data, versus a liver and hepatocyte subset. As described in the Methods section, two TADs (an $\sim$800 kb region on chromosome 7 containing at least one liver-specific gene (ABCB4) (Table 3) and an $\sim$835 kb region on chromosome 12 containing

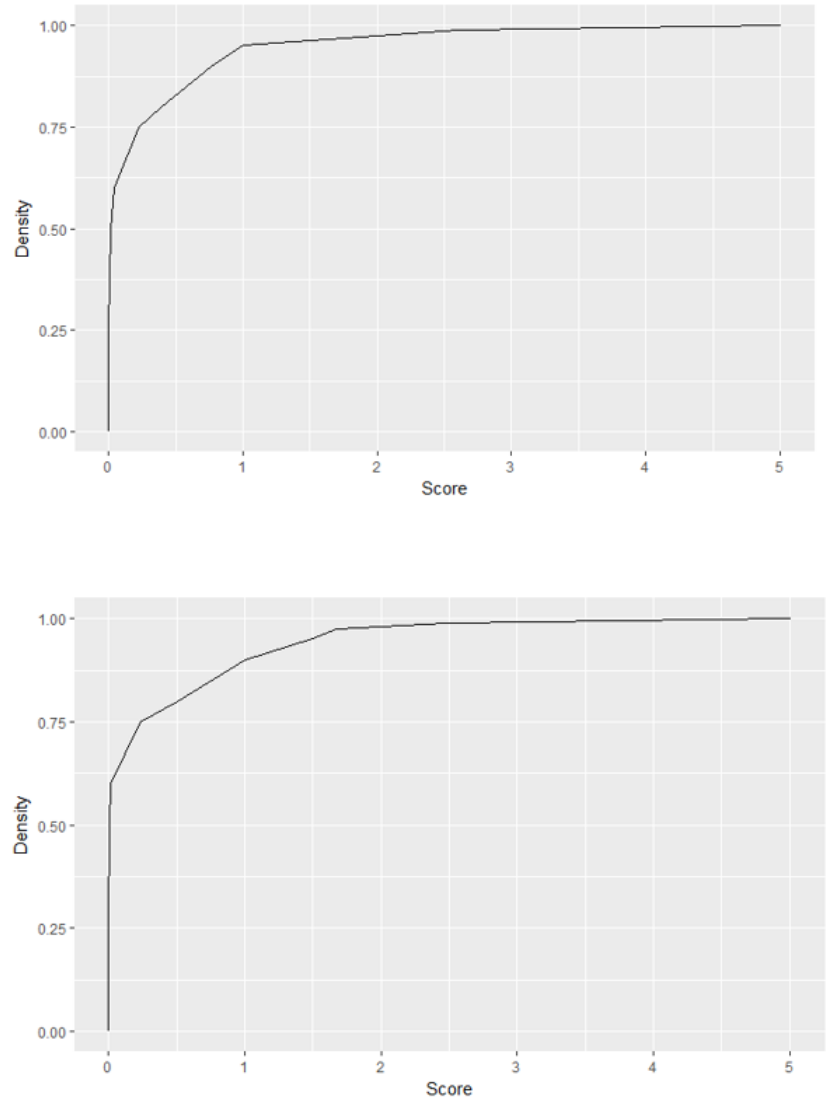



**Figure 14: The cumulative distribution charts of individual nucleotide scores from two TADs.** All nucleotides within TADs containing the gene ABCB4 (top) and NEFM (bottom) have different discrete scores, although the overall distribution pattern remains similar.

mostly brain-specific (NOS1) or housekeeping genes, Table 4) were analyzed.

In the more liver-specific TAD, 31 RRs were reported using the combination of

all datasets as features and the heuristic algorithm from the methods. In contrast, using only liver and hepatocyte samples as features resulted in the prediction of 64 RRs. The overlap between the sets was only 5. While the all-feature RRs were spread across the TAD, 64% of liver-specific feature enhancers covered 29% of the TAD. The covered portion of the TAD included the genes ABCB4, ABCB1, and the promoter region of RUNDC3B, all which are expressed in the liver, according to RNA sequencing data from the Human Protein Atlas, GTEx, and FANTOM5.

**Table 3:** Summary of OnTarget regulatory region predictions for the TAD containing the gene ABCB4. Few identified liver-specific RRs overlap those predicted using all datasets. Furthermore, liver-specific RRs tend to cluster together, rather than those from all datasets, which are spread randomly throughout the TAD. In general, the fewer number of datasets used, the less strict OnTarget becomes with RR boundaries, as seen by the difference in RR size between datasets.

| **ABCB4 TAD** | **Liver datasets** | **All datasets** |
|---|---|---|
| **Median RR size (bp)** | 232 | 73 |
| **Regions identified by OnTarget** | 64 | 31 |
| **Overlapping regions** | 8% | 16% |
| **RR localization** | clustered | sparse |

In the second TAD, 54 RRs were reported when using all datasets as features, while 75 RRs were identified using liver-specific datasets, with an overlap of 25 RRs. Unlike patterns seen in the TAD on chromosome 8, both ubiquitous and liver-specific RRs were spread throughout the regions.

## 3.5 Comparing the Designs Between Bespoke and Semi-automated Approaches

To assess the reliability of the semi-automated approach to reproduce the designs generated by hand, I compared the resulting designs (see (Table 6)). For the analysis of our successful design, I chose Ple345 and Ple346 due to their success in targeting retinal ganglion cells, with no apparent off-target expression. Furthermore, these genes used five different RRs (six in total, as both MiniPromoters use the same four enhancer RRs but different promoter RR) which allowed for more

**Table 4:** Summary of OnTarget regulatory region predictions for the TAD containing the gene NOS1. There is greater overlap between liver-specific and all datasets within this TAD. All identified RRs were scattered throughout the TAD, not clustering around any particular gene, unlike the RRs proposed in the ABCB4 TAD. Similarly to the previous TAD, lower dataset counts results in less stringent RR boundaries in a liver-specific context, while remaining almost constant using all datasets.

| **NOS1 TAD** | **Liver datasets** | **All datasets** |
|---|---|---|
| **Median RR size (bp)** | 300 | 71 |
| **Regions identified by OnTarget** | 75 | 54 |
| **Overlapping regions** | 33% | 46% |
| **RR localization** | sparse | sparse |

regions to be compared.

Using all data available on UCSC at the time of bespoke design, we identified 5 enhancer RRs and 2 promoter RRs in a 224 kb region surrounding the NEFM and NEFL genes. A subset of 4 out of 5 enhancer RRs were included in the finalized designs.

OnTarget analyzed across the full ∼640 kb TAD, which surrounded both genes (as well as two long non-coding RNA transcripts, one microRNA, and the partial 3' end of another protein-coding gene). Knowing that the expected expression should be seen in both retinal ganglion cells and basal ganglia in the brain, I analyzed feature data from the following primary cell and tissue datasets: neuronal stem cells, neuronal progenitor cells, neuron, brain, midbrain, eye, and retina. Within the original 640 kb TAD, 15 RRs were identified. A subset of 10 RRs (66% of all identified RRs) were located within the 224 kb region originally used for manual RR identification. Analyzing the same 640 kb TAD using all datasets, mimicking our bespoke MiniPromoter design, 17 RRs were identified. A subset of 9 RRs were found within the 224 kb search space. Out of the 7 manually identified RRs, 4 (from the brain and eye datasets, or 3 from all datasets) overlapped with those found by OnTarget using the selected features. Relaxing the scoring threshold to the $98^{th}$ percentile, 5 RRs from the brain and eye datasets were recovered. Reassuringly, the other two RRs could be identified at the $95^{th}$ percentile score cutt-off (Table 5).

For an example of an unsuccessful design, I chose a subset of designs based

**Table 5:** Regulatory Regions Identified for NEFM/NEFL Bespoke MiniPromoters in Comparison to Regulatory Predictions Predictions from OnTarget.

| | **MiniPromoter Design (Manual, All Available Datasets)** | **OnTarget (All Available Datasets)** | **OnTarget (Brain/Eye Datasets)** |
|---|---|---|---|
| **Regulatory region search space (in kb)** | 224 | 640 | 640 |
| **Number of identified regions** | 7 | 17 (9 in 224kb) | 15 (10 in 224kb) |

on the gene DDC, constructs Ple56, Ple57, Ple58, and Ple59. These were early constructs designed for the paper by Portatles-Casamar *et al.* in 2010[60], before the implementation of our current bespoke design process. No expression was seen in the brain for any of these MiniPromoters.

Out of four designs, three enhancer RRs were used in combination with one of two similar promoter RRs. The search space for RRs was limited to the flanking genes of DDC, spanning about 20 kb upstream and 5 kb downstream. OnTarget analyzed across the TAD, a region of ∼720 kb, using the same datasets described above. Out of the 18 brain and eye-specific RRs predicted by OnTarget, only one region partially overlapped all tested MiniPromoter regions. Unsurprisingly, this partial overlap was across the promoter RR of the MiniPromoter, although the predicted RR is much more compact than the tested regions, potentially indicating the original promoter RRs contained elements unimportant to, or unfavorable for expression in the brain. Interestingly, there was a large overlap of predicted RRs in the brain and eye specific datasets compared with all data. Out of the 15 RRs selected while using all datasets, only 2 did not overlap RRs from the tissue-specific set.

Based on the above results, I expect that regions predicted by OnTarget in cell-specific contexts should raise the likeliness of MiniPromoters producing expression of the reporter gene. It is therefore of interest to predict which bespoke designs still awaiting testing would be more likely to be successful. I chose Ple368, one of the newest MiniPromoter constructs designed based on the GAD2 gene with the previously described datasets. While OnTarget used a ∼560 kb TAD as its search

space, our MiniPromoter analysis was done using a ∼260 kb region which was based off a successful mouse BAC targeting *Gad2*[9]. Out of 13 RRs predicted by OnTarget in the cell-specific datasets, 9 were found in the search space used for the bespoke design of Ple368. Inversely, only 1 RR was predicted within that smaller search space, while the other 8 OnTarget-predicted RRs fell within the larger TAD. Both RRs selected by hand for Ple368 completely overlapped OnTarget-predicted RRs using the brain and eye-specific datasets. It is therefore our prediction that this MiniPromoter will express our reporter gene in at least a subset of regions of the brain and eye.

**Table 6:** Summary of OnTarget Regulatory Region Predictions for Bespoke MiniPromoters.

| | **Ple345 & Ple346** | **Ple56 to Ple59** | **Ple368** |
|---|---|---|---|
| **MiniPromoter result** | Success | Fail | Untested |
| **MiniPromoter search space (kb)** | 224 | 125 | 260 |
| **OnTarget search space (kb)** | 640 | 720 | 560 |
| **Number of original RRs identified** | 7 | 5 | 2 |
| **OnTarget brain/eye RR overlap with MiniPromoters** | 4 | 2 | 2 |
| **OnTarget all datasets RR overlap with MiniPromoters** | 3 | 0 | 1 |

# Chapter 4

# Conclusion & Future Work

In the research activities of this thesis, we were able to create 49 MiniPromoters designed to drive expression of a reporter gene in cell-specific contexts, of which a subset of 20 constructs were tested *in vivo* to validate the design process. Based on an initial set of bespoke designs, we were able to recreate the design logic within a semi-automated pipeline. As expected, analyses using cell type-specific datasets results in different designs than those incorporating all possible datasets. We have performed initial tests of the semi-automatic approach, supporting the use of cell-type specific information. The bioinformatics approaches within the thesis are of importance in the field of gene therapy, as the use of small, specific promoters not only increases the therapeutic capacity, but also restricts the delivery of the therapeutic to relevant cells.

We recognize, however, the need for future developments of our tool. First, we must continue to test OnTarget as our validation set of tested MiniPromoters grows. In this sense, we must ensure that our current weighting scheme continues to preferentially detect RRs from successful MiniPromoter experiments. Furthermore, OnTarget should identify *cis*-regulatory sequences that were not selected by human designers in the unsuccessful experiments. We also plan to test OnTarget for prediction of enhancers described in literature.

Next, we will continue to collect cell-specific datasets as they become available. Unsurprisingly, most experiments are performed on immortalized cell lines. Although these are good starting points, they may not accurately recapitulate ex-

pression in natural conditions. There is a dearth of tissue and cell-specific TF experiments.

Finally, we hope to implement an important additional feature into future releases of the OnTarget tool. Specifically, we hope to include the ability to modify each cis-regulatory sequence in order to modulate the strength of TF binding sites. Previous research, as well as our own observations from Ple331, has shown that small base pair changes in TF binding sites can dramatically affect expression levels. OnTarget will scan identified RRs and provide suggestions of minimal alteration of the endogenous sequence to create a high-affinity binding site for a desired TF. This system is based off of previous lab expertise in the creation of the MANTA database[51] (DNA alterations impacting TF binding), and the upkeep of JASPAR[52] (TF binding profile database). Importantly, this feature must also ensure that sequence modification does not destroy other important binding sites, or create undesired sites which may lead to unexpected and undesired off-target expression after MiniPromoter delivery.

I have shown preliminary work for successfully identifying cis-regulatory sequences and creating functional MiniPromoters for therapeutic delivery in small viral vectors. This approach lead to the creation of our semi-automated tool, OnTarget, to perform the same task. While there is still work to be done, implementing OnTarget as outlined in this thesis should ultimately lead to better and faster identification of *cis*-regulatory sequences and designing of MiniPromoters.

# Bibliography


[1] Viral Plasmids and Resources. → pages 3, 5, 6

[2] N. al Yacoub, M. Romanowska, N. Haritonova, and J. Foerster. Optimized production and concentration of lentiviral vectors containing large inserts. *The Journal of Gene Medicine*, 9(7):579–584, jul 2007. ISSN 1099498X. doi:10.1002/jgm.1052. URL http://doi.wiley.com/10.1002/jgm.1052. → pages 5

[3] R. Andersson. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays*, 37(3):314–323, mar 2015. ISSN 02659247. doi:10.1002/bies.201400162. URL http://doi.wiley.com/10.1002/bies.201400162. → pages 9, 10, 12

[4] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, T. F. Consortium, A. R. R. Forrest, P. Carninci, M. Rehli, and A. Sandelin. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, mar 2014. ISSN 0028-0836. doi:10.1038/nature12787. URL http://www.nature.com/doifinder/10.1038/nature12787. → pages 11, 19

[5] A. Asokan, D. V. Schaffer, and R. Jude Samulski. The AAV Vector Toolkit: Poised at the Clinical Crossroads. *Molecular Therapy*, 20(4):699–708, apr 2012. ISSN 15250016. doi:10.1038/mt.2011.287. URL http://linkinghub.elsevier.com/retrieve/pii/S1525001616305342. → pages 6

[6] A. A. Avilion. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes & Development*, 17(1):126–140, jan 2003. ISSN 08909369. doi:10.1101/gad.224503. URL http://www.genesdev.org/cgi/doi/10.1101/gad.224503. → pages 12

[7] A. J. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, mar 2011. ISSN 1001-0602. doi:10.1038/cr.2011.22. URL http://www.nature.com/doifinder/10.1038/cr.2011.22. → pages 13

[8] K. Benihoud. Adenovirus vectors for gene delivery. *Current Opinion in Biotechnology*, 10(5):440–447, oct 1999. ISSN 09581669. doi:10.1016/S0958-1669(99)00007-5. URL http://linkinghub.elsevier.com/retrieve/pii/S0958166999000075. → pages 5

[9] S. Besser, M. Sicker, G. Marx, U. Winkler, V. Eulenburg, S. Hülsmann, and J. Hirrlinger. A transgenic mouse line expressing the red fluorescent protein tdtomato in gabaergic neurons. *PloS one*, 10(6):e0129934, 2015. → pages 57

[10] L. Biasco, A. Ambrosi, D. Pellin, C. Bartholomae, I. Brigida, M. G. Roncarolo, C. Di Serio, C. von Kalle, M. Schmidt, and A. Aiuti. Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell. *EMBO Molecular Medicine*, 3(2):89–101, feb 2011. ISSN 17574676. doi:10.1002/emmm.201000108. URL http://embomolmed.embopress.org/cgi/doi/10.1002/emmm.201000108. → pages 5

[11] R. M. Blaese, K. W. Culver, A. D. Miller, C. S. Carter, T. Fleisher, M. Clerici, G. Shearer, L. Chang, Y. Chiang, P. Tolstoshev, J. J. Greenblatt, S. A. Rosenberg, H. Klein, M. Berger, C. A. Mullen, W. J. Ramsey, L. Muul, R. A. Morgan, and W. F. Anderson. T lymphocyte-directed gene therapy for ADA- SCID: initial trial results after 4 years. *Science (New York, N.Y.)*, 270 (5235):475–80, oct 1995. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/7570001. → pages 2

[12] J. A. Blake, J. T. Eppig, J. A. Kadin, J. E. Richardson, C. L. Smith, and C. J. Bult. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research*, 45(D1): D723–D729, jan 2017. ISSN 0305-1048. doi:10.1093/nar/gkw1040. URL

https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1040. → pages 19

[13] M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708–715, 2004. → pages 21

[14] R. C. Boucher, M. R. Knowles, L. G. Johnson, J. C. Olsen, R. Pickles, J. M. Wilson, J. Engelhardt, Y. Yang, and M. Grossman. Gene Therapy for Cystic Fibrosis Using E1-Deleted Adenovirus: A Phase I Trial in the Nasal Cavity. University of North Carolina at Chapel Hill, Chapel Hill, North Carolina. *Human Gene Therapy*, 5(5):615–639, may 1994. ISSN 1043-0342. doi:10.1089/hum.1994.5.5-615. URL http://www.liebertonline.com/doi/abs/10.1089/hum.1994.5.5-615. → pages 2

[15] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2): 311–322, 2008. → pages 13

[16] J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. In *Current Protocols in Molecular Biology*, pages 21.29.1–21.29.9. John Wiley & Sons, Inc., Hoboken, NJ, USA, jan 2015. doi:10.1002/0471142727.mb2129s109. URL http://doi.wiley.com/10.1002/0471142727.mb2129s109. → pages 14

[17] L. C. Byrne, Y. J. Lin, T. Lee, D. V. Schaffer, and J. G. Flannery. The expression pattern of systemically injected AAV9 in the developing mouse retina is determined by age. *Molecular therapy : the journal of the American Society of Gene Therapy*, 23(2):290–296, 2015. ISSN 1525-0024 (Electronic). doi:10.1038/mt.2014.181. → pages 36, 49

[18] M. Cavazzana-Calvo. Gene Therapy of Human Severe Combined Immunodeficiency (SCID)-X1 Disease. *Science*, 288(5466):669–672, apr 2000. ISSN 00368075. doi:10.1126/science.288.5466.669. URL http://www.sciencemag.org/cgi/doi/10.1126/science.288.5466.669. → pages 2

[19] S. Chira, C. S. Jackson, I. Oprea, F. Ozturk, M. S. Pepper, I. Diaconu, C. Braicu, L.-Z. Raduly, G. A. Calin, and I. Berindan-Neagoe. Progresses

towards safe and efficient gene therapy vectors. *Oncotarget*, 6(31): 30675–30703, oct 2015. ISSN 1949-2553. doi:10.18632/oncotarget.5169. URL http://www.oncotarget.com/fulltext/5169. → pages 5

[20] A. V. Cideciyan, T. S. Aleman, S. L. Boye, S. B. Schwartz, S. Kaushal, A. J. Roman, J.-j. Pang, A. Sumaroka, E. A. Windsor, J. M. Wilson, et al. Human gene therapy for rpe65 isomerase deficiency activates the retinoid cycle of vision but with slow rod kinetics. *Proceedings of the National Academy of Sciences*, 105(39):15112–15117, 2008. → pages 16

[21] L. J. Core, J. J. Waterfall, and J. T. Lis. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, 322(5909):1845–1848, dec 2008. ISSN 0036-8075. doi:10.1126/science.1162228. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1162228. → pages 10

[22] C. N. de Leeuw, F. M. Dyka, S. L. Boye, S. Laprise, M. Zhou, A. Y. Chou, L. Borretta, S. C. McInerny, K. G. Banks, E. Portales-Casamar, M. I. Swanson, C. A. D'Souza, S. E. Boye, S. J. Jones, R. A. Holt, D. Goldowitz, W. W. Hauswirth, W. W. Wasserman, and E. M. Simpson. Targeted CNS delivery using human MiniPromoters and demonstrated compatibility with adeno-associated viral vectors. *Molecular Therapy - Methods & Clinical Development*, 1:5, 2014. ISSN 23290501. doi:10.1038/mtm.2013.5. URL http://linkinghub.elsevier.com/retrieve/pii/S2329050116300705. → pages 2, 17

[23] C. N. de Leeuw, A. J. Korecki, G. E. Berry, J. W. Hickmott, S. L. Lam, T. C. Lengyell, R. J. Bonaguro, L. J. Borretta, V. Chopra, A. Y. Chou, C. A. D'Souza, O. Kaspieva, S. Laprise, S. C. McInerny, E. Portales-Casamar, M. I. Swanson-Newman, K. Wong, G. S. Yang, M. Zhou, S. J. M. Jones, R. A. Holt, A. Asokan, D. Goldowitz, W. W. Wasserman, and E. M. Simpson. rAAV-compatible MiniPromoters for restricted expression in the brain and eye. *Molecular Brain*, 9(1):52, 2016. ISSN 1756-6606. doi:10.1186/s13041-016-0232-4. URL http://www.ncbi.nlm.nih.gov/pubmed/27164903{%}5Cnhttp://molecularbrain.biomedcentral.com/articles/10.1186/s13041-016-0232-4. → pages 2, 17, 20

[24] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, apr 2012. ISSN

0028-0836. doi:10.1038/nature11082. URL http://www.nature.com/doifinder/10.1038/nature11082. → pages 14, 15

[25] M. Edelstein. Gene Therapy Clinical Trials Worldwide, 2017. URL http://www.abedia.com/wiley/index.html. → pages 1, 3, 5

[26] A. El-Aneed. An overview of current delivery systems in cancer gene therapy. *Journal of Controlled Release*, 94(1):1–14, jan 2004. ISSN 01683659. doi:10.1016/j.jconrel.2003.09.013. URL http://linkinghub.elsevier.com/retrieve/pii/S0168365903004462. → pages 5

[27] G. Elgar and T. Vavouri. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in genetics*, 24(7):344–352, 2008. → pages 2

[28] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696):636–40, oct 2004. ISSN 1095-9203. doi:10.1126/science.1105136. URL http://www.ncbi.nlm.nih.gov/pubmed/15499007. → pages 12, 13, 19

[29] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, sep 2012. ISSN 1476-4687. doi:10.1038/nature11247. URL http://www.ncbi.nlm.nih.gov/pubmed/22955616http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3439153. → pages 12, 13, 19

[30] J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, feb 2012. ISSN 1548-7091. doi:10.1038/nmeth.1906. URL http://www.nature.com/doifinder/10.1038/nmeth.1906. → pages 15, 19

[31] FANTOM Consortium and the RIKEN PMI and CLST (DGT), A. R. R. Forrest, H. Kawaji, M. Rehli, J. K. Baillie, M. J. L. de Hoon, V. Haberle, T. Lassmann, I. V. Kulakovskiy, M. Lizio, M. Itoh, R. Andersson, C. J. Mungall, T. F. Meehan, S. Schmeier, N. Bertin, M. Jørgensen, E. Dimont, E. Arner, C. Schmidl, U. Schaefer, Y. A. Medvedeva, C. Plessy, M. Vitezic, J. Severin, C. A. Semple, Y. Ishizu, R. S. Young, M. Francescatto, I. Alam, D. Albanese, G. M. Altschuler, T. Arakawa, J. A. C. Archer, P. Arner, M. Babina, S. Rennie, P. J. Balwierz, A. G. Beckhouse, S. Pradhan-Bhatt, J. A. Blake, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A. M. Burroughs, A. Califano, C. V. Cannistraci, D. Carbajo, Y. Chen,

M. Chierici, Y. Ciani, H. C. Clevers, E. Dalla, C. A. Davis, M. Detmar, A. D. Diehl, T. Dohi, F. Drabløs, A. S. B. Edge, M. Edinger, K. Ekwall, M. Endoh, H. Enomoto, M. Fagiolini, L. Fairbairn, H. Fang, M. C. Farach-Carson, G. J. Faulkner, A. V. Favorov, M. E. Fisher, M. C. Frith, R. Fujita, S. Fukuda, C. Furlanello, M. Furino, J.-i. Furusawa, T. B. Geijtenbeek, A. P. Gibson, T. Gingeras, D. Goldowitz, J. Gough, S. Guhl, R. Guler, S. Gustincich, T. J. Ha, M. Hamaguchi, M. Hara, M. Harbers, J. Harshbarger, A. Hasegawa, Y. Hasegawa, T. Hashimoto, M. Herlyn, K. J. Hitchens, S. J. Ho Sui, O. M. Hofmann, I. Hoof, F. Hori, L. Huminiecki, K. Iida, T. Ikawa, B. R. Jankovic, H. Jia, A. Joshi, G. Jurman, B. Kaczkowski, C. Kai, K. Kaida, A. Kaiho, K. Kajiyama, M. Kanamori-Katayama, A. S. Kasianov, T. Kasukawa, S. Katayama, S. Kato, S. Kawaguchi, H. Kawamoto, Y. I. Kawamura, T. Kawashima, J. S. Kempfle, T. J. Kenna, J. Kere, L. M. Khachigian, T. Kitamura, S. P. Klinken, A. J. Knox, M. Kojima, S. Kojima, N. Kondo, H. Koseki, S. Koyasu, S. Krampitz, A. Kubosaki, A. T. Kwon, J. F. J. Laros, W. Lee, A. Lennartsson, K. Li, B. Lilje, L. Lipovich, A. Mackay-Sim, R.-i. Manabe, J. C. Mar, B. Marchand, A. Mathelier, N. Mejhert, A. Meynert, Y. Mizuno, D. A. de Lima Morais, H. Morikawa, M. Morimoto, K. Moro, E. Motakis, H. Motohashi, C. L. Mummery, M. Murata, S. Nagao-Sato, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, K. Nakazato, E. van Nimwegen, N. Ninomiya, H. Nishiyori, S. Noma, S. Noma, T. Noazaki, S. Ogishima, N. Ohkura, H. Ohimiya, H. Ohno, M. Ohshima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. A. Ovchinnikov, A. Pain, R. Passier, M. Patrikakis, H. Persson, S. Piazza, J. G. D. Prendergast, O. J. L. Rackham, J. A. Ramilowski, M. Rashid, T. Ravasi, P. Rizzu, M. Roncador, S. Roy, M. B. Rye, E. Saijyo, A. Sajantila, A. Saka, S. Sakaguchi, M. Sakai, H. Sato, S. Savvi, A. Saxena, C. Schneider, E. A. Schultes, G. G. Schulze-Tanzil, A. Schwegmann, T. Sengstag, G. Sheng, H. Shimoji, Y. Shimoni, J. W. Shin, C. Simon, D. Sugiyama, T. Sugiyama, M. Suzuki, N. Suzuki, R. K. Swoboda, P. A. C. 't Hoen, M. Tagami, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, Z. Tatum, M. Thompson, H. Toyodo, T. Toyoda, E. Valen, M. van de Wetering, L. M. van den Berg, R. Verado, D. Vijayan, I. E. Vorontsov, W. W. Wasserman, S. Watanabe, C. A. Wells, L. N. Winteringham, E. Wolvetang, E. J. Wood, Y. Yamaguchi, M. Yamamoto, M. Yoneda, Y. Yonekura, S. Yoshida, S. E. Zabierowski, P. G. Zhang, X. Zhao, S. Zucchelli, K. M. Summers, H. Suzuki, C. O. Daub, J. Kawai, P. Heutink, W. Hide, T. C. Freeman, B. Lenhard, V. B. Bajic, M. S. Taylor, V. J. Makeev, A. Sandelin, D. A. Hume, P. Carninci, and Y. Hayashizaki. A promoter-level mammalian expression atlas. *Nature*, 507 (7493):462–70, mar 2014. ISSN 1476-4687. doi:10.1038/nature13182.

URL http://www.ncbi.nlm.nih.gov/pubmed/24670764http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4529748. → pages 7, 11

[32] T. R. Flotte. Gene Therapy Progress and Prospects: Recombinant adeno-associated virus (rAAV) vectors. *Gene Therapy*, 11(10):805–810, may 2004. ISSN 0969-7128. doi:10.1038/sj.gt.3302233. URL http://www.nature.com/doifinder/10.1038/sj.gt.3302233. → pages 6

[33] M. K. Foecking and H. Hofstetter. Powerful and versatile enhancer-promoter unit for mammalian expression vectors. *Gene*, 45(1):101–5, 1986. ISSN 0378-1119. URL http://www.ncbi.nlm.nih.gov/pubmed/3023199. → pages 6

[34] S. F. Gilbert. *Developmental Biology*. Sinaur Associates, Sutherland, MA, 6 edition, 2000. ISBN 0-87893-243-7. → pages 7

[35] G. Gill. Regulation of the initiation of eukaryotic transcription. *Essays In Biochemistry*, 37:33–43, may 2001. ISSN 0071-1365. doi:10.1042/bse0370033. URL http://essays.biochemistry.org/lookup/doi/10.1042/bse0370033. → pages 11, 13

[36] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, 17(6): 877–885, jun 2007. ISSN 1088-9051. doi:10.1101/gr.5533506. URL http://www.genome.org/cgi/doi/10.1101/gr.5533506. → pages 14

[37] S. Goverdhana, M. Puntel, W. Xiong, J. Zirger, C. Barcia, J. Curtin, E. Soffer, S. Mondkar, G. King, J. Hu, S. Sciascia, M. Candolfi, D. Greengold, P. Lowenstein, and M. Castro. Regulatable gene expression systems for gene therapy applications: progress and future challenges. *Molecular Therapy*, 12(2):189–211, aug 2005. ISSN 15250016. doi:10.1016/j.ymthe.2005.03.022. URL http://linkinghub.elsevier.com/retrieve/pii/S1525001605001450. → pages 1, 5, 16

[38] S. Hacein-Bey-Abina. LMO2-Associated Clonal T Cell Proliferation in Two Patients after Gene Therapy for SCID-X1. *Science*, 302(5644):415–419, oct 2003. ISSN 0036-8075. doi:10.1126/science.1088547. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1088547. → pages 3

[39] N. Heintz. Gene Expression Nervous System Atlas (GENSAT). *Nature Neuroscience*, 7(5):483–483, may 2004. ISSN 1097-6256. doi:10.1038/nn0504-483. URL http://www.nature.com/doifinder/10.1038/nn0504-483. → pages 16, 19

[40] J. W. Hickmott, C.-y. Chen, D. J. Arenillas, A. J. Korecki, S. L. Lam, L. L. Molday, R. J. Bonaguro, M. Zhou, A. Y. Chou, A. Mathelier, S. L. Boye, W. W. Hauswirth, R. S. Molday, W. W. Wasserman, and E. M. Simpson. PAX6 MiniPromoters drive restricted expression from rAAV in the adult mouse retina. *Molecular Therapy*, 3(June):16051, 2016. ISSN 2329-0501. doi:doi:10.1038/mtm.2016.51. → pages 2, 17, 20

[41] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476, mar 2012. ISSN 1548-7091. doi:10.1038/nmeth.1937. URL http://www.nature.com/doifinder/10.1038/nmeth.1937. → pages 15, 19

[42] T. Hollon. Researchers and regulators reflect on first gene therapy death. *Nature Medicine*, 6(1):6–6, jan 2000. ISSN 1078-8956. doi:10.1038/71545. URL http://www.nature.com/doifinder/10.1038/71545. → pages 2

[43] M. A. Kay, J. C. Glorioso, and L. Naldini. Viral vectors for gene therapy: the art of turning infectious agents into vehicles of therapeutics.Kay, M. A., Glorioso, J. C., & Naldini, L. (2001). Viral vectors for gene therapy: the art of turning infectious agents into vehicles of therapeutics. Natur. *Nature medicine*, 7(1):33–40, jan 2001. ISSN 1078-8956. doi:10.1038/83324. URL http://www.nature.com/doifinder/10.1038/83324http://www.ncbi.nlm.nih.gov/pubmed/11135613. → pages 3, 5, 6

[44] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, may 2002. ISSN 1088-9051. doi:10.1101/gr.229102. URL http://www.genome.org/cgi/doi/10.1101/gr.229102. → pages 16, 19

[45] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, C. J. Mungall, E. Arner, J. Baillie, N. Bertin, H. Bono, M. de Hoon, A. D. Diehl, E. Dimont, T. C. Freeman, K. Fujieda, W. Hide, R. Kaliyaperumal, T. Katayama, T. Lassmann, T. F. Meehan, K. Nishikata, H. Ono, M. Rehli, A. Sandelin, E. A. Schultes, P. t Hoen, Z. Tatum, M. Thompson, T. Toyoda,

D. W. Wright, C. O. Daub, M. Itoh, P. Carninci, Y. Hayashizaki, A. Forrest, and H. Kawaji. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology*, 16(1):22, 2015. ISSN 1465-6906. doi:10.1186/s13059-014-0560-6. URL http://genomebiology.com/2015/16/1/22. → pages 11, 19

[46] D. G. Lupiáñez, M. Spielmann, and S. Mundlos. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics*, 32 (4):225–237, apr 2016. ISSN 01689525. doi:10.1016/j.tig.2016.01.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0168952516000044. → pages 15

[47] A. Lusser and J. T. Kadonaga. Chromatin remodeling by ATP-dependent molecular machines. *BioEssays*, 25(12):1192–1200, dec 2003. ISSN 0265-9247. doi:10.1002/bies.10359. URL http://doi.wiley.com/10.1002/bies.10359. → pages 13

[48] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015. → pages 31

[49] C. A. Maguire, S. H. Ramirez, S. F. Merkel, M. Sena-Esteves, and X. O. Breakefield. Gene therapy for the nervous system: challenges and new strategies. *Neurotherapeutics*, 11(4):817–839, 2014. → pages 16

[50] G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*, 7(1):29–59, sep 2006. ISSN 1527-8204. doi:10.1146/annurev.genom.7.080505.115623. URL http://www.annualreviews.org/doi/10.1146/annurev.genom.7.080505.115623. → pages 9

[51] A. Mathelier, C. Lefebvre, A. W. Zhang, D. J. Arenillas, J. Ding, W. W. Wasserman, and S. P. Shah. Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome biology*, 16:84, apr 2015. ISSN 1474-760X. doi:10.1186/s13059-015-0648-7. URL http://www.ncbi.nlm.nih.gov/pubmed/25903198http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4467049. → pages 59

[52] A. Mathelier, O. Fornes, D. J. Arenillas, C.-y. Chen, G. Denay, J. Lee, W. Shi, C. Shyr, G. Tan, R. Worsley-Hunt, A. W. Zhang, F. Parcy,

B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–D115, jan 2016. ISSN 0305-1048. doi:10.1093/nar/gkv1176. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1176. → pages 12, 59

[53] R. Mundade, H. G. Ozer, H. Wei, L. Prabhu, and T. Lu. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle*, 13(18): 2847–2852, sep 2014. ISSN 1538-4101. doi:10.4161/15384101.2014.949201. URL http://www.tandfonline.com/doi/full/10.4161/15384101.2014.949201. → pages 12, 13

[54] L. Naldini. Gene therapy returns to centre stage. *Nature*, 526(7573):351–60, oct 2015. ISSN 1476-4687. doi:10.1038/nature15818. URL http://www.ncbi.nlm.nih.gov/pubmed/26469046. → pages 1, 3

[55] D. B. Nikolov and S. K. Burley. RNA polymerase II transcription initiation: a structural view. *Proceedings of the National Academy of Sciences of the United States of America*, 94(1):15–22, jan 1997. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/8990153http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC33652. → pages 12

[56] H. Niwa, K. Yamamura, and J. Miyazaki. Efficient selection for high-expression transfectants with a novel eukaryotic vector. *Gene*, 108(2): 193–9, dec 1991. ISSN 0378-1119. URL http://www.ncbi.nlm.nih.gov/pubmed/1660837. → pages 6

[57] E. Ntini, A. I. Järvelin, J. Bornholdt, Y. Chen, M. Boyd, M. Jørgensen, R. Andersson, I. Hoof, A. Schein, P. R. Andersen, P. K. Andersen, P. Preker, E. Valen, X. Zhao, V. Pelechano, L. M. Steinmetz, A. Sandelin, and T. H. Jensen. Polyadenylation siteinduced decay of upstream transcripts enforces promoter directionality. *Nature Structural & Molecular Biology*, 20(8): 923–928, jul 2013. ISSN 1545-9993. doi:10.1038/nsmb.2640. URL http://www.nature.com/doifinder/10.1038/nsmb.2640. → pages 9

[58] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin,

J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, jan 2016. ISSN 0305-1048. doi:10.1093/nar/gkv1189. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1189. → pages 19

[59] L. A. Pennacchio, W. Bickmore, A. Dean, M. A. Nobrega, and G. Bejerano. Enhancers: five essential questions. *Nature Reviews Genetics*, 14(4): 288–295, mar 2013. ISSN 1471-0056. doi:10.1038/nrg3458. URL http://www.nature.com/doifinder/10.1038/nrg3458. → pages 9

[60] E. Portales-Casamar, D. J. Swanson, L. Liu, C. N. de Leeuw, K. G. Banks, S. J. Ho Sui, D. L. Fulton, J. Ali, M. Amirabbasi, D. J. Arenillas, N. Babyak, S. F. Black, R. J. Bonaguro, E. Brauer, T. R. Candido, M. Castellarin, J. Chen, Y. Chen, J. C. Cheng, V. Chopra, T. R. Docking, L. Dreolini, C. A. D'Souza, E. K. Flynn, R. Glenn, K. Hatakka, T. G. Hearty, B. Imanian, S. Jiang, S. Khorasan-zadeh, I. Komljenovic, S. Laprise, N. Y. Liao, J. S. Lim, S. Lithwick, F. Liu, J. Liu, M. Lu, M. McConechy, A. J. McLeod, M. Milisavljevic, J. Mis, K. O'Connor, B. Palma, D. L. Palmquist, J. F. Schmouth, M. I. Swanson, B. Tam, A. Ticoll, J. L. Turner, R. Varhol, J. Vermeulen, R. F. Watkins, G. Wilson, B. K. Wong, S. H. Wong, T. Y. Wong, G. S. Yang, A. R. Ypsilanti, S. J. Jones, R. A. Holt, D. Goldowitz, W. W. Wasserman, and E. M. Simpson. A regulatory toolbox of MiniPromoters to drive selective expression in the brain. *Proc Natl Acad Sci U S A*, 107(38):16589–16594, 2010. ISSN 1091-6490. doi:10.1073/pnas.1009158107. URL http://www.ncbi.nlm.nih.gov/pubmed/20807748. → pages 2, 17, 56

[61] P. Preker, J. Nielsen, S. Kammler, S. Lykke-Andersen, M. S. Christensen, C. K. Mapendano, M. H. Schierup, and T. H. Jensen. RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. *Science*, 322(5909):1851–1854, dec 2008. ISSN 0036-8075. doi:10.1126/science.1164096. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1164096. → pages 9

[62] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(Database):D61–D65, jan 2007. ISSN 0305-1048. doi:10.1093/nar/gkl842. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl842. → pages 19

[63] Rare-diseases.ca. CIHR: New Emerging Team for Rare Diseases, 2016. URL http://rare-diseases.ca/. → pages 1

[64] Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, feb 2015. ISSN 1476-4687. doi:10.1038/nature14248. URL http://www.ncbi.nlm.nih.gov/pubmed/25693563http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4530010. → pages 13, 19

[65] N. P. Rodrigues, A. J. Tipping, Z. Wang, and T. Enver. GATA-2 mediated regulation of normal hematopoietic stem/progenitor cell function, myelodysplasia and myeloid leukemia. *The International Journal of Biochemistry & Cell Biology*, 44(3):457–460, mar 2012. ISSN 13572725. doi:10.1016/j.biocel.2011.12.004. URL http://linkinghub.elsevier.com/retrieve/pii/S1357272511003396. → pages 12

[66] N. V. Rozhkov. Global Run-On Sequencing (GRO-seq) Library Preparation

from Drosophila Ovaries. *Methods in molecular biology (Clifton, N.J.)*, 1328:217–30, 2015. ISSN 1940-6029. doi:10.1007/978-1-4939-2851-4_16. URL http://www.ncbi.nlm.nih.gov/pubmed/26324441. → pages 11

[67] L. Samaranch, E. A. Salegio, W. San Sebastian, A. P. Kells, K. D. Foust, J. R. Bringas, C. Lamarre, J. Forsayeth, B. K. Kaspar, and K. S. Bankiewicz. Adeno-Associated Virus Serotype 9 Transduction in the Central Nervous System of Nonhuman Primates. *Human Gene Therapy*, 23(4):382–389, apr 2012. ISSN 1043-0342. doi:10.1089/hum.2011.200. URL http://online.liebertpub.com/doi/abs/10.1089/hum.2011.200. → pages 6

[68] D. Schmidt, M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield, and D. T. Odom. ChIP-seq: Using high-throughput sequencing to discover proteinDNA interactions. *Methods*, 48(3):240–248, jul 2009. ISSN 10462023. doi:10.1016/j.ymeth.2009.03.001. URL http://linkinghub.elsevier.com/retrieve/pii/S1046202309000474. → pages 12

[69] R. SCOLLAY. Gene Therapy. A Brief Overview of the Past, Present, and Future. *Annals of the New York Academy of Sciences*, 953a(1 NEW VISTAS IN):26–30, dec 2001. ISSN 0077-8923. doi:10.1111/j.1749-6632.2001.tb11357.x. URL http://doi.wiley.com/10.1111/j.1749-6632.2001.tb11357.x. → pages 3

[70] C. Sheridan. Gene therapy finds its niche. *Nature Biotechnology*, 29(2): 121–128, feb 2011. ISSN 1087-0156. doi:10.1038/nbt.1769. URL http://www.nature.com/doifinder/10.1038/nbt.1769. → pages 5, 6

[71] T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776–15781, dec 2003. ISSN 0027-8424. doi:10.1073/pnas.2136655100. URL http://www.pnas.org/cgi/doi/10.1073/pnas.2136655100. → pages 10, 11

[72] P. L. Sinn, S. L. Sauter, and P. B. McCray. Gene Therapy Progress and Prospects: Development of improved lentiviral and retroviral vectors design, biosafety, and production. *Gene Therapy*, 12(14):1089–1098, jul 2005. ISSN 0969-7128. doi:10.1038/sj.gt.3302570. URL http://www.oncotarget.com/fulltext/5169http://www.nature.com/doifinder/10.1038/sj.gt.3302570. → pages 5

[73] A. Smit, R. Hubley, and P. Green. RepeatMasker Open-4.0. 2013-2015 ., 2013. URL http://www.repeatmasker.org/. → pages 19

[74] J. A. Stamatoyannopoulos. Illuminating eukaryotic transcription start sites. *Nature Methods*, 7(7):501–503, jul 2010. ISSN 1548-7091. doi:10.1038/nmeth0710-501. URL http://www.nature.com/doifinder/10.1038/nmeth0710-501. → pages 7

[75] C. E. Thomas, A. Ehrhardt, and M. A. Kay. Progress and problems with the use of viral vectors for gene therapy. *Nature Reviews Genetics*, 4(5): 346–358, may 2003. ISSN 14710056. doi:10.1038/nrg1066. URL http://www.nature.com/doifinder/10.1038/nrg1066. → pages 1, 2, 3, 5

[76] N. D. Trinklein, S. F. Aldred, S. J. Hartman, D. I. Schroeder, R. P. Otillar, and R. M. Myers. An abundance of bidirectional promoters in the human genome. *Genome research*, 14(1):62–6, jan 2004. ISSN 1088-9051. doi:10.1101/gr.1982804. URL http://www.ncbi.nlm.nih.gov/pubmed/14707170http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC314279. → pages 9

[77] N. L. van Berkum, E. Lieberman-Aiden, L. Williams, M. Imakaev, A. Gnirke, L. A. Mirny, J. Dekker, and E. S. Lander. Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. *Journal of Visualized Experiments*, (39), may 2010. ISSN 1940-087X. doi:10.3791/1869. URL http://www.jove.com/index/Details.stp?ID=1869. → pages 15

[78] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, 22(9):1798–812, sep 2012. ISSN 1549-5469. doi:10.1101/gr.139105.112. URL http://www.ncbi.nlm.nih.gov/pubmed/22955990http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3431495. → pages 20

[79] J. Wang, J. Zhuang, S. Iyer, X.-Y. Lin, M. C. Greven, B.-H. Kim, J. Moore, B. G. Pierce, X. Dong, D. Virgil, E. Birney, J.-H. Hung, and Z. Weng. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic acids research*, 41(Database

issue):D171–6, jan 2013. ISSN 1362-4962. doi:10.1093/nar/gks1221. URL http://www.ncbi.nlm.nih.gov/pubmed/23203885http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3531197. → pages 20

[80] P. Washbourne and A. McAllister. Techniques for gene transfer into neurons. *Current Opinion in Neurobiology*, 12(5):566–573, oct 2002. ISSN 09594388. doi:10.1016/S0959-4388(02)00365-3. URL http://linkinghub.elsevier.com/retrieve/pii/S0959438802003653. → pages 2

[81] T. Wirth, N. Parker, and S. Ylä-Herttuala. History of gene therapy. *Gene*, 525(2):162–169, aug 2013. ISSN 03781119. doi:10.1016/j.gene.2013.03.137. URL http://linkinghub.elsevier.com/retrieve/pii/S0378111913004344. → pages 1, 2

[82] Z. Wu, A. Asokan, and R. J. Samulski. Adeno-associated Virus Serotypes: Vector Toolkit for Human Gene Therapy. *Molecular Therapy*, 14(3): 316–327, sep 2006. ISSN 15250016. doi:10.1016/j.ymthe.2006.05.009. URL http://linkinghub.elsevier.com/retrieve/pii/S1525001606002048. → pages 6

[83] C. Yang, E. Bolotin, T. Jiang, F. M. Sladek, and E. Martinez. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, 389(1): 52–65, mar 2007. ISSN 03781119. doi:10.1016/j.gene.2006.09.029. URL http://linkinghub.elsevier.com/retrieve/pii/S0378111906006238. → pages 7

[84] M. A. Zabidi, C. D. Arnold, K. Schernhuber, M. Pagani, M. Rath, O. Frank, and A. Stark. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 518(7540):556, 2015. → pages 9

[85] C. Zincarelli, S. Soltys, G. Rengo, and J. E. Rabinowitz. Analysis of AAV serotypes 1-9 mediated gene expression and tropism in mice after systemic injection. *Molecular therapy : the journal of the American Society of Gene Therapy*, 16(6):1073–1080, jun 2008. ISSN 1525-0016. doi:10.1038/mt.2008.76. URL http://www.sciencedirect.com/science/article/pii/S1525001616317324http://linkinghub.elsevier.com/retrieve/pii/S1525001616317324. → pages 6