Identification and Characterization of *Cis*-Regulatory Elements in the Human Genome

by

Mel Englund

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Genetics and Genomics)
in The University of Michigan
2023

Doctoral Committee:

Associate Professor Alan P Boyle, Chair
Professor Anthony Antonellis
Associate Professor Sue Hammoud
Professor Diane M Robins
Professor Trisha Wittkopp

Mel Englund
melyssae@umich.edu
ORCID iD: 0000-0003-3551-2877

To the family we choose and who chooses us.

Yana & Mary,

Peggy, Jean, and Jim

To my new family,

Chad,

Milo and Jackson

To my lab family, and to Torrin

I wouldn't be here without any of you

ACKNOWLEDGEMENTS

So many truly wonderful people were part of the process of this work, starting of course with the members of the Boyle lab. I have learned so much from all of you, have been honored to be able to work alongside you and have had so much fun getting to know you. I cannot wait to see where we all end up, and to see the Boyle lab continue to grow. Thank you in particular to my wet lab family, Torrin, Sierra, Camille and Jessica. You have been such a part of my life for the last five years, you have made this work easier, encouraged me, and taught me so much. Who I become as a scientist will always a little bit come from all of you. Torrin, you practically dragged me into the Boyle lab, and refused to let me give up, so thank you. You have never not been a loyal, supportive friend. To Chris, partner in commiseration and thesis crime, thanks for helping me feel I was not alone during this final phase. Also to Nandini, what a gem you are, I am so lucky to have met you and I already miss our conversations. And of course, thank you to Alan, who gave me a second chance in many ways. Your consistent support, patience, and faith in your students are the reasons I have made it this far. It has not always been easy, but it has been an incredible journey and I am honored to have had the chance to be in your lab. You make the lab a welcoming place, a family (and I swear I am not just saying that because of the fun parties). I am so excited to see how the lab has grown and to see how it continues to grow in the future..

To the Department of Human Genetics, to all the students, staff, faculty, thank you for training me, supporting me. This has been such an incredible experience. I genuinely look back on my days in HG541 with fondness. To Karen, Molly, and Ashley who were unfailingly kind and helpful to confused graduate students. Thank you particularly to Dr. Tony Antonellis and Dr. John Moran for building the Genetics Training Program. I hope you have some idea of how meaningful and enriching that program is. It taught me to think of myself as a scientist and other students as colleagues, made me want to live up to that ideal. Thanks to Dr. John Moran and Dr. Shigeki Iwase for their

generosity with lab resources. Thank you to Dr. Jacob Kitzman, for giving me a foundation I continue to use and for teaching me to think clearly, critically, and deeply about all the details of my experiments. To Becca and Sadie, my friends in PhD life, I am so glad to have met you. Thank you for all your kindness and support.

This work is also dedicated to my family, related and adopted. Yana and Mary. You have not done a PhD, but after all the long phone calls you might be qualified for an honorary one. I love you, you are my family, thank you for sharing this with me, supporting me. Mary, you kept a crazy promise and literally moved me here, and made the start of this journey a treasured memory. Yana, I cannot think of any words that are enough to express my gratitude. You are my family. Thank you. To my partner Chad, who met me just around prelims and still decided to stay, I don't know why, but I am so glad you did. Your quiet, patient presence and support have been such a comfort. I am so grateful to have found you and been able to go through this with you, I cannot imagine how I could have been so lucky in you.

Peggy, thank you from the bottom of my heart for giving me space to be, to grow, to shine, the way you saw I could. Thank you for what you sacrificed to help us, for being the best of what a mom can be. You are strong and kind and you shaped who I am. You are my real life guardian angel. To Jim and Jean. I am so lucky to have had more than the usual number of parents in my life, for all the different phases. You also gave me space to grow, you believed in me, supported me, took care of the things I could not, when you did not have to. Taught me to drive, to budget, and to recognize the finite nature of time particularly where it applies to homework. Our conversations are so precious to me. I am so lucky to have you all as my family and I cannot express enough how much I would not be here without you.

I would also like to thank and acknowledge my funding received through the NIH Training Grant Michigan Predoctoral Training in Genetics (T32GM007544).

TABLE OF CONTENTS

LIST OF FIGURES

Figure

CHAPTER I

CHAPTER II

CHAPTER III

CHAPTER IV

ABSTRACT

Maintaining precise spatiotemporal control of gene expression patterns is essential for the proper functioning of cells, tissues, and organisms. In eukaryotic cells, this control is established through the use of multiple systems, which regulate expression at the levels of chromatin, DNA, RNA and proteins. At the DNA level, regulation is controlled by *cis*-regulatory elements (CREs): modular non-coding sequences with varying functions mediated through the binding of transcription factors. Variation within these non-coding sequences is increasingly understood to contribute to human phenotype and disease, however mapping and characterizing CREs is challenging, as non-coding sequences comprise 98.8% of the human genome. For this reason, in addition to their flexibility and scalability, episomal reporter assays have been, and continue to be, the primary tool used to test for CRE function in non-coding sequences.

As our appreciation of the complexity and interconnectedness of *cis*-regulatory systems increases, so does the complexity of the assays designed to interrogate CRE function. However, with increasing complexity comes the potential for confounding effects within assay systems. In Chapter 1 of this thesis, I review the roles of different CRE classes in cellular regulation, the types of assays used to characterize CREs, and address the ways in which the transcription factor, CRE, and chromatin layers of regulation interconnect. I discuss how models are needed that account for the interactions of elements within and across regulatory systems, and for the role of silencers in these systems, in a genomic context. In Chapter 2, I discuss several considerations for plasmid-based reporter assay design in light of this increasing diversity and complexity. I provide supporting data for the impact of each component, discuss how it can impact interpretation, provide models for improving design, and demonstrate the utility of plasmid-based systems for modeling CRE mechanisms. Due to the potential functionality of each component in a complex plasmid system, tools that

support full, rather than partial plasmid sequence validation are needed. I address this in Chapter 3, where I present OnRamp, a combined protocol and analysis toolset that leverages the long-read nature of the nanopore sequencing platform to facilitate rapid, affordable, and accessible multiplexed full-plasmid sequencing.

Finally, in Chapter 4, I use a modified regulatory assay panel to characterize enhancer, silencer, and enhancer blocker activity across a single regulatory topologically associating domain (TAD) of the human genome, containing the genes *PRDM1* (crucial to B-, NK- and T-lymphocyte differentiation) and *ATG5* (an essential autophagy-related gene). Using assay data and previously generated high-throughput datasets, I generate a model of regulation in this region which incorporates chromatin, CRE, and transcription-factor level systems to account for the differential regulation of the two genes both within the TAD and across two cancer cell lines - K562 (myelogenous leukemia) and HepG2 (hepatocellular carcinoma).

Together, this work contributes to the improved design and fidelity of plasmid-based reporter assays for the study of *cis*-regulatory elements and generates a functional model for regulatory dynamics in a previously relatively uncharacterized region of the human genome containing genes important for basic cellular and immune function.

CHAPTER I
Introduction

The importance to proper cell function of maintaining precise control of every aspect of gene expression is evident through the many overlapping regulatory systems active in eukaryotic cells, where gene expression is mediated at the levels of chromatin, DNA, RNA, and protein.

In eukaryotic genomes, DNA is packaged in chromatin; it is wrapped around nucleosomes, histone octamers composed of histones H2A, H2B, H3, and H4 in duplicates [1]. DNA packaged in nucleosomes is less accessible to binding of RNA polymerase II (RNA Pol II) and the transcription factors (TFs) needed to mediate function [2]. This provides a layer of regulatory control, based on whether DNA is wrapped in nucleosomes or not. Additionally, nucleosome-wrapped DNA can be packed more tightly and be less accessible (heterochromatin), or less tightly packed and more accessible (euchromatin), and as a result more available for displacement of nucleosomes and transcriptional activation, providing an additional layer of regulation [2].

At the DNA level, *cis*-regulatory elements (CRE) drive regulation by mediating rates of transcription initiation in a timing, signal-responsive, and cell-type-specific manner [3,4]. CREs are non-coding DNA sequences that can be located adjacent to, within the intronic sequences of, or distal from, genes. Their function is determined by the multiple factors that bind them in a sequence-specific manner and secondary factors that bind in a non-sequence specific manner [5]. At the level of epigenetic modifications of DNA, methylation of cytosines [6] can prevent gene expression [7,8] through inhibition of transcription factor binding or recruitment of repressive factors. DNA methylation primarily occurs at C-G base pairs or CpGs, regulates X-inactivation, and is involved in silencing retroviral elements [9], imprinting [10], and tissue-specific gene regulation [11]. (See Moore, Le and Fan 2013 for a review [12]).

Regulation at the level of RNA includes alternative splicing of exons, where segments of coding sequence are joined and interspersed non-coding ‘intronic’ sequence is removed to produce different isoforms of the same gene [13,14]. RNA is also regulated through alternative polyadenylation [15], the addition of post-transcriptional modifications [16–18], controlling its localization [19], and microRNA-mediated targeted transcript degradation [20]. (See Licatalosi and Darnell 2010 for a review on this topic [21]). Once translated, cellular regulation continues, targeting proteins through co-binding of other factors, covalent modifications, degradation, and transportation [22].

By the application of all of these layers of regulation, sometimes simultaneously, cells maintain precise control of gene expression throughout development and differentiation, as well as during responses to transient environmental signals. **Figure 1** illustrates how all of these regulatory layers coincide. While all of these systems are important to cell function, here I will focus on the role of *cis*-regulatory elements (CRE) in gene regulation, as well as chromatin-mediated regulation to the degree that it interacts and overlaps with the functions of *cis*-regulatory control.

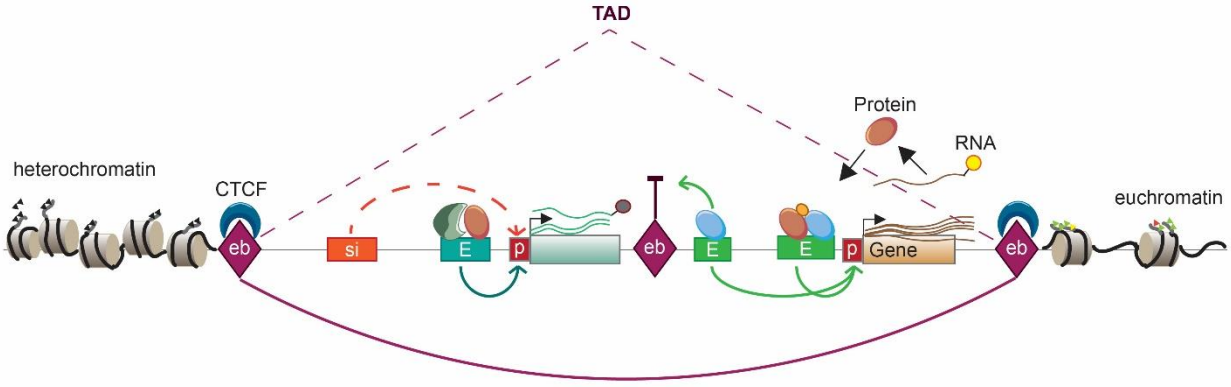


Figure 1.1 Coordinated layers of regulation in eukaryotic cells
Chromatin state, transcription factor binding, CRE regulation, RNA and protein modifications all contribute to regulation. eb – enhancer blocker, si – silencer, p – promoter, e – enhancer, TAD – topologically associating domain

1.1 *Cis*-Regulatory Elements

Mutations impacting non-coding *cis*-regulatory elements contribute to human disease through alteration of gene expression patterns. Of the disease-associated variants identified through genome-wide association studies, 88% fall in non-coding regions of the human genome [23]. CRE mutations are also increasingly being

understood to play a role in cancer biology [24]. A number of different enhancer mutations have been demonstrated to contribute to human diseases, including beta-thalassemia [334], Burkitt's lymphoma [335], and polydactyly [336].

Cis-regulatory elements (CRE) are non-coding DNA sequences that function to regulate genes at the level of transcription, through functions which are mediated by factors that bind these sequences. Within this definition, there are different types of CREs, differentiated by direction of effect on transcription and function. These element types include promoters, enhancers, silencers, enhancer blockers, and insulators [3]. While here I focus on the roles of non-coding CREs, it should be noted that coding CREs have also been identified. Exonic enhancers, sequences which both code for proteins and act as enhancers are an example of this [337]. Together, sets of multiple CREs of different types coordinate to form the regulatory unit of a gene, and multiple genes with their regulatory units can be organized into larger domains. Below I provide an overview of: what we do and do not know about CRE mechanisms of action (section 1.2), the assays and systems we use to answer these questions (and their advantages and limitations) (section 1.3), the combinatorial logic of interactions between CREs of the same and different classes and how this relates to chromatin organization (section 1.4), and discuss models for how these interactions and different layers of regulation interact at the domain and cell levels, as well as themes and ongoing questions in the field (section 1.5).

1.2 CRE Classes and Their Mechanisms

1.2.1 Promoters

RNA Polymerase II promoters are the sequences through which signals for control of gene expression are organized. They contain a transcription start site, a binding site for RNA Pol II, the enzyme which transcribes DNA into RNA. They also contain binding sites for general transcription factors (GTFs) and Mediator, which together with RNA Pol II form the pre-initiation complex (PIC) [25]. These sites are collectively referred to as the core promoter region [26,27]. GTFs facilitate TSS recognition, RNA Pol II recruitment, and promoter recognition [28]. The PIC is sufficient to drive low (basal) levels of transcription which are often not sufficient for biological

function [29]. In order to achieve increased transcription, transcription factors can bind at sites upstream of the core promoter in the proximal promoter region ~200bp upstream of the core promoter [30] to form multiple synergistically-acting regulatory units [31]. These transcription factors typically have a DNA binding domain and an activating domain and can bind other co-factors that function through protein-protein interactions [32,33]. Repressive factors can also bind at these promoter-proximal sites, to repress transcription [34].

RNA Pol II promoters have bi-directional transcription activity, however transcripts in the anti-sense direction are not productive and remain short, while in the coding direction elongation occurs [35,36]. These promoters also have distinctive histone modification patterns, characterized by nucleosome depletion, and H3K4me3 & H3K27ac at flanking histones [37–39]. There are many excellent reviews available covering additional details on promoters, transcription factors, and the processes of transcription [40,41]. Eukaryotic cells also make use of RNA Pol I and III promoters, which control transcription of ribosomal RNAs (Pol I) and transfer RNAs, 5S ribosomal RNA, and snoRNAs (Pol III) [338]. Here, I focus primarily on the role of RNA Pol II promoters in mediating transcription initiation through acting as the site for integration of regulatory signals from transcription factors, enhancers and silencers.

1.2.2 Enhancers

Enhancers are DNA sequences which bind activating transcription factors and increase transcription at their target promoters. They are largely considered independent, modular elements. They can be located distally or proximally to their target promoters and can perform their function regardless of placement (are position-independent) relative to the promoter [42]. They are also orientation-independent [43,44]. Enhancers are associated with a histone modification pattern which differs from that of promoters, and includes high H3K4me1 vs H3K4me3 ratios, and H3K27ac [45,46]. They are also often bound by CBP-p300 histone acetyltransferase [47,48].

Enhancers are made up of multiple transcription factor binding sites (TFBS), bound by TFs. Through the combinatorial logic of TFBS arrangement, the spacing, type, orientation, and number of TFBS are variably constrained and impact overall function

[55]; [273]. The presence of multiple TFBS in enhancers allows different cell types expressing different suites of TFs to use the same enhancers. It also allows cells multiple inputs into the activation of an enhancer. Through the need for synergistic or additive interactions of multiple TFs, a cell can limit enhancer activation to more specific states. For instance activation can be limited to a specific stage of cell cycle, as well as in response to a specific extracellular signal, or the presence of another gene product. The precise architecture and logic of these arrangements has not yet been fully elucidated. There exist different models for how TFBS combine to determine enhancer function. These models describe the different degrees to which groups of TFs act as independent modular units, versus having functions that depend on interdependent interactions between factors, and the degree to which protein-DNA vs protein-protein interactions contribute to function (further detail provided in Chapter 4) [49–51].

In order to act on promoters, where regulatory information is integrated, the majority of models for enhancer action agree that enhancers must come into proximity of their target promoter, where they bring activating factors into contact with the PIC machinery, recruit chromatin modifiers, facilitate RNA Pol II pause release, and/or increase local concentrations of RNA Pol II [52]. Broadly, enhancers have been shown to recruit TFs and bring them into contact with the promoter, suggesting a role for TF-TF synergistic interactions [53,54]. Precisely how enhancers mediate increases in transcription is still not fully understood, and remains one of the more important questions in the field of regulation [55]. Studies of the kinetics of enhancer activation and how they relate to transcription output are based on the observation that enhancer-activated transcription is not continuous but rather occurs in ‘bursts’ [56]. The number of RNA Pol II complexes that transcribe a gene simultaneously during a single burst can be increased or decreased, and is dependent on the promoter structure [57]. However the frequency with which a burst of transcription is activated is modulated by enhancer activation [58,59].

There are also different models for how enhancers achieve the proximity needed to activate promoters. In the tracking model, factors are recruited at the enhancer and travel along the DNA until reaching the target promoter [60]. Linking is similar to tracking except that a chain of connected factors bind each other along the DNA between

enhancer and promoter [61] to establish the progression of the signal rather than of a polymerase. In the looping model, DNA folds to bring the enhancer and its target promoter into contact [62]. While tracking would not be feasible over long distances as it would interfere with, or be interfered with, by any other bound factors or genes along the way, there is evidence for RNA Pol II tracking from a classically studied enhancer, HS2, and its target promoter [63,64]. Current evidence supports looping as the primary genomic mechanism for ensuring enhancer-promoter proximity [65,66]. Enhancer-promoter contacts are supported by *in vitro* studies [67] as well as by chromatin conformation capture assays (3C/4C/HiC), where genomic DNA is restriction-digested and ligated *in situ*, then sequenced, to detect which sequences are located proximally in nuclear space. The frequency of interactions can then be quantified using the number of reads containing a pair of sequences together as a proxy [68].

Similar to promoters, enhancers also undergo transcription [69]. The widespread detection of enhancer RNAs (eRNAs) [70–72], has recently been made possible by the development of techniques for capture of small RNAs genome-wide [73]. Whether these enhancer RNAs are independently functional, or a passive byproduct of RNA Pol II localization to enhancers [74], and what causes promoters but not enhancers to undergo productive elongation in one direction [75] are all important ongoing areas of investigation. eRNA presence has been used as a marker of enhancer activity and strength [76] and through correlation with mRNA expression timing and location, used to link enhancers to their potential target promoters [77].

1.2.3 Silencers

Despite the first examples of silencers being identified around the same time as early examples of enhancers [78], silencers have been relatively under-studied, particularly in mammalian models, until very recently. One reason for this may be that it is possible to model cell-type specific gene regulation without the inclusion of an independent CRE silencer class, discussed below (section 1.5). Prior to 2020, there were only sparse individual examples of silencer elements. Some of the more well-studied silencer examples include the *HMRE* silencer in yeast [78], the *VRE* silencer in *Drosophila [79]*, the *CD4* intronic silencer [80], and the constitutive T39 silencer element

[81], more recently profiled by Qi et al. Additionally, fewer repressive transcription factors have been identified than activating transcription factors. Perhaps the most well-known repressive factor (TFs which mediate silencing) is REST (repressor element 1-silencing transcription factor or neuron-restrictive silencer factor (NRSF)), which represses neuronal genes in non-neuronal cell types [82]. Recently however, a number of high-throughput silencer assays in *Drosophila [83]* and in mammalian cells [84,85] as well as predictive computational models generated in these and other studies [86,87] have rapidly increased the number of characterized silencer elements. This has contributed to a better understanding of the diversity, widespread distribution over the genome, and similarities and differences to enhancers, of silencer elements.

Silencers are currently seen as being repressive counterparts to enhancers, with analogous roles and mechanisms. Silencers comprise DNA sequences that bind repressive factors and mediate down-regulation of gene expression [88]. Like enhancers, they can be located proximal or distal to target genes and can be found in both intronic and intergenic sequences [85]. They function independent of orientation, but reports of their position-independence in an episomal context have been mixed [81,89], and loop to contact their target promoters [90]. Silencers also have tissue-specific expression patterns that vary [84,85]. Their sequences are conserved to about the same degree as enhancer sequences, and they are similarly enriched for the presence of GWAS SNPs (single-nucleotide polymorphisms from genome-wide association studies) [84], supporting their relevance to human disease.

While some papers have proposed a link between silencer elements and specific histone marks and factors, including H3K27me3 and the enzyme PRC2 which places this mark (polychrome repressive complex 2) [83–87] and HP1-bound H3K9me3 [85], this link is tentative. H3K27me3 enrichment was not significant in functional studies and enrichment for H3K9me3 was weak. This inability to determine silencer-associated chromatin modification patterns is one factor which has limited the study of silencers, as they can greatly assist in the prioritization of regions for testing, as they have in enhancers [45,46]. This inability to find enriched marks might also in itself provide information regarding a current question in the field: whether silencers form a single unified class with similar mechanisms of action, or are made up of a number of different

classes of elements with differing mechanisms but the same direction of effect. Perhaps no unifying histone mark can be identified because silencers as a class are more diverse than enhancers. For instance, some silencers are dual enhancer-silencer elements (see section 1.2.5). This also reflects the weakness of using chromatin modifications alone as predictive. In the case of enhancers, a focus on 'classical' enhancer marks might bias studies to one set of enhancers that share similar behavior while excluding others which behave differently and have different modifications [91].

These same high-throughput silencer studies also searched for TF-binding or motif enrichment for known repressive factors. They found enrichment for the repressor Snail in *Drosophila*, REST motifs in HepG2, KLF12 binding in K562 cells, and AP2 motifs in both [46,83,92].

Silencers are believed to act through binding of repressive transcription factors [93] or recruitment of factors which place repressive chromatin modifications like HSK27me3. Models of active enhancer interference by silencers, or repression of enhancing by binding of competing repressive TFs within an enhancer have also been proposed [94]. Which of these mechanisms are present, for which silencers and in which combinations, and if there are different silencers separated by mechanism, are all important and ongoing areas of inquiry. See recent reviews for additional information on silencers [85,94,95].

1.2.4 Enhancer Blockers and Insulators

Enhancer blockers are defined through their ability to prevent communication between an enhancer and a promoter in a position-dependent manner - they function only when placed between an enhancer and its target promoter. Early work on enhancer-blocking elements completed in *Drosophila* established basic principles of function through studies of the impacts of SuHw-binding sites in the *gypsy* retrotransposon on expression in the *yellow* locus [96].

Barrier insulators function as boundaries for chromatin domains, preventing the spread of repressive chromatin states across the insulator, protecting adjacent genes from repression through chromatin compaction. As discussed in the beginning of this chapter, DNA packaged in nucleosomes is less accessible to binding of functional

factors [1]. The exception to this is pioneer transcription factors, which are able to bind nucleosome-wrapped DNA as a first step in opening an enhancer region for activation [97]. Nucleosome-wrapped DNA can be either more (heterochromatin) or less (euchromatin) tightly packaged [2], providing a layer of regulation at the chromatin level. Heterochromatin and euchromatin are associated with repressive or active transcription and histone modifications, respectively [98,99], so protection of actively transcribed genes from the spreading of repressive chromatin by insulators is essential to proper regulation of expression. Studies in *Drosophila* have also been instrumental in shaping our understanding of insulator function - the scs and scs' insulator elements that flank the hsp-70 locus in *Drosophila* are important models of insulator activity [100] and were used to establish the design of enhancer-blocking assays [101].

Unlike in *Drosophila* where a number of different factors have been linked to enhancer blocking/insulator function [102], in vertebrates, CTCF, a conserved zinc-finger binding protein [103], is the primary protein factor involved, so a majority of research into enhancer blocker and insulator functions in humans focuses on CTCF-binding elements. While enhancer blocker and barrier activity are intertwined in human models and often discussed together, these functions are separable [104]. The first characterized example of an enhancer blocker in vertebrates, the cHS4 element found at the 5' end of the chicken β-globin locus, displays both of these characteristics - CTCF binding, and separable enhancer blocker and insulator functions [105,106], and functions in *Drosophila* and human cells [107]. (See Chapter 2, section 2.6.4 for a detailed discussion and data related to cHS4 activity in a plasmid context). Although the majority of enhancer blocker/insulator sites characterized to date are CTCF-binding, not all are. Notable exceptions include GATA repeat sequences studied in human cell lines [108], and tRNA genes [109,110].

Enhancer blocker (eb) activity can be studied using plasmid-based assays, where a single element is separately placed upstream of, or between, an enhancer and promoter, and eb activity is determined by positional effect. If the element is an enhancer blocker, a decrease in expression occurs only when it is placed between the enhancer and promoter (discussed further in Chapter 2) [89,105,111]. Tests for barrier insulator activity require chromatin context, and so use integrated constructs with

flanking insulators, where the readout is repression, or protection from repression, of the integrated reporter by spreading native heterochromatin [106,109,112].

The genomic roles and proposed mechanisms of insulators and enhancer blockers have significant overlap. As there are many genes and regulatory elements throughout the genome, and many regions that must be properly repressed or activated in order to maintain correct patterns of activation, mechanisms are needed to separate adjacent domains with differing activities. Enhancer blockers constrain enhancer (and silencer) activity to specific genes based on their placement, and insulators prevent spread of repressed chromatin states across neighboring genes and regulatory elements (discussed further in section 1.4). There are a number of proposed mechanisms for this activity, which are discussed in detail in Chapter 2, section 2.6.4 [113].

Briefly, the models for enhancer blocker activity involve either direct contact, a processive signal mechanism, or looping of paired enhancer blockers. In the processive model, (as in the enhancer tracking model discussed above [114]) some signal directed from the enhancer to the promoter is disrupted by factors present at the intervening enhancer blocker [113]. ‘Direct contact’ describes a model where enhancer blockers directly contact promoters, creating a physical DNA structure that prevents enhancer contact with the promoter [104]. In the looping model, pairs of enhancer blockers bind to each other, effectively looping out the intervening DNA sequence, increasing the proximity of elements within the loop but decreasing proximity of elements inside the loop to those outside it [115]. Within this loop model, the ability of just a single element to block enhancer-promoter communication on a plasmid is explained by observations of CTCF-dependent tethering of chromatin (or plasmids) to nucleolar surfaces, which would physically separate segments DNA on either side of the bound site [115].

There are many more aspects to enhancer blocking and insulator activity than can be covered here, in part because studies of these elements necessarily require investigating mechanisms of enhancer-promoter communication, and are related to organizational principles of cells at many different levels, tying together DNA and chromatin, regulatory elements and genes, and domains at the level of chromosomes, megabases and multi-kilobase loops.

1.2.5 Overlap Between CRE Classes

While the separation of regulatory elements into groups based on direction of effect and position-dependence is useful and reflects actual underlying differences, it is also important to note that there are not always distinct lines of separation between these classes. There are a number of features of promoters, enhancers, silencers, and enhancer blockers that overlap. For instance, as discussed above, despite the utility of histone modification patterns for narrowing potential enhancer candidates for functional testing, not all enhancers share these 'classical' modification patterns [91]. In fact, while these marks were initially identified as useful in distinguishing between enhancers and promoters, recent comparisons show that these marks occur on a spectrum and that the differences between some enhancers and some promoters are subtle, with regards to both chromatin marks [116] and especially with the discovery of transcription at enhancers [117,118]. Additionally, many enhancers can act as promoters *in vitro* [119], and promoters can act as enhancers [120].

Another emerging complication for CRE class distinction is driven by the discovery of elements with dual silencer-enhancer activity. These elements behave as enhancers in one cell type and silencers in another. Dual silencer-enhancers are not uncommon, as they were observed in all of the recent high-throughput silencer studies which tested elements across multiple tissue contexts [83–85]. Possibly, the diversity across both enhancer and silencer classes could be explained by the presence of these dual CREs which are included in both groups. This might also explain the similarity of silencers and enhancers (orientation- and position-dependence, looping, similar conservation) and support a model of silencer mechanism that is the same as enhancers, but where cognate repressive factors and repressive chromatin modifiers are looped into promoter proximity instead of activating factors. Whether this is the case, whether all enhancers and silencers have dual potential, or whether these represent separate adjacent elements [95], or identical elements with differing TF binding, are all important and ongoing areas of research.

Finally, as discussed above (and below), CTCF is central to enhancer blocker and insulator activity in mammals. However CTCF has been observed to also have activating and repressing activities, and involvement in a variety of other cellular

processes [121]. This is not to indicate that there is no difference between element classes, but rather that the difference between them may be less than previously thought, and the similarities and differences may in fact provide additional insight to which characteristics drive activity and which are markers of activity.

1.3 Assays for Characterization of CREs

Cis-regulatory element activities have been studied using a variety of assay types, both low- and high-throughput. These fall primarily into four groups: episomal assays, integrated assays (in cell lines), deletion/interference assays, and animal models, with some overlap between the groups.

Episomal assays are a fundamental tool for the study of CREs and both low- and high-throughput assays have been instrumental in many important discoveries in enhancer [122], promoter [120], and silencer biology [84]. Episomal regulatory assays typically have a test element inserted into a plasmid in conjunction with other regulatory elements depending on the assay, and use a reporter gene as the readout for expression. In low-throughput assays, typical reporters include luciferase, GFP and beta-galactosidase [81,89]. In high-throughput assays (massively parallel reporter assays, or MPRAs) mRNA is frequently used as the readout, as it can be quantitated using next-generation sequencing, or fluorescence-activated cell sorting (FACS) [123]. The advantage of plasmid-based assays is the ease with which a cloned construct can be created *in vitro*, the lack of integration, meaning there is low likelihood of interference with normal cell function due to DNA integration, and the scale; in high-throughput assays, thousands of elements can be tested at once [123]. The primary limitation is the temporary, episomal nature. In most episomal assays plasmids are not maintained across cell divisions, and as they are not integrated, an element's activity may differ from its genomic activity due to lack of chromatin context (discussed further in Chapter 2). Low-throughput assays lack scale, but have high fidelity, where high-throughput assays sacrifice increased false positive and false negative rates, in exchange for scale.

Integrated assays have the advantage of better reflecting native chromatin dynamics, as elements are inserted into chromatin using lentivirus or transposases. In the majority of cases, however, while integration does add chromatin context, it does

not add the actual *in situ* native chromatin context of that particular element, if constructs are randomly integrated [124]. Additionally, without proper protection using enhancer blockers, assays using integration can also be subject to positional variation due to heterochromatin spreading or impacted by neighboring regulatory elements, adding a degree of uncertainty to results. Animal models, typically in mouse, *Drosophila*, or zebrafish, also use integrated constructs and are by necessity typically low-throughput, however they can provide important data on developmental patterns of CRE activity and spatiotemporal specificity [125,126].

Deletion/disruption assays have been made possible at large scales due to the development of CRISPR/Cas9 technology [127], and provide advantages beyond episomal and integrated systems. The CRISPR/Cas9 system can be used to direct targeted deletions or to recruit repressive or activating co-factors to non-coding genomic regions [128,129]. This allows for manipulation of elements in their native context, on a high-throughput scale, and importantly is one of the few reliable ways to connect a CRE to its target promoter. It additionally tests CRE necessity for function, where episomal assays test only sufficiency. Limitations of this method are the potential for cryptic off-target mutations at other genes or sites which impact readout, and the need for a selectable phenotype [130].

Through the use of a variety of assays, different aspects of CRE function have been tested, however episomal assays remain a primary tool for interrogation of CRE action and mechanism. In Chapter 2 of this thesis, I discuss in greater detail design principles and issues for episomal reporter assays and the power of plasmid-based assays for modeling CRE activity.

1.4 CRE Interactions and Chromatin Domains

Many of the above described assays focus on characterizing a single element type or mechanism at a time, particularly high-throughput assays, due to limitations of scale. However in a genomic context, gene expression is dynamically regulated by multiple elements, and across multiple layers of regulation, all of which can vary between cell types or cell states.

1.4.1 Multi-Element Interactions

Similar to how the function of an enhancer is the result of the combined activities of multiple TFs and the interactions and synergy between those factors, the expression of a gene is often mediated by multiple interacting CREs [131]. These form what I call the gene's regulatory unit, the collection of all the CREs contributing to its expression in a given cell type and state. Each gene promoter is contacted by 4.75 enhancers on average, and in addition, enhancers can be components of multiple regulatory units at once - an estimated 1/4 of enhancers contact two or more promoters [132]. The potential for enhancer blocker-promoter contact is built into the direct-contact model of enhancer blocking, and promoter-promoter contacts by one estimate made up 9% of chromatin contacts across multiple cell lines [133].

Having many elements contributing to regulation is postulated to provide both redundancy and specificity to gene expression. In many cases, certain enhancers are essentially redundant [134], and can compensate in the case that another enhancer loses function through mutation [135]. And similar to using multiple TFs expressed as a result of varying signals or processes in a cell, the requirement for multiple enhancers allows for a system that is controllable in terms of timing and degree of expression through the use of multiple inputs responsive to different conditions [55,135,136].

1.4.2 Controlling Element Interactions

Given that many elements can interact across many regulatory units, and that enhancer and silencer action can be mediated from a great distance, clearly cellular mechanisms must be present which limit these interactions, otherwise expression of every gene in a domain would be activated by every element, to the same level. There are three primary mechanisms that limit CRE-CRE contact: proximity, specificity, and barriers and boundaries. Proximity as a limiting factor seems as if it would be inconsistent with the ability of enhancers, silencers and promoters to loop over long distances. However proximity of a CRE and promoter does have a contributing effect on expression, within the constraints of boundaries (see below) or over shorter distances. In Chapter 2, section 2.4.4 I present support for a decrease in expression mediated by increasing enhancer-promoter distance on a scale of hundreds of bases in a plasmid

context. Studies from two other groups support this effect on a 100bp scale [137], and at a larger genomic scale [138]. Additionally, studies of shadow enhancers (redundant enhancers) show a tendency for the primary, active, enhancer in a redundant pair to be the enhancer located closer to the promoter [139]. The second limiting factor is specificity. Enhancer-promoter specificity is thought to be mediated by promoter core sequences and compatibility between enhancer- and promoter-bound transcription factors, and is an area of ongoing research [140,141]. Silencer-promoter specificity has also been observed [83].

1.4.3 Barriers: Insulators, TADs, and Looping

The final factor involves the presence of chromatin loops and segments which form discrete domains with different regulatory properties, across the genome. There is much fascinating work being done in this area which cannot possibly be covered in one section alone; I provide a simple overview below.

Chromosomal organizational structures are generated through looping and folding of chromosomes such that active and repressed regions are not necessarily contiguous along the chromosome, and such that there are multiple layers of organization from large- to small-scale, associated with physical and functional structures. At the largest size scale, individual chromosomes are separated into multiple A/B compartments associated with transcriptional activity/repression, respectively [142]. Within these compartments, chromatin is further organized into topologically associated domains (TADs), defined by chromatin capture methods to be regions where there is a higher degree of contact among the sequences within the domain than between sequences inside and outside of it [143]. TADs form a fundamental functional, as well as physical, organizing unit of the cell. TADs can contain multiple genes and their regulatory units, but CREs within a domain rarely contact any genes outside the TAD [144] This limits cross-interactions to those few genes within the domain, rather than all genes on the chromosome. Domain structure is also associated broadly with segments of chromatin activation or repression and with degrees of expression of genes within that domain [145,146]. TADs are formed through looping of DNA mediated by cohesin, and flanked by enhancer-blocking insulator elements bound by cohesin [147,148].

This structure provides an additional piece to the explanation of CTCF-bound enhancer blocker and insulator mechanisms, and positions them in their genomic role. These loops not only limit interactions with elements outside the loop, but also are thought to promote CRE-promoter communication by decreasing distances for sequences inside the loop [113]. Finally, within TADs, more cell-type specific loop structures, including sub-TADS and enhancer-promoter contact loops can form, through similar Cohesin-mediated mechanisms [149]. While TAD structures are largely conserved across tissues, sub-TADs show more cell-type specific variation [113]. This final organizing principle also forms the structure within which enhancer-promoter specificity and proximity can operate, to provide an additional layer of gene-CRE specificity among the multiple regulatory units in a TAD, and without which those principles might be insufficient to mediate proper limitation of gene-CRE contact.

1.5 Models for Regulatory Logic Across a Cell

Across the aspects of CRE activity and function and layers of regulation discussed in this chapter, a number of common themes and principles of regulation emerge. The first are the principles of robustness and precision through multiplicity [4,51,55]. At every level of regulation discussed here, there are multiple inputs which combine together to produce a functional output. At the highest level, there are regulatory systems for each stage of the process - chromatin states, CRE DNA elements and DNA methylation, RNA, and protein. Chromatin states involve multi-layered organizational systems, and histone modifications are not typically found alone, but rather sets of modifications combine to establish the ultimate profile for each element type. At the level of a gene regulatory unit, multiple CREs regulate the same gene, and CREs can act on multiple genes. Even within a CRE like an enhancer, multiple transcription factors bind and through their collective states and interactions determine the element’s function. The presence of multiple inputs seems to serve two purposes: it provides redundancy, which creates robustness against the impacts of mutation and variation, and it allows each element or gene unit to be precisely controlled, through the requirement for coordination of multiple elements (or TFs) each of which can be modulated in response to different inputs (timing, cell type, signaling).

The second theme is that of specificity, additivity, and synergy. It is perhaps not surprising that since CRE function is driven by bound factors, the overall principles of these elements reflect principles of interaction for their components. Within an enhancer and within the set of CREs regulating a gene, interactions can be element-specific, and the combination of two TFs or two CREs can give an output that is additive (the sum of the activities of the individual parts), sub-additive, or synergistic (more than the sum of the parts). However, across both individual and collective CREs, we do not yet know enough about the mechanisms or driving factors of these synergistic and specific interactions to predict when they will occur, between what elements. Answering these questions also has the potential to improve our understanding of the mechanisms of CRE activity.

Adding another layer, these regulatory systems have dynamic activities across cell lines, and within cell populations. The state of an element, gene, or sub-TAD, can vary across evolution, development, cell type, and cell state. Additionally, there is often an aspect of cell-to-cell variation relevant to each of the topics addressed above, which was not discussed here. Novel single-cell technologies reveal the degree to which assessments of function and expression have actually been measurements of averages across a population [150], which can sometimes mask our ability to discern mechanisms.

The final theme is the interdependence and interactions of the regulatory layers. While models of regulatory activity necessarily dissect and examine individual components in order to clearly answer specific questions, all of these activities occur in a cell concurrently and many are intertwined. CREs are bound by TFs which interact with a promoter where transcription is ongoing. These CREs are all affected by, and themselves affect, histone modifications and chromatin states. Enhancer blockers, enhancers, silencers all fold and loop DNA, impacting at the level of chromatin as well as sequence and factor.

This theme of interaction and interdependence also touches on an important question in silencer biology, that of its role in the larger logic of cell regulation. One reason silencers may have been understudied for the last few decades is related to the enhancer model of eukaryotic regulation, in which gene regulation can almost entirely

be managed through the gain or loss of enhancing [151]. In this model, silencers are not frequently necessary as expression can be reduced through enhancer loss. Additionally, if enhancer loss is insufficient, heterochromatinization is available for repression. So what is the role of silencers in this picture? Clearly, given recent studies, silencers are prevalent and biologically relevant. One explanation is evolutionary in nature. It is much simpler to disrupt a function through a single mutation than to gain a novel function. So by mutation of silencer sequences, gene activation and corresponding gain of function could occur in evolution more rapidly than through creation of a novel enhancer [152]. Johnson et al. provide an example of this occurring in the *Drosophila yellow* gene [153]. Another is that silencers might provide more precise, specific, and signal-responsive regulation than heterochromatin, especially in regions where many genes are in close proximity and only one should be silenced. What determines which mechanisms of silencing are used and when, is not yet understood.

In Chapter 4 of this thesis, I investigate regulatory interactions for a specific region of the human genome, and place intra- and inter-CRE interactions within their chromatin, TAD, and gene regulatory unit context. I use this region as a foundation for studying the way these principles of regulation interact, and to begin to provide a model for one way that intra-TAD gene silencing can be mediated, by testing for both enhancer, silencer and enhancer blocker activity in elements in the region.

1.6 Conclusion and Overview

Many important questions remain regarding the mechanisms by which CREs mediate function, the ways CREs interact and coordinate amongst themselves and with the other layers of regulation in a cell, to ultimately determine gene expression in response to various signals and across cell types. In order to answer these questions, two crucial components are needed: robust and well-designed assays for CRE characterization, and genomic model systems that allow for the integration of information across the layers of regulation for a single region. In this thesis I address both of these components. In Chapter 2, I discuss the importance of plasmid-based reporter assays in CRE characterization and their utility in modeling CRE behavior. I detail a number of considerations for plasmid assay design, and address the issues and

complications that can arise from a lack of understanding of the intricacies of plasmid design. I present ways to address these issues using controls and design modifications. In Chapter 3, I address the importance of full-plasmid sequence validation for capturing potentially functional plasmid backbone variation and present a combined protocol and data analysis toolset designed to facilitate this validation which leverages novel long-read nanopore sequencing technology. In Chapter 4, I apply these principles of assay design and validation to the development of a modified assay panel for the characterization of both positive and negative regulatory CRE activity. I apply this assay to the identification of *cis*-regulatory elements in the *PRDM1-ATG5* regulatory domain in human cell lines, and leverage available genomic datasets to characterize these elements. Using these datasets, I develop a model for regulatory activity in this region that incorporates CRE activity, chromatin states, and TAD structure dynamics to account for differential regulatory activity which occurs within the region and between the cell lines. I also characterize combinatorial relationships between the CREs in the region and the HS2 enhancer, and dissect the structure of DHS16, the strongest enhancer tested and a candidate for regulation of *ATG5*.

Together, these chapters provide a framework for improved design of one of the primary tools used for CRE validation, present a method for rapid full-plasmid sequence validation which has the potential to improve the robustness of these tools, and apply these principles to characterize regulation in a region of the human genome, containing genes important to cellular function, in a way that integrates different models of regulatory control to provide a foundation for future studies of the domain.

CHAPTER II

Reporter Assay Design for the Study of *Cis*-Regulatory Elements

2.1 Abstract

Plasmid-based reporter assays were used to characterize some of the earliest known *cis*-regulatory elements (CRE) and continue to be important tools for the study of CRE activity today. Their usefulness and ubiquity is due to the ease and precision with which we can manipulate them *in vitro* using recombinant DNA technology, our ability to generate trillions of plasmid copies using *Escherichia coli (E. coli)* for replication, and the development of techniques for transferring plasmids into many different cell and tissue types. As our understanding of the diversity of *cis*-regulatory element types and activities has expanded, the number and complexity of plasmid-based reporter assays has increased in proportion. As this complexity increases, so has the potential for interactions of plasmid components with assay elements in ways that can interfere with interpretation of results. Below I discuss a number of considerations related to plasmid design and usage including: differences between plasmids and chromatinized DNA, the importance and elusiveness of non-functional sequence, proper design of controls, the effect of inter-element spacing, the potential for interactions between plasmid components, enhancer blocker usage in plasmid design, and the impacts of transfection on cell biology. I present an overview of each concept with data supporting or expanding on each consideration, discuss how it impacts interpretation of results, and suggest actions that can be taken to mitigate this impact.

2.2 The Power of Plasmids

The centrality of plasmids as a tool in molecular genetics is due in large part to three main factors. The first is the ease of producing large quantities of these recombinant plasmids cheaply using bacteria. In 1970, Mandel and Higa published a method for transformation of circular or linear phage DNA into *E. coli* which involved

making the *E. coli* chemically ‘competent’ to directly take up DNA [154]. In 1972, Cohen et al. used this method to achieve the uptake of plasmid DNA containing antibiotic resistance genes [155]. *E. coli*, like many bacteria, have a single circular chromosome and replicate extremely rapidly compared to most eukaryotic cells [156]. These factors combine to allow for efficient plasmid replication through leveraging existing bacterial biology. Recombinant plasmids are made to contain both a bacterial origin of replication and an antibiotic resistance gene in addition to their other components for functional testing in eukaryotic cells. Plasmids are transformed into bacterial cells which have been made chemically or electro-competent. The cells copy the plasmid alongside their own genome due to the origin of replication, and maintain it in the presence of antibiotic media which usually allows only cells carrying the plasmid antibiotic-resistance gene to survive. Through their rapid doubling times, a population of transformed bacteria can reach numbers in the range of 10^9-10^{10} cells overnight, with each cell carrying one or more plasmid copies [157]. These cells are then lysed and plasmid DNA isolated for downstream experiments. The advantage of this process is that it is significantly cheaper, easier, and maintains sequences with higher fidelity, than making a comparable amount of plasmid through *in vitro* processes, such as polymerase chain reactions (PCR).

The second factor is the ease with which recombinant plasmid DNA can be manipulated, even at the level of single nucleotides, to create desired products. Similar to the leveraging of bacterial biology to make plasmid factories, plasmid sequence manipulation *in vitro* is also largely based on innovative co-opting of existing biological enzymatic processes. Restriction enzymes, proteins which cut at sequence-specific sites, were originally isolated from bacteria that use them to target and fragment the genomes of bacteriophages, and are one of the earliest tools developed for molecular cloning [158]. Other cloning enzymes include ligases, phosphorylases, transposases and polymerases. An indicator of the ongoing importance of molecular cloning techniques is the continuing development of new methods for manipulation of DNA *in vitro*, such as Gateway cloning (which leverages the lambda phage integration and excision process) [159], Gibson assembly [160], and EMMA assembly [161].

The third factor contributing to the centrality of plasmids as a tool in molecular genetics is our ability to insert plasmids into target cell lines or tissues, using a variety of techniques where genes on the plasmids can be transcribed, translated and their function tested. Similarly to enzymatic techniques for plasmid cloning, a variety of methods have been developed for delivery of DNA into cell systems [162]. These can be categorized broadly into three approaches. Biological techniques co-opt viruses as vehicles to either insert DNA into the host genome or maintain it as a separate circular moiety. Using modified, usually fully or partially inactivated, viral sequences and packaging proteins, target DNA is delivered into cells as part of the viral genome. Chemical transfection relies on molecules including calcium phosphate, cationic lipids or polymers, which have positive charges and form complexes with negatively-charged plasmid DNA. The chemical-DNA complexes are then attracted to negatively charged cell membranes, allowing DNA to be delivered through the cell membrane and to the nucleus via mechanisms that are likely similar to endocytosis. Physical methods include direct injection, or laser-mediated poration or electroporation, which porate cell membranes allowing the DNA to enter the cell [163]. All of these methods have different advantages and drawbacks, in terms of safety, efficacy, skill, and equipment requirements, but the diversity of approaches allows researchers to choose an appropriate protocol based on the nucleic acid and cell type in use. This has allowed for the testing of plasmid constructs in a large number of cells, tissues, and organisms.

The factors listed above have contributed to the importance of plasmids as a tool in many fields of molecular biology, for many functions, including the expression and purification of proteins [164], gene therapies [165], vaccines [166], and modification of native DNA through delivery of TALEN [167], or CRISPR systems [168]. Here, however, I will focus primarily on their use in reporter assays for the testing of *cis*-regulatory elements in mammalian cell lines.

2.3 Principles of Reporter Assay Design

The first well-studied example of an orientation- and position- independent transcriptional enhancer, a 72bp repeat from the Simian Virus 40 (SV40) genome, was characterized using a circular vector containing recombinant DNA [44]. This plasmid

carried the SV40 repeats and a genomic fragment containing the rabbit hemoglobin Beta-Globin gene, and was transfected into human HeLa cells. Expression of the plasmid with or without the SV40 repeats was measured by use of a radioactive probe specific to the rabbit Beta-Globin mRNA showing that the SV40 ‘enhancer’ could increase transcription levels 200-fold. Since this first study, the basic principles of plasmid-based reporter assay design have remained largely unchanged, reflecting the power and utility of these assays in studying *cis*-regulatory elements.

While the core elements of plasmid-based reporter assay design are largely the same across most assays, these elements, like the tools used to create the plasmids themselves, have been rearranged, modified and expanded upon over the decades, producing a variety of different assays for studying various aspects of *cis*-regulatory activity. The basic components of a reporter assay are the reporter gene, whose expression is the readout, a promoter to drive expression of the reporter gene (usually a minimal, ubiquitous promoter - except in the case of promoter assays), and the DNA sequence that is inserted into the assay to test for its ability to alter expression of the reporter gene **(Figure 2.1)**. It is important to note that plasmids containing the reporter assay sequences will also carry a bacterial origin of replication, antibiotic resistance gene, clonal insertion sequences for test elements like restriction or Gateway sites, and the DNA sequences connecting these components and the assay components, all of which are collectively referred to as the plasmid ‘backbone’ and are typically considered inert (addressed in **section 2.4.2** below).

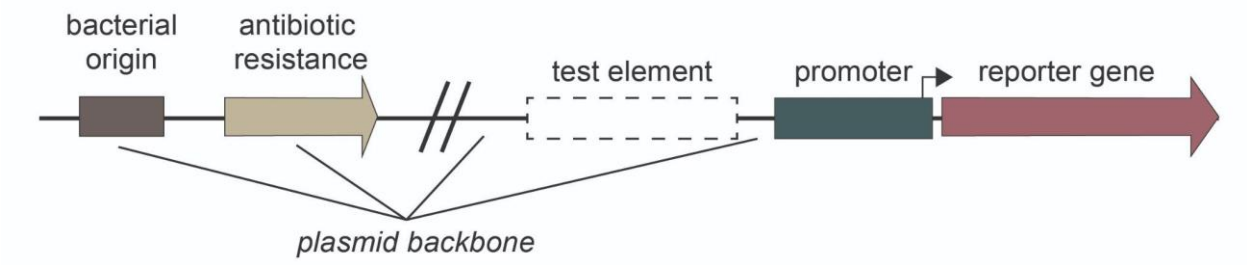


Figure 2.1 Reporter assay plasmid components

Many variations and expansions on this core design have been generated, and the differences largely depend on whether the assay is episomal or integrated (targeted or random insertion), constitutive or inducible, low- or high-throughput, applied to whole

organisms vs cell lines, cloning method used for test element insertion, and whether the readout is RNA, protein, or a functional assay(see reviews from [130] and [169] for more detail on specific reporter assay designs). Despite the diversity of their design, the similarities between all these assays are the core assumptions about element and assay function on which they rely, and which are important to understand in order to correctly interpret results from reporter assays.

The first of these assumptions is that the same basic molecular principles apply to element function in a reporter assay as in a native context (the primary exception to this - chromatin, is discussed below). As an example, the same sequence that acts as a transcription factor binding site (TFBS) in an enhancer's native site, can act as a binding site for that factor when the element is placed in the reporter assay (there are exceptions to this based on whether the TFBS in that enhancer are modular or not [49]). Fundamentally this assumption states that it is possible for us to obtain results from episomal reporter assays that are informative about native enhancer function.

The second principle is that of comparison. Interpretation of almost any reporter assay relies on the comparison of two conditions - one without the tested element (the control or ground state) and one with the tested element. Changes in expression or function are measured by comparison of one state to another. This makes the use of proper controls crucial to the correct interpretation of reporter assay results. Built into this principle of comparison, are two additional assumptions: first that the only difference generated between the control and the test plasmid(s) is the test element itself, and second that the change in expression seen in the test condition is due to the regulatory function of the test element, not a change in the relationship between the other components of the plasmid (the impact is active not passive).

2.4 Impact of Reporter Assay Design on Assumptions & Interpretation

While plasmid-based reporter assays have proved highly valuable in generating large amounts of functional data on many potential *cis*-regulatory elements, these assays are also prone to false positives and negatives, and issues with replicability. Not only do reporter assay results not always replicate native activity [124], results sometimes are not replicable even within the same cell type when validating high-

throughput results with low-throughput assays. For example, Pang and Snyder et al. [85] showed that only 4/5 elements which tested positive for silencer activity in their high-throughput assay replicated silencer activity in a low-throughput assay. Below, I address some potential causes for these issues with reporter assays, and discuss how they impact assay interpretation, how this relates to the key assumptions stated above, and ways these problems can be addressed. I will also present relevant data from my research demonstrating some of these potential issues. I examine the following:

2.4.1 The Impact of Context on Regulatory Element Activity

The primary caveat to the assumption of shared molecular function across reporter assay and native contexts relates to chromatin. In eukaryotic cells, DNA is packaged in chromatin made up of 147bp segments of DNA wrapped around nucleosomes, which are in turn made up of histones H2A, H2B, H3 and H4 [170], [171]. This chromatin itself acts as a form of regulation - DNA wrapped around nucleosomes is inaccessible to factors needed for activation of expression [172,173]. However, whether transient episomal constructs are chromatinized in a way similar to genomic DNA is not well-understood.

One study reported the presence of nucleosome-like particles on transiently transfected plasmids, as measured by micrococcal nuclease (MNase) digest and Southern blot, however the banding pattern they saw was ‘anomalous’ compared to the banding pattern of genomic DNA prepared similarly [174]. Another group used a 5.7kb plasmid as their model system for measuring the topology of nucleosomes in mammalian cells, indicating that plasmids are chromatin-competent. In this case, the plasmid was designed to be maintained throughout replication, and so was

chromatinized through the replication-dependent chromatin assembly process [175], however typical episomal assays do not use plasmids that are maintained through replication. More recent results support not only some form of chromatinization of plasmids, but also the functional impact of regulatory elements on plasmid nucleosomes, and demonstrate chromosome-like histone-mark spreading and the functional effects of nucleosome depletion in a plasmid [176].

If plasmids *are* differently chromatinized or non-chromatinized, how this would impact interpretation of reporter assays is also unknown. Given that it is known that chromatin-remodeling, histone-modifying, and histone-reading factors are recruited to enhancers during their activation [177], it seems reasonable to suppose that chromatin state could play a crucial role in CRE function. In that case, a lack of similar histone structure in plasmids could lead to differences in CRE function in an episomal context. Inoue et al. [124] attempted to address this question by testing 2236 candidate enhancer elements side-by side in either integrated (using lentivirus) or episomal (using a non-integrating mutated lentivirus) assays. Their results did show differences in activity between episomal and integrated activity of elements - the Spearman correlation for normalized RNA/DNA scores between replicates within the integrated or not integrated expression groups (0.944, 0.908 respectively) was higher than between the two groups (0.785). They also observed that data from integrated contexts more strongly correlated with genomic annotations indicative of activity. However, the presence of a strong correlation (0.785) between the integrated and non-integrated values from this test, can lead to a slightly different conclusion; while clearly there are differences in episomal vs integrated assays, and these must be accounted for, a majority of the time results from episomal assays *are* similar to those from integrated assays. This is especially surprising if the transiently transfected plasmids have significantly different chromatin structure.

While episomal assays should not be considered 100% reflective of native function, they can still be largely useful, especially for studies at scale, or where disruption of the native genome through random integrations is not desirable. Episomal assays (as discussed in Chapter 1) are tests of sufficiency, and these results support the importance of secondary functional tests, particularly disruption of elements in their

native context, to reinforce reporter assay results and provide data on necessity for gene expression. Additionally, a lack of heterochromatinization in plasmids may be advantageous. Within a given cell, a CRE sequence may be sufficient for function, but not accessible. However, when placed in an episomal assay context, it may be able to function. This allows for identification of not only the active CREs within a cell line, but of those poised for activity, perhaps on expression of a specific TF in response to cell signaling. Additionally, deletion of this element might not reveal a related change in gene expression and so would be missed by a necessity test.

The other aspect of context is cellular context - specifically, the suite of transcription factors expressed by a particular cell type that can drive cell-type CRE specificity [131,178]. A candidate sequence may be sufficient for regulatory activity broadly, but inactive in a particular cell line due to lack of its activating transcription factors. Additionally, there is evidence for enhancer-silencer elements with dual activity that are cell-context dependent, meaning that even classification of element type may be context-specific [179]. This can be addressed by testing elements across multiple cell lines and tissue types. However, as testing across all cell types is not usually feasible, reporter assays should be used to identify sequences capable of potential regulatory activity, not to eliminate sequences as potential elements, except within the context of that cell line. Whether an element is a CRE or not, as determined from reporter assay data, must always be discussed and considered within the framework of these aspects of context. A CRE is sufficient for regulatory (silencer, enhancer) activity, in a specific cell type or types, given it is in accessible chromatin and the requisite protein factors necessary for function are present within that cell type.

2.4.2 The Role of Plasmid Backbone Elements in Contributing to Expression

Plasmid sequences which include bacterial sequences or inter-element DNA are collectively considered ‘backbone’ sequences. These sequences are often treated as non-functional and not relevant to assay design, as opposed to regulatory and gene sequences which are carefully checked for sequence fidelity. As a result, there can be variability in the presence and types of these sequences across different plasmids, depending on which backbone is used, and on the sequences being shuffled in and out

attached to the ends of functional elements during cloning. However, these sequences can be functional and influence expression through sequence context.

An example of a backbone introducing novel error comes from the cloning of an SV40 promoter upstream of a gene in one of our plasmids, which introduced an ATG that combined with downstream sequences to create a complete truncated open reading frame with a stop codon just upstream of, and in a different reading frame from, the gene **(Figure 2.2a)**. Due to prioritization of the upstream reading frame during translation and its overlap with the start of the gene's reading frame, this error reduced expression to the same level as a promoterless plasmid **(Figure 2.2b)**. Removal of this upstream ATG completely restored expression to levels similar to that of another SV40 promoter plasmid **(Figure 2.2c)**.

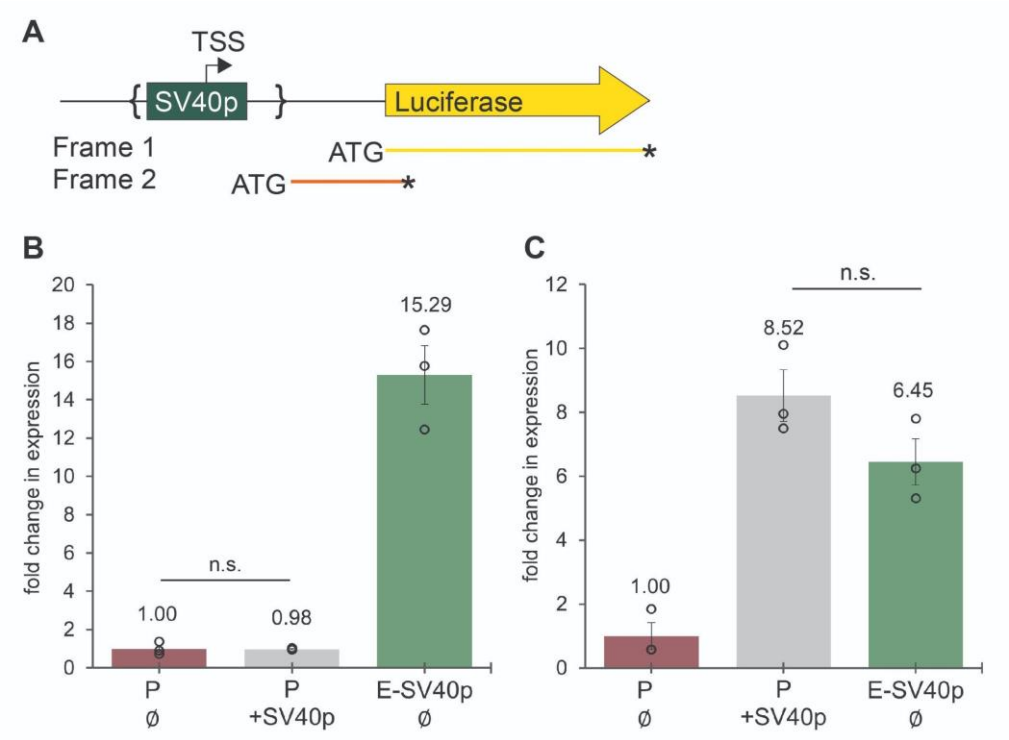


Figure 2.2 Impact of upstream alternate reading frame
a. Diagram showing upstream truncated reading frame: truncated frame (orange line), correct reading frame (yellow). { } show SV40 promoter insertion site. * are stop codons. **b and c.** Fold change in luciferase activity of plasmid with (**b**) or without (**c**) truncated reading frame (gray bar). P ∅: promoter assay with no insert. P +SV40p: promoter assay with SV40 promoter inserted into cloning site. E ∅: enhancer assay with SV40 promoter, no enhancer insert (and no reading frame error). Three biological replicates (open circles), error bars are standard error, t-test * = $p<0.05$, $p\geq0.05$ = n.s.

All expression data presented in this chapter were generated using electroporation of Firefly luciferase-expressing plasmids and Renilla co-transfection controls in the K562 myelogenous leukemia cell line, in at least three biological replicates (see section 2.6 for further details on transfections and analysis).

This error was not initially detected as the contributing sequences were located entirely in the plasmid backbone, and in a different frame, so the gene's reading frame remained intact. The detection of this error was made possible due to the use of matched controls during assay testing. The SV40 promoter was inserted as a control (P +SV40p) for a promoter testing assay (P ∅), and was compared to expression of an enhancer assay that contained a fixed SV40 promoter in a similar position (E ∅), but did not contain the reading frame error. While errors like this are actively selected against in the genome due to their disruptive nature [180], they are relatively easy to accidentally introduce during molecular cloning. Without the use of appropriate controls, a lack of expression when testing other promoters could have been attributed to a lack of promoter function in this cell line rather than a failure of the assay.

In another case, the 5' untranslated region (UTR) sequence just upstream of the reporter gene in a plasmid contributed strongly to expression, as determined by the impact of deleting this element. Deletion of the 160bp just upstream of the luciferase gene (not including the Kozak sequence) roughly halves expression **(Figure 2.3)** compared to a plasmid without this deletion.

This result is consistent with what is known of 5'UTRs in genomic contexts, as they can regulate translational efficiency [181]. Another possible interpretation for this result is that this region is acting as a proximal enhancer (at the level of transcription initiation). In this plasmid, the 5'UTR region contained a variable 20bp barcode sequence used for identification in a plasmid library, as has been used for high-throughput regulatory assays (massively parallel reporter assays or MPRAs). However this result indicated the possibility that the variation of the barcode sequences themselves, rather than the tested element, could contribute to changes in expression. This would negate one of our basic assumptions - that changes in expression are due to insertion of the test element only.

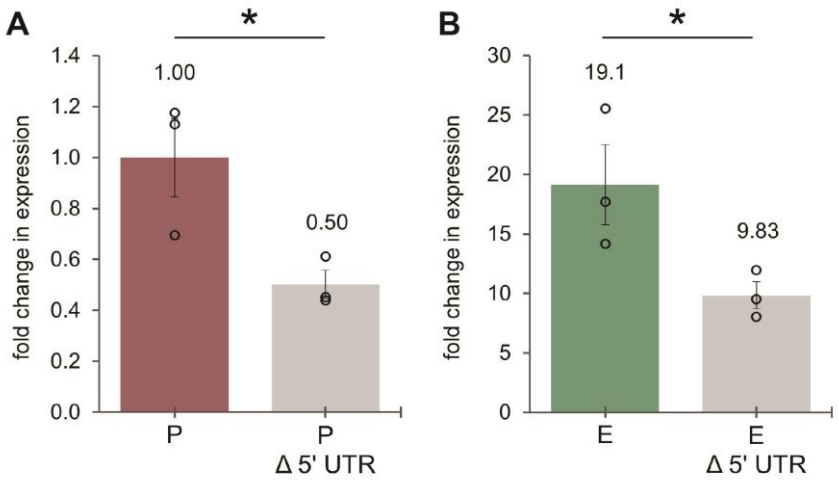


Figure 2.3 Impact of 5'UTR deletion on expression
Fold change in luciferase activity of **a.** promoter control plasmid (**P**) and **b.** enhancer control plasmid (**E**) vs versions of each plasmid with deletion of their 5'UTR (**Δ 5'UTR**). Three biological replicates (open circles), error bars show standard error, t-test * = $p<0.05$, $p\geq0.05$ = n.s.

To prevent this, the barcode region could be moved to the 3'UTR, where it is located in many current MPRAs such as STARR-seq [122]. However, results from various labs have found sequence-specific effects of barcodes on reporter expression [182]. A recent paper from the Shendure lab systematically tested the impact of barcode and test element positional variation and found that placement of test elements and barcodes in the 3'UTR led to much lower replicability (mean r=0.54 vs 0.9 between replicates) as opposed to 5'UTR placements [183]. They also found enrichment in their 3'UTR-barcoded MPRA results for features correlated with RNA stability, not enhancer function. This supports the idea that 3' placement leads to prioritization of elements that impact functions related to 3'UTR impact on transcript stability, rather than independent, pre-transcription CRE activity. This issue can be addressed in part by the use of multiple barcodes per test element, to allow for averaging of an element's measurement across contexts with multiple differing barcodes to control for their individual effects [184].

In contrast to the strong impact of deleting the small 160bp 5'UTR region of a plasmid, in a similar test, deletion of a large 1.5kb region, upstream of the promoter, resulted in no significant change in expression **(Figure 2.4)**. While a deletion this large (1.5kb of a 6.8kb plasmid) was expected to impact expression, it did not, indicating that

changes from altering plasmid sequence can be highly sequence- and function-dependent.

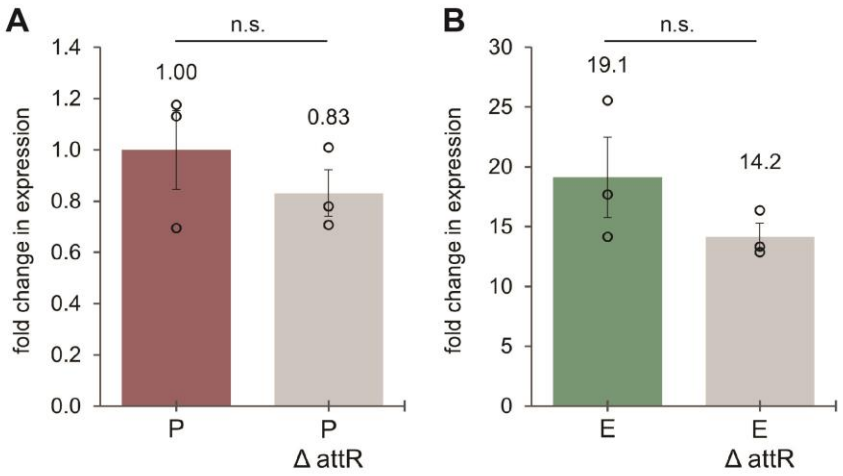


Figure 2.4 Impact of Gateway site on deletion
Fold change in luciferase activity of **a.** promoter control plasmid (**P**) and **b.** enhancer control plasmid (**E**) vs versions of each plasmid with deletion of the attR Gateway insertion site and intervening CmR sequence cassettes (**Δ attR**). Three biological replicates (open circles), error bars show standard error, t-test * = $p<0.05$, $p\geq0.05$ = n.s.

As a final example of the effect of plasmid backbone sequences that are assumed to be 'non-functional', Muerdter et al. found that transcripts on the STARR-seq reporter assay were primarily originating from the bacterial origin of replication (f1 ori) and not the intended core promoter [185]. The bacterial origin was acting as a highly competitive promoter, in a mammalian system. While the group provided a solution to this issue by co-opting the origin as their core promoter, this is not a long-term solution, as it would mean only testing enhancers for activity using a single, non-mammalian promoter. Additionally, the bacterial origin is not typically sequenced or validated following cloning, since a loss of function would lead to non-replication of the plasmids during bacterial growth. Thus it is assumed that successful plasmid growth implies an intact origin sequence. However, should mutations occur that impact its promoter, but not its origin, function, or occur during cloning steps subsequent to bacterial growth, these would not typically be captured by standard validation methods and could impact expression.

All of the above results indicate the potential functional consequences and complexity of non-coding backbone plasmid sequence. This is particularly important due to these regions being typically considered non-functional, leading to the potential for alterations to the sequences propagating from lab-to-lab as components are cloned into plasmids and sequences are not fully validated. This can result in variability between plasmids, even those listed in publications as using the same backbone sequence, depending on the origin of the plasmid. A solution to address this is to make sequencing validation of entire plasmids, not just key functional regions, standard (see Chapter 3) and to standardize publishing full plasmid sequences as a part of methods in publications. The larger solution to the issue of backbone complexity and its potential confounding effects on reporter assay results is, first, a broader understanding of the potential impacts of these sequences. Second, the application of appropriate testing when modifying 'non-functional' plasmid regions to determine functional impact, and most importantly, the use of appropriate positive and negative controls, as discussed below.

2.4.3 Controls and 'Non-Functional' Sequence

The use of proper controls in regulatory assays is crucial for interpretation of results, as interpretation relies on comparison as discussed above. In order to discuss this more generally, I will define positive controls as controls that have known function matching that of the type of element being tested (so for an enhancer, a positive control will increase expression, but a silencer or enhancer blocker positive control will decrease function) and negative controls as sequences that have no function in the assay context.

Negative controls allow for comparison to occur as they are the base state against which increases or decreases in function are measured. One choice for a negative control in low-throughput assays is the use of the reporter assay plasmid, **'as-is'**, prior to insertion of a test element through cloning. This seems at first to be the most simple form of control, as the only change should be due to the presence of the test element. However, it is entirely possible that this element can be placed into the middle of backbone plasmid sequence and act as a cryptic enhancer, or provide sequence

context that increases TF binding strength adjacent to the insertion site. This could lead to apparent silencer activity of an actually neutral element, as insertion would reduce expression, or non-activity of a weak enhancer if the enhancer activity and the loss due to insertion are balanced. (see 2.4.4 below for a special case of this related to enhancer-promoter spacing). In one case, we see a false positive for silencer function, and in the other a false negative for enhancer function. The possibility for this type of issue is strongly supported by the importance of surrounding sequence context on CRE function. This was demonstrated by Klein et al., who tested putative 651 enhancers either using a minimal core sequence of 192bp, or larger regions of 354bp and 678bp, containing more of the elements' genomic context, in an MPRA. They found low correlation between expression values (normalized RNA/DNA) for the long vs short versions of the same element (r=0.53, vs r=0.94 between replicates of the same set) [183].

This impact can also depend on the type of cloning site used. In restriction digest methods, disruption of surrounding sequence is minimal. However another commonly used method, Gateway cloning, which is useful for the ability to rapidly transfer one test sequence into multiple assays or sites, element insertion changes flanking sequence. In **Figure 2.5**, the potential functional impact of this process is demonstrated by comparing expression from a plasmid prior to, and after, conversion of attR sites to their corresponding attB sequences as a result of the Gateway reaction insertion/excision process. In this case, two things are occurring; attR to attB site sequence conversion, and swapping of the DNA inside the attR sites (containing 1.3 kb of sequence including a bacterial selection marker under control of a bacterial promoter) for the DNA between the attB sites (a 50bp sequence from a 'minimal' entry vector) **(Figure 2.5a)**. Since both of these have different expression **(Figure 2.5b)**, it is difficult to determine which is the appropriate 'baseline' reporter assay negative control. The attR version is the assay in its original form, however the attB version is more similar to the state of the plasmid after insertion of a test element. Additionally, the attR version contains bacterial functional elements (selection marker and promoter) and given that some bacterial sequences may be functional in mammalian cells [185], these could impact expression. However, as seen in **Figure 2.4**, deletion of the entire attR cassette resulted in no

significant change in expression in another plasmid backbone (P vs P ΔattR and E vs E ΔattR), indicating that it could be a better baseline control, being seemingly non-functional.

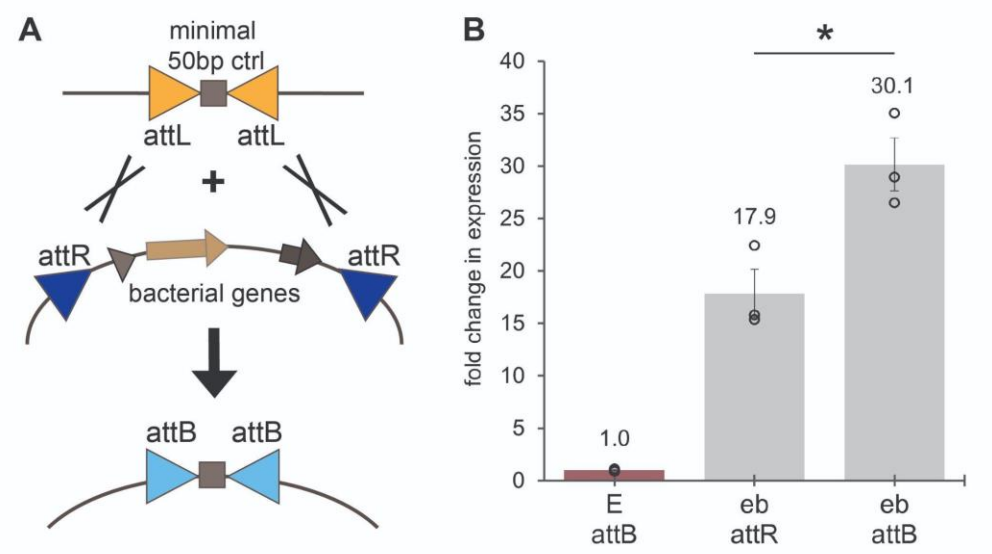


Figure 2.5 Comparison of alternate Gateway site controls
a. Diagram showing Gateway reaction and inserts used in these plasmids. **b.** Fold change in luciferase activity of enhancer blocker (**eb**) assay before Gateway reaction (**attR**) or after Gateway reaction to inert 50bp 'minimal' entry plasmid (**attB**). **eb** plasmids contain, in order: [HS2 enhancer-Gateway site-SV40 promoter-Luciferase]. Values are normalized to the **E** plasmid (attB-SV40 promoter-Luciferase). Three biological replicates (open circles), error bars show standard error, t-test * = $p<0.05$, $p\geq0.05$ = n.s.

An alternative to this type of 'as-is' negative control is the use of **'neutral' or 'non-functional' sequences** as controls. These controls have been used in both low-throughput assays [107], and high-throughput assays [183].They are inserted in the same manner as a test element, and the assay carrying the neutral element insertion is used as the control. This version better controls for passive impacts of sequence insertion into the cloning site. The drawback to this approach is the difficulty in finding 'non-functional' sequences. Given the small size of TF binding motifs (6-12bp [178]), it is hard to find a 200bp DNA sequence without potential TF binding sites. Even commonly used 'scrambled' controls, where a test sequence is randomly rearranged [186], could contain new TF motifs. Many of these might be for TFs not expressed in the chosen cell type. However, this raises another issue - due to the cell-type specificity of

CREs, a sequence may be neutral in one cell type, but not another, and so there might not be universal neutral sequences.

Neutral sequence should be determined on a case-by-case basis by testing for functionality in the assay context. **Figure 2.6** shows functional testing of two types of neutral sequence I designed for use here and in my own CRE assay (see Chapter 4). r33 is a 750bp sequence from a set of 50 randomly generated sequences, chosen for its lack of low complexity regions, and relatively low number of potential TF motifs relative to the other 49 sequences. Lbda is a 754bp sequence taken from the reverse strand of an intron from a Lambda phage gene. Neither control shares significant homology with human genomic sequences and both have ~50% GC content. Both were inserted in three different assay contexts by restriction digest, tested, and compared to the original plasmids with no insert (ø).

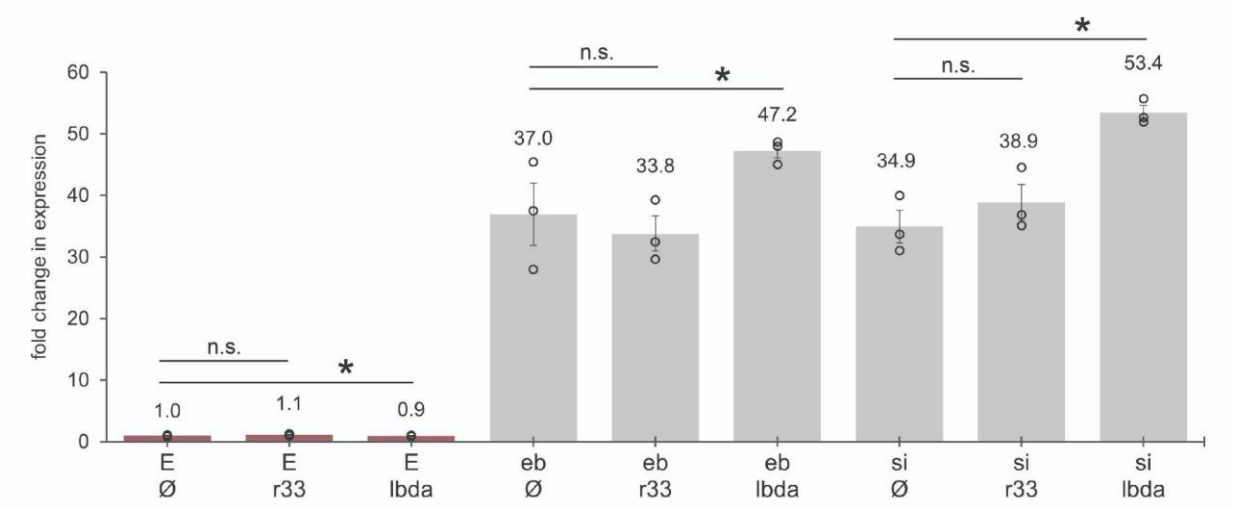


Figure 2.6 Impact of non-functional control sequences on expression
Fold change in luciferase activity relative to E-ø control. **ø:** no control sequence inserted, **r33**: random negative control inserted, **lbda**: lambda negative control inserted. In three different assays - **E**: enhancer assay (control-SV40p), **eb**: enhancer blocker assay (HS2e-control-SV40p), **si**: silencer assay (control-HS2e-SV40p). Three biological replicates (open circles), error bars show standard error, t-test * = $p<0.05$, $p\geq0.05$ = n.s.

The addition of the r33 control did not result in a change in expression relative to the original plasmid in any assay. The addition of the lbda spacer did increase expression in two cases and decrease it in another. This would indicate that the r33 could be an excellent negative control within this assay context in this cell type (K562 cells). The three assays used here are arranged following the same element order throughout this chapter except where otherwise indicated: Promoter Assay (P) = insert-

gene, Enhancer Assay (E) = insert-SV40p-gene, Silencer Assay (SI) = insert-HS2e-SV40p-gene, Enhancer Blocker Assay (EB) = HS2e-insert-SV40p-gene. Lowercase e as in HS2e and lowercase p as in SV40p stand for enhancer and promoter respectively.

Similar approaches have been used to generate non-functional negative controls in high-throughput reporter assays, on a larger scale. One MPRA in HepG2 cells [183] used 100 negative controls based on evidence of their lack of function in a previous MPRA in the same cell type [187]. Not every element might be non-functional but the majority should, and the use of many controls can allow for estimation of the range of variation from largely neutral sequences to help determine rates of false positives.

A caveat to the non-functional sequence approach is that when testing neutral sequence functionally, one faces the same potential confounding factors of initial insertion impact and necessary comparison to the assay plasmid pre-insertion as mentioned above for the 'as-is' assay control. While this is a concern, it cannot be taken to the extreme that therefore no controls can ever be relied upon or useful data generated. Negative controls are useful when designed with consideration for the assay and cloning method, tested functionally, and interpreted appropriately. This caveat is also reason to combine regulatory assay data with other lines of evidence for function, such as genomic deletion assays and transcription factor binding, in order to reinforce evidence for function, as discussed in Chapter 1.

Positive controls are important both in establishing functionality of the assay and to provide biologically significant context for signals. Positive controls help to avoid obvious false negative results due to a malfunctioning assay. Demonstrating the assay is functioning using a known functional element can support a lack of signal in an assay as originating from a lack of function for a given test element, rather than a failure of another part of the assay. They also provide important biological context and allow for better comparison across datasets and publications. A 0.2-fold or a 10-fold increase could be significant or not, depending on the assay and readout. This is hard to determine without comparison to the expression of a known, biologically relevant element. While there may be variation in the baseline expression from one reporter assay construct to another, some useful comparisons can be made if both use the same

control enhancer, or silencer, to benchmark expression. Some important considerations for positive control elements are that they be functional in the chosen cell type and compatible with other CREs in the assay.

2.4.4 The Impact of Element Spacing on Function

Enhancers are considered to be position- and orientation-independent, as supported by early studies in enhancer biology [44,63,188]. This also seems to hold true for some silencers [189], but not others [190]. However conclusions from these studies focused largely on whether elements were or were not position-independent, not to what extent fine gradations in distance could impact functional levels. Looking at the details of results from these papers, differences *can* be detected. In one paper, moving an HS2 enhancer 5.8kb upstream in a plasmid increased expression 20% compared to when it was placed adjacent to the assay's promoter [63]. The possibility of this positional effect, and its relevance to genomic and episomal contexts was supported in a recent paper by Zuin et al [138]. By leveraging the random nature of transposon insertions, they were able to create 264 cell lines, each with one enhancer inserted into a different position in a fairly even distribution across a 560kb topologically associated domain (TAD) containing a central eGFP gene. They found that increasing enhancer-promoter distance decreased eGFP expression, up to 10 fold. Another paper found a 50% decrease in expression on moving an enhancer from 6.5kb to 9kb away from its promoter in *Drosophila* that was mitigated by placing enhancer blockers [191].

Should this effect of relative CRE-promoter distance also hold in a reporter assay, this could create an exception to our assumption that changes in reporter assay expression can be attributed solely to test element function (rather than a passive effect of insertion). Any assay that requires cloning of an element between other elements, depending on the insertion size and cloning method, could alter the distance of said elements. While strongly active CREs might overcome any effect of distance-mediated change in expression, weaker ones might be missed depending on their strength and direction of effect, relative to that of a positional effect. Both examples above are tested on a multi-kb scale, but the size of tested elements in reporter assays usually ranges

from 100bp-1kb. Below, I present data supporting the potential for an impact of enhancer-promoter distance on expression at reporter-assay scale distances.

I initially observed a possible distance-dependent expression effect in my data while testing differences in expression across plasmid controls. **Figure 2.7a** expands on **Figure 2.5** to show the impact of attR vs attB sites in silencer vs enhancer blocker assays. When comparing these two assays, an unexpected effect was revealed - an opposite direction of effect of the impact of the Gateway reaction (attR → attB). One explanation for this could be the difference in size of the attR vs attB controls and the influence that has on the distance of the enhancer from the promoter. In addition to having different sequence content, the sizes of the intervening sequences are very different for the attR vs attB controls (1363bp vs 102bp inclusive of the Gateway sites, respectively) **(Figure 2.7b)**.

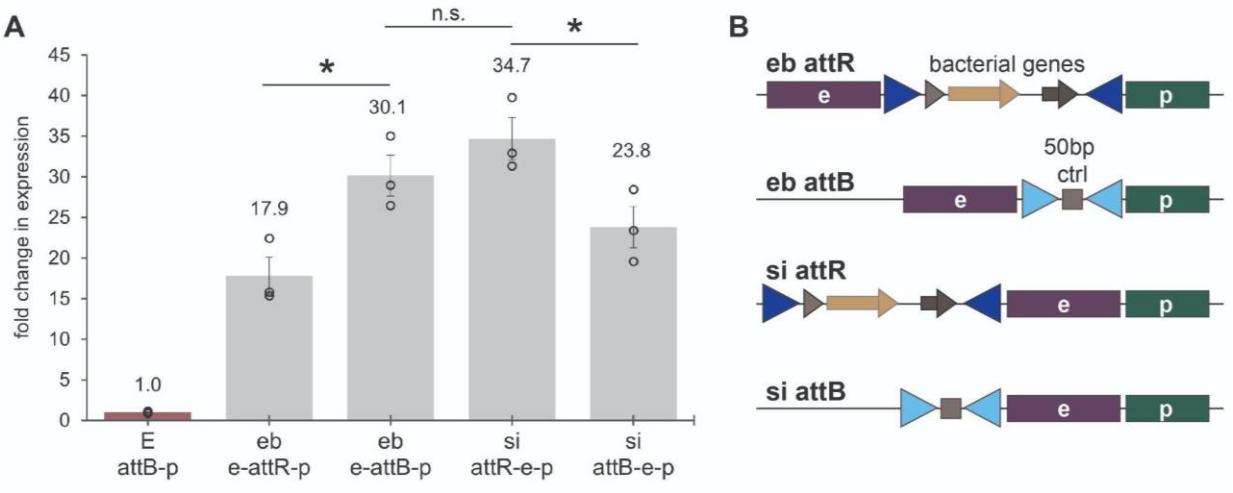


Figure 2.7 Positional effect of two Gateway site controls
a. Fold change in luciferase activity for silencer (**si**) or enhancer blocker (**eb**) reporter assays with either no Gateway insert (**attR**), or a control minimal sequence inserted (**attB**). Fold change is over enhancer control plasmid (**E**). **b.** Diagram showing plasmid layouts. 'e' is HS2 enhancer, 'p' is SV40 promoter, dark blue triangles are attR sites, light blue are attB. Three biological replicates (open circles), error bars show standard error, t-test * = $p<0.05$, $p\geq0.05$ = n.s.

As a result, in the case of the enhancer blocker (eb) assay, the attB conversion moves the enhancer closer to the promoter. This does not, however, account for the impact observed in the silencer assay, where there is an effect in the opposite direction. It is also not possible to distinguish whether this effect is due to the passive effect of enhancer and promoter distance, or some cryptic functional element in either the attR/B

or intervening sequences. Another explanation could be the presence of a cryptic weak enhancer blocker in the sequence between the attR sites which reduces enhancer-promoter (e-p) communication in the eb position and increases it in the silencer position by enforcing directionality of e-p communication in the optimal forward direction **(more on this in section 2.4.5)**.

In order to interrogate the possibility of an effect of enhancer-promoter distance, I designed a test specifically to test for the impact of element spacing, while maintaining identical sequence content, to exclude confounding differences from the attR/B example. **Figure 2.8** shows results from testing this assay as well as a diagram of the assay design. In the assay, the enhancer and promoter are initially separated by the 750bp r33 neutral sequence tested in **Figure 2.7**, acting as a 'spacer.' r33 is split into three 250bp fragments, such that the enhancer can be cloned closer to the promoter in 250bp increments, while maintaining the same overall sequence context (see diagram). Each 250bp spacer fragment is also tested independently with the enhancer and promoter to determine whether the separate elements show different independent activity levels (white bars in **Figure 2.7**). There is a difference between the expression level of the A and B 250bp spacer fragments (e-A-p vs e-B-p; t-test p=0.024), but it is not strong enough to account for the observed full change in expression. Surprisingly, results from this test do show an almost 2-fold increase in luciferase activity that correlates with the enhancer moving closer to the promoter. Interestingly, this jump seems to be discrete, not continuous, as the 250bp and 10bp conditions are roughly the same, as are the 500bp and 750bp conditions, in terms of expression (the differences in expression are not statistically significant).

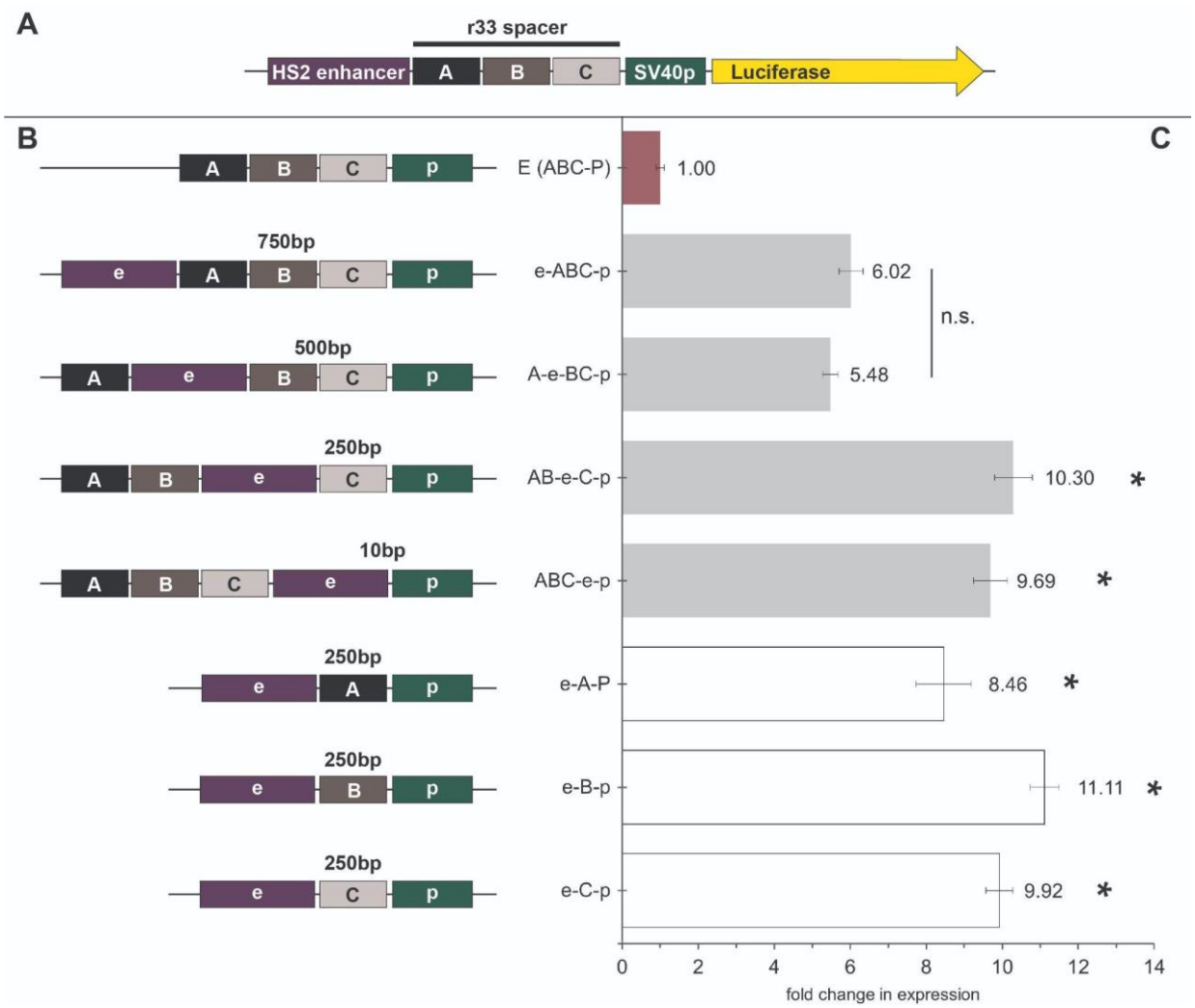


Figure 2.8 Impact of enhancer-promoter spacing on expression
a. Diagram of assay at top. A, B, C are 250bp fragments of 'neutral' spacer sequence r33. **b.** Simplified plasmid diagrams listed adjacent to their corresponding expression values (panel **c**). Enhancer-promoter distance is listed above each diagram. **e**=enhancer, **p**=promoter. Fold change measured to promoter-only control (red bar). t-test * = $p<0.05$, $p\geq0.05$ = n.s. vs condition e-ABC-p.

Additional support for the idea of an e-p distance effect in a plasmid context comes from Davis et al. In their paper, Davis et al. used the scale of an MPRA to dissect and quantify the sequence context, distance, and site number dependencies of transcription factor activity at a very fine level in both episomal and genomic contexts [137]. They found that as a pair of TF binding sites (c-AMP response elements, or CRE sites) at a fixed 10bp distance from each other is moved further from a minimal promoter while maintaining the same sequence context, transcription levels decrease with distance, with a roughly 2.2-fold drop-off in signal once the elements reach about

147-176bp from the promoter. This is consistent with previous work on CRE distance-specificity in a genomic context [192]. This is also very close to the roughly 2-fold drop-off at a 250bp distance in my data shown in **Figure 2.8**.

While this distance effect would not be relevant for design of promoter, and some enhancer, regulatory assays, especially high-throughput assays which size-match elements and controls due to oligo synthesis limitations, it could impact others. In enhancer blocker assays, where test elements must be inserted between an enhancer and promoter, variable fragment size relative to controls could lead to false positives or negatives. A larger element may reduce activity due to increasing e-p distance, rather than enhancer blocker function creating a false positive result. A false negative could be caused by the decreased distance balancing out the effect of a smaller, weak enhancer blocker. Alternatively, in enhancer or silencer assays, it is possible that testing much larger sequences of 1kb compared to 200bp fragments would place the core functional unit of an enhancer or silencer further from the promoter, creating this effect. In either of these cases, setting a standard for all reporter assays of using uniform test element sizes with size-matched controls would control for this distance effect.

Understanding the underlying mechanism for this effect not only has implications for aspects of reporter assay design, but could also reveal useful information about the mechanisms of enhancer (and silencer) action. A key aspect of the molecular mechanism of enhancer activity is the need for physical proximity with the target promoter, facilitated by DNA looping [62]. However, possibly for enhancers directly adjacent to a promoter, looping of DNA is not necessary as bound TFs are in close enough contact to activate transcription. The distance at which expression drops off in a relatively binary fashion around 150-250bp, described in results shown here and in the Davis paper, could represent the distance at which proximity is insufficient and looping becomes necessary for contact to continue [137]. It could follow that the corresponding decrease in expression is due to looping (at least in a plasmid context) being less efficient for activation than permanent proximity of an enhancer and promoter. In support of this, a study using an *in vitro* model system to study *E. coli* regulatory elements found that at 110bp distance, enhancer-promoter communication was efficient

regardless of plasmid state, but that at 2.5kb supercoiling of plasmid DNA greatly increased expression over relaxed DNA, presumably due to a mechanism similar to looping where the enhancer and promoter are brought into proximity [193]. A continuous decrease in expression with increasing distance is unexpected given that abundant evidence points to successful enhancer function even at great distances from promoters in the genome. This distance is also coincidentally consistent with the estimated ~200bp size of TF-binding promoter-proximal regulatory regions [35]. This suggests that enhancer-promoter spacing in episomal assays does seem to accurately model genomic behavior and provides support for plasmid assays' usefulness in studying CRE activity.

2.4.5 Inter- and Intra-Plasmid Interactions

In addition to having individual activities that are context- and position-dependent, CREs in reporter assays can have activities that arise from and are dependent on the interactions of all the elements. While reporter assays can be designed very simply, the simplest being a promoter assay (a gene and a test element), a number of more complex assays have also been constructed. These complex designs have allowed researchers to manipulate various aspects of regulation, like timing of location of expression [194,195]. They can allow for co-expression of multiple genes [196], which greatly increases the possible design intricacy [197]. However, as more elements are added, the possibility of complex, potentially confounding, interactions between assay components increases. Below I present an example of one of these effects from my data **(Figure 2.9)** and discuss and test two possible cases of complex interactions that could cause such an effect. In **section 2.4.6** I discuss enhancer blockers and their role in limiting these effects in reporter assays as well as in other plasmid-based systems.

Figure 2.9 shows two unexpected, concerning effects in our plasmid controls. First, the P-∅ plasmid lacks a promoter upstream of the luciferase gene, yet has significant expression (>5x) over background signal (cells only - in all other figures this was subtracted from signal during normalization and is not shown), and almost twice the activity of a plasmid from a different background context with an SV40 promoter (pGL3-

SV40p). Second, the SV40p in this assay context (E ∅-SV40p) is 28x higher than the SV40p in the pGL3 context. While there can be variation between different plasmid contexts, this does not seem likely to cause a difference this large, nor should it cause expression from a promoterless gene.

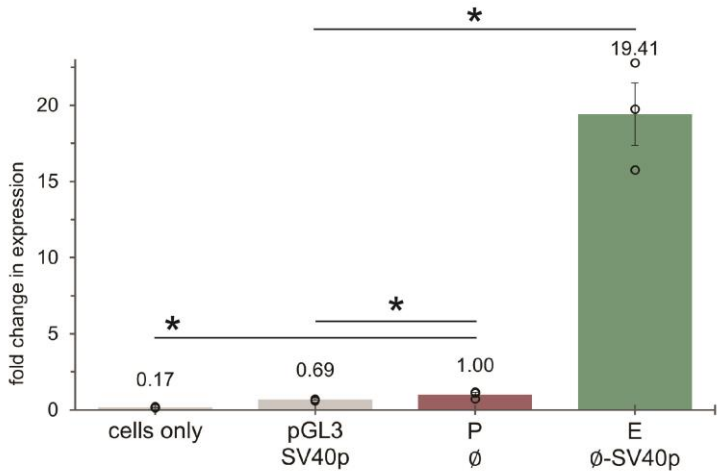


Figure 2.9 Expression in a promoterless cassette
Fold change in luciferase activity over '**P ∅**' control. **P=** plasmid with no test promoter inserted (red), **E =** plasmid with no test enhancer inserted upstream of SV40 promoter. **pGL3-SV40p** is pGL3 vector with SV40 promoter (different plasmid backbone than P/E). '**Cells only**' shows signal from non-transfected cells. Three biological replicates (open circles), error bars show standard error, t-test * = $p<0.05$, $p\geq0.05$ = n.s.

We first considered a possible cause of this expression could be due to *trans*-plasmid interactions between co-transfected plasmids within the same cell. The data shown here originates from cells co-transfected with a plasmid containing the CRE assay, which uses Firefly luciferase as the readout, and a control plasmid bearing the SV40 enhancer and promoter, driving Renilla luciferase expression. Renilla luciferase expression is used to normalize for variation due to cell death and electroporation efficiency (see Methods). We wished to determine whether it was possible for the SV40 enhancer/promoter on the Renilla plasmid to impact expression on the Firefly plasmid. This seemed possible, given that during bulk electroporation or transfection of a cell population, there is not an even distribution of one plasmid entering one cell [198]. Additionally, in a chromatin context, *trans*-chromosomal regulation (transvection) is an established phenomenon [199], indicating that physically separate pieces of DNA can cross-regulate.

In order to test for this possibility, a co-transfection test was performed (**Figure 2.10**). A minimal plasmid was used, that carries only necessary bacterial elements and a cloning site (pX) and has no luciferase genes. The strong, ubiquitous enhancer human cytomegalovirus (CMV) enhancer (that is active in K562 cells [200]) was inserted into the cloning site to create pX-CMVe. pX with (pX-CMVe) or without the CMV enhancer (pX-∅) was co-transfected alongside a plasmid containing Firefly luciferase under the control of an SV40 promoter (E-∅), or a CMVe-SV40p-Firefly plasmid (E-CMVe).

pX-∅ and pX-CMVe when transfected individually showed no expression as expected. pX-∅ was used as a control for any impact that co-transfection itself might have on expression (for instance competing for cellular uptake with its co-transfected partner). As shown in **Figure 2.10**, these co-transfected plasmids do not seem to be capable of *trans*-regulation. Co-transfected pX-CMVe does not increase expression of E-∅, and while it does seem to slightly increase expression of E-CMVe, this is be due to a slight decrease in the pX-∅ control, and the increase is not nearly the 28x observed in **Figure 2.9**. This suggests that while co-transfected cells can contain multiple plasmids, either cross-regulation does not occur, perhaps due to plasmids being separated in the relatively large nuclear space, or that if it does occur, it is not at a frequency significant enough to impact population-level readouts.

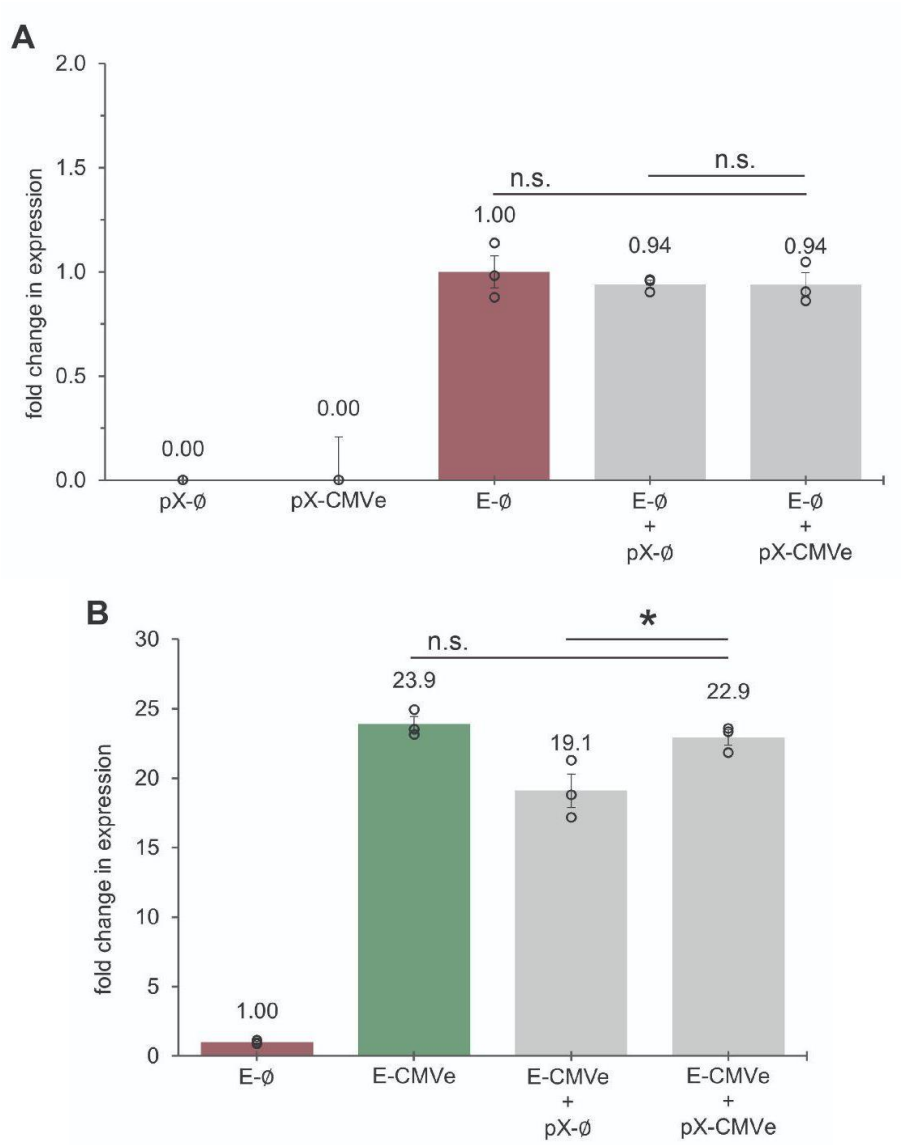


Figure 2.10 Test for inter-plasmid activity using co-transfection
Co-transfection of **a.** SV40 promoter-only **(E-Ø - red)** or **b.** CMV enhancer-SV40 promoter **(E-CMVe - green)** -containing Enhancer Assay plasmids **(E),** and another plasmid, pX, without **(pX-Ø)** or with a CMVe enhancer **(pX-CMVe)**. pX is a minimal plasmid containing no other regulatory elements and no luciferase gene. + indicates a 1:1 copy-number co-transfection of the plasmids listed above and below the + sign. Normalized to E-Ø in both charts. Three biological replicates (open circles), error bars show standard error, t-test * = $p<0.05$, $p\geq0.05$ = n.s.

Having established that inter-plasmid interactions are an unlikely cause of the effect shown in **Figure 2.9**, we next looked at intra-plasmid interactions. The data in **Figure 2.9** is derived from plasmids which were designed as a complex multi-cassette expression system used for testing of positive and negative CREs in a high-throughput assay. These plasmids contain a secondary expression cassette (defined as a set of

CREs immediately upstream of a gene, the gene, and its polyA signal) upstream of the luciferase expression cassette where test elements are inserted (see diagram in **Figure 2.11a)**. We decided to investigate whether the upstream CREs were impacting downstream expression. I tested this possibility by generating a series of plasmids with deletions of the promoter, enhancer or both, from the upstream cassette in both the promoter (P) and enhancer (E) assay plasmids and measured luciferase expression.

Deletion of the upstream SV40 promoter in the P construct **(Figure 2.11b)** did not significantly decrease expression from the downstream promoterless luciferase. Deletion of the upstream CMV enhancer, or upstream CMVe and SV40p, however, ablated expression to levels not significantly higher than background signal. In the E assay (**Figure 2.11c**), deletion of the upstream CMVe or upstream CMVe+SV40p reduced expression levels ~20x, bringing the signal much closer to the expected range, as compared to the pGL3-SV40p vector (data not shown, 0.55x fold change in this experiment). These results strongly support intra-plasmid activity as the cause of the unexpected expression patterns in these plasmids.

As both cassettes are needed for proper functioning of this assay, we next wanted to determine the potential mechanism behind this intra-plasmid activity in order to determine how to best address it. Two potential mechanisms are represented in Figure 2.11a. The first (red dotted line) is read-through of RNA polymerase II and transcription machinery, where transcription does not stop at the polyA signal, and continues downstream through the luciferase gene [201]. Promoters are known to be able to drive expression of multiple genes in a single plasmid [202]. In this model, deletion of the upstream SV40p would prevent loading of transcription machinery and prevent transcription of both genes. However, this is not consistent with the lack of change in expression in both plasmids after SV40p deletion. The second (blue dotted line), represents a looping model, where through looping or supercoiling of the plasmid (as discussed above) onto itself allows the CMV enhancer to regulate both the upstream and downstream SV40 promoters [193,203]. This model is supported by the strong decrease in expression in both plasmids upon CMVe deletion. This does not explain the promoterless expression in the P ΔSV40p plasmid, but that could be explained by the presence of a cryptic promoter site in the backbone sequence 5’ of the

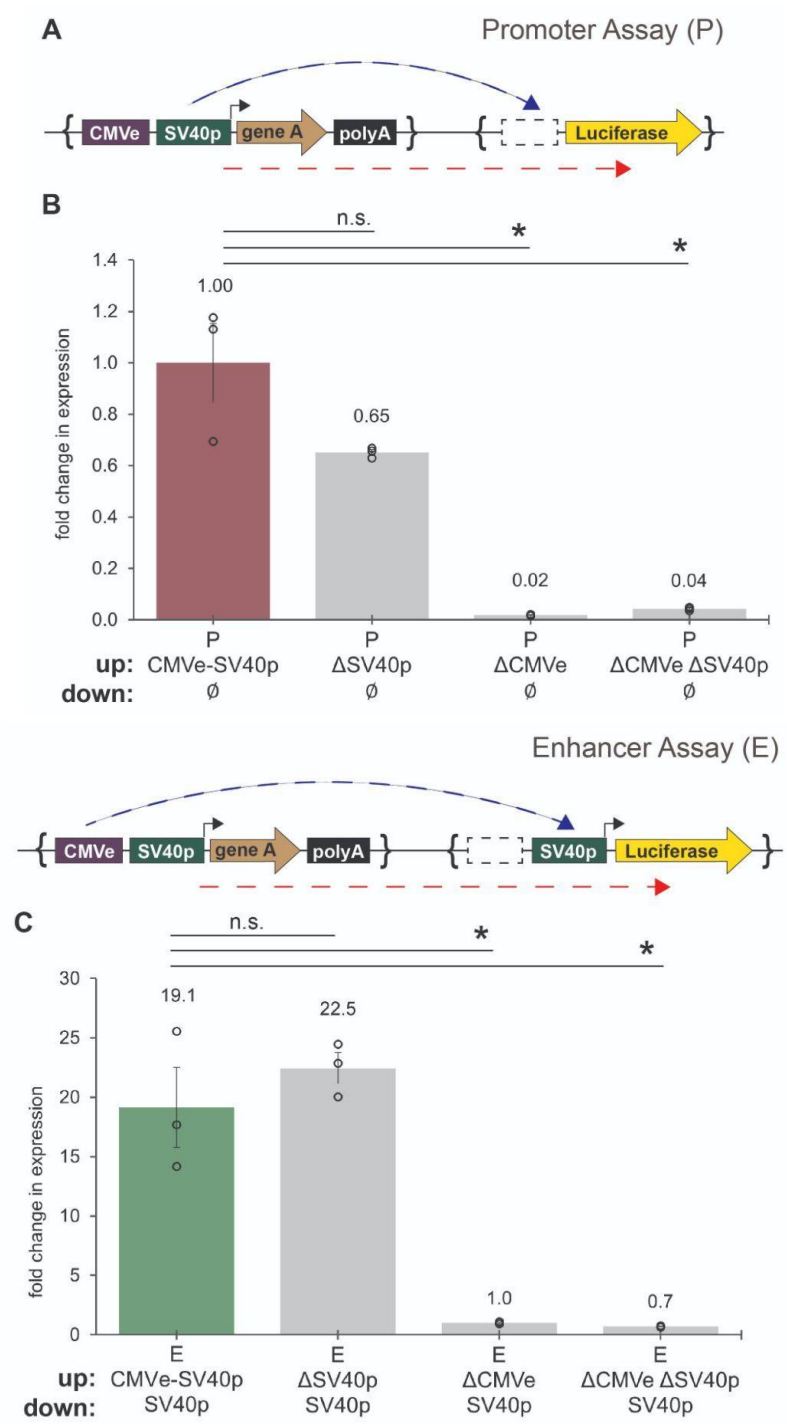


Figure 2.11 Impact of upstream CREs on downstream luciferase expression in a plasmid
a. Diagrams showing layout of promoter **(P)** and enhancer **(E)** assays. Brackets define upstream and downstream expression cassettes, dashed boxes are element insert sites. Dashed lines show potential cause of activity in downstream cassette: looping activity (blue) or read-through transcription (red). **b** and **c**. Fold change in luciferase activity for full plasmids, or plasmids with upstream SV40p, CMVe, or both, deleted (Δ). Y-axis: top line text is assay type, middle line shows CRE deletions from the upstream cassette, bottom shows CRE present in downstream cassette. Three biological replicates (open circles), error bars show standard error, t-test * = $p<0.05$, $p\geq0.05$ = n.s.

luciferase gene (Figure 2.3 supports some positive functional activity for this region) or the bacterial origin of replication acting as a promoter further upstream (not shown on the diagram) as described in Muerdter et al. [185].

These results represent another example of the importance of both proper controls and an understanding of the complexities of assay design, and interrogating unexplained aberrations in assay control data. Proper interpretation of data from the full high-throughput form of this assay set absolutely requires that elements from the upstream cassette be unable to regulate the downstream luciferase gene, as it uses a signal inversion between the two cassettes. Regulation of the downstream gene by an upstream test element would cause a negation of the signal inversion effect, masking the effect of the upstream test regulatory elements and causing high false-negative rates. This cross-communication is likely related to the plasmid looping in a way that mimics the effects of enhancer- or silencer-promoter genomic looping. While this is a source of complication for more complex assay designs, it is also another example of a way in which similar principles of CRE behavior apply across episomal and genomic assays. A solution to the issues caused by intra-plasmid interactions is the use of enhancer blockers as part of plasmid backbones (discussed below).

2.4.6 The Use of Enhancer Blockers in Plasmid Contexts

Enhancer blockers are particularly important features of plasmids for three main types of studies: multi-cassette assays, studies using randomly genomic-integrating plasmids, and studies of enhancer blockers themselves. In all of these, proper use of enhancer blockers in the correct context is important for preventing inter-plasmid or plasmid-chromatin interactions which could confound assay interpretation.

In any assay with multiple expression cassettes, where it is necessary for correct interpretation that the cassettes are regulated separately or where promoter interference is an issue, the cassettes must be separated [204,205]. This can be done by separating one cassette to a different plasmid, separating them and integrating one into the cell or animal genome, or by placing enhancer blockers to separate the cassettes on the same plasmid. Each of these approaches has different advantages and drawbacks. Using enhancer blockers on either an episomal or plasmid construct

has the advantage of guaranteeing that the two cassettes are kept in equal copy numbers when transfected or integrated.

In cases where a plasmid is randomly integrated, either by lentivirus [137,183] or transposase [206], there is the potential for position-variegation effects [207] and cross-regulation. Due to either the spread of heterochromatin which suppresses the inserted sequence, or due to the activity of native CREs near the inserted element, variation in expression of inserts can occur that is unrelated to activity from the plasmid itself. One study estimated a 26% false negative rate in their MPRAs due to integration effects when using a lentivirally integrated assay [208]. In the case of gene therapy, these effects are especially important, as this can make the inserted construct ineffective or cause disease due to elements within the insert mis-regulating genes near the insert site [209]. The use of enhancer-blocking insulators inserted into the plasmid, flanking the desired insert, can help mitigate these effects by preventing spreading of heterochromatin (insulator activity) and contact of enhancers or silencers with insert elements (enhancer blocker activity) [112].

The last case is the use of plasmid-based reporter assays to study the biology and mechanisms of enhancer blockers. These studies also necessarily elucidate some of the ways that enhancer-promoter communication occurs in a plasmid context, as this communication must be present in order to test the ability of enhancer blockers to disrupt it [210]. Plasmid-based approaches have been used by a number of studies on enhancer blockers [193,203,211].

One of most well-studied enhancer blockers in both plasmid and chromatin context, the **chicken hyper-sensitive site 4 (cHS4) element**, has been used in all three of these types of studies and is also the first enhancer blocker to be characterized in vertebrates. It is located at the 5’ end of the chicken β-globin locus, is bound by CTCF, and displays both classical insulator and enhancer-blocking activities in *Drosophila* and human cells [107]. Mapping of the 1.2kb cHS4 element identified a 250bp sequence responsible for the majority of its activity [212], and within that a smaller CTCF-binding fragment, **F2/3**, which is necessary and sufficient for its enhancer blocker activity [105], but does not encode insulator activity [106]. Below, I present data

using this F2/3 minimal enhancer blocker in the plasmid shown in **Figure 2.11** to demonstrate how enhancer blockers can be used to mitigate cross-expression cassette activity. I test the orientation- and position-specificity of these elements both singly and in pairs and discuss the implications of results for the design of enhancer blocker assays. I also compare similarities and differences between my results and those of past studies into this element and other similar enhancer blockers.

Figure 2.12a shows a diagram of the E plasmid from **Figure 2.11,** with the three insertion sites I used for enhancer blocker testing. The enhancer blocker is made of 4 tandem, same-orientation copies of the F2/3 element from cHS4, as in Bell et al.'s paper, which showed that 4 copies were significantly stronger than one copy [105]. **Figure 2.12b** shows the location of this element within cHS4, and **Figure 2.12c** shows the 4x version, which I call F2/3^4 (indicated as >> (fw) or << (rv) in bar charts).

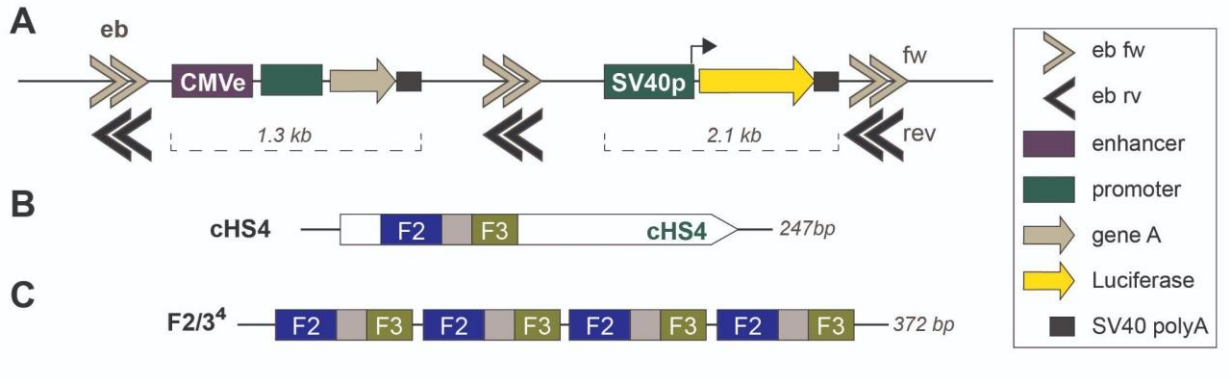


Figure 2.12 Enhancer blocker test plasmid design
a. Diagram of plasmid used for enhancer blocker (eb) testing showing the three possible eb insertion sites and two orientations. Readout of expression uses luciferase gene (yellow). **b.** Diagram of cHS4 eb. Locations CTCF (F2) and SP2 (F3) binding footprints and intervening sequence within cHS4 shown in blue, gold, and gray respectively. **c.** F2/3^4 eb construct of 4 tandem fw orientation copies of the minimal enhancer-blocking F2/3 footprint region of cHS4.

I first tested the F2/3^4 enhancer blocker as a single insert, in either orientation, at all three positions **(Figure 2.13)**. The results show that a single element is capable of enhancer-blocking activity against CMVe, but only when placed between the two cassettes. This activity is independent of orientation, reducing activity by 52% in the forward (fw) orientation, (CMV >> SV40p) and 59% in the reverse (rv) (CMV << SV40p) (differences n.s.).

At first glance, this represents precisely the canonical activity expected of an enhancer in an enhancer blocker assay. The tested element decreases expression only when placed between the enhancer CMVe and the promoter driving expression of our reporter gene (luciferase). It only decreases expression by roughly half, but does not completely reduce it to promoter-only levels. One explanation is that in a circular plasmid, enhancer activity can be bi-directional. This is supported by a paper from the same group that characterized cHS4, which shows that complete blocking of enhancer activity can be obtained by linearizing the plasmid, or by flanking the enhancer on both sides with enhancer blockers [106].

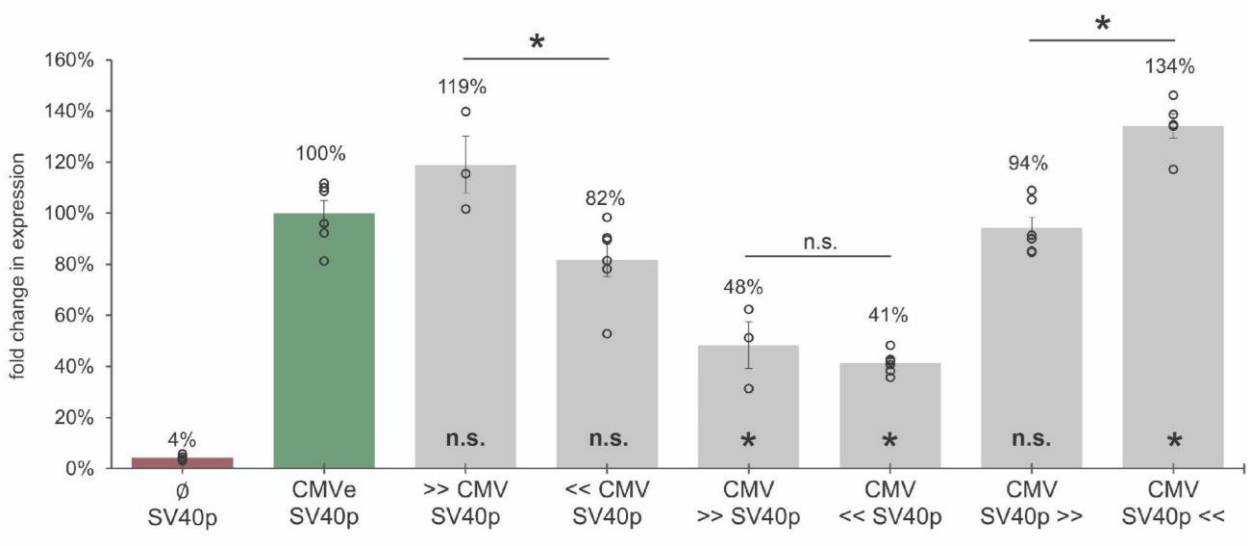


Figure 2.13 F2/3[4] enhancer blocker position- and orientation-dependent activity
Values shown as percents normalized to **CMVe-SV40p** construct with no enhancer blockers (green). **Ø SV40p** is promoter-only control (red). CMVe-SV40p = 25x relative to Ø SV40p. Three or more biological replicates (open circles), error bars show standard error, t-test * = $p<0.05$, $p\geq0.05$ = n.s. relative to CMVe-SV40p (at base of grey bars).

This result has important implications for the design of enhancer blocker assays. It is possible that weaker enhancer blockers might be missed (false negatives), if, when placed between the enhancer and promoter, the enhancer is essentially able to 'escape' blocking by acting in the other direction around the plasmid. This is consistent with my results when testing two tandem cHS4 elements (cHS4[2]) in the same plasmid and positions as F2/3[4] (CMVe >> SV40p) (data not shown). The cHS4[2] element does not reduce activity unless a second copy is added 3' of the Luciferase gene. When the

SV40p-luciferase cassette is flanked by cHS4^2 on both sides, activity is reduced by 34% **(Figure 2.14)**. A single cHS4 is likely much more representative of the strength of genomic enhancer blockers than F2/3^4 or cHS4^2. So for detection of genomic enhancer blockers it is important to constrain enhancer directionality either by linearizing plasmids prior to transfection (this can decrease transfection efficiency - [213]) or to place a 'fixed' enhancer blocker on the outside of the enhancer relative to the promoter. (It should be noted that this is specific to testing enhancer blocker, not insulator, activity, where constructs are chromosomally integrated).

Another explanation for the inability of F2/3^4 to completely prevent CMVe-SV40p communication and for the insufficiency of cHS4^2 for any enhancer blocking, may be the strength of the CMV enhancer. CMV (human cytomegalovirus) is a commonly used ubiquitously active strong enhancer [200] that increases expression 25-fold in K562 in this plasmid. **Figure 2.14** shows the relative strength of a pair of cHS4^2 elements at blocking the CMVe vs an HS2 enhancer inserted in place of the CMVe (same plasmid as shown in **Figure 2.12**). HS2e, an erythroid-specific enhancer from the β-globin locus [214], increases expression 10-fold over the promoter-only control (SV40p), so is about half as strong as CMVe, and correspondingly, the enhancer blockers are about twice as effective against it. Enhancer-enhancer blocker specificity has also been reported for the *Drosophila gypsy* insulator [215]. The implications of this for enhancer blocker assay design are that for identifying novel genomic enhancer blockers, moderate-to-weak enhancers might be better suited to detecting genomic-level enhancer blocker activity.

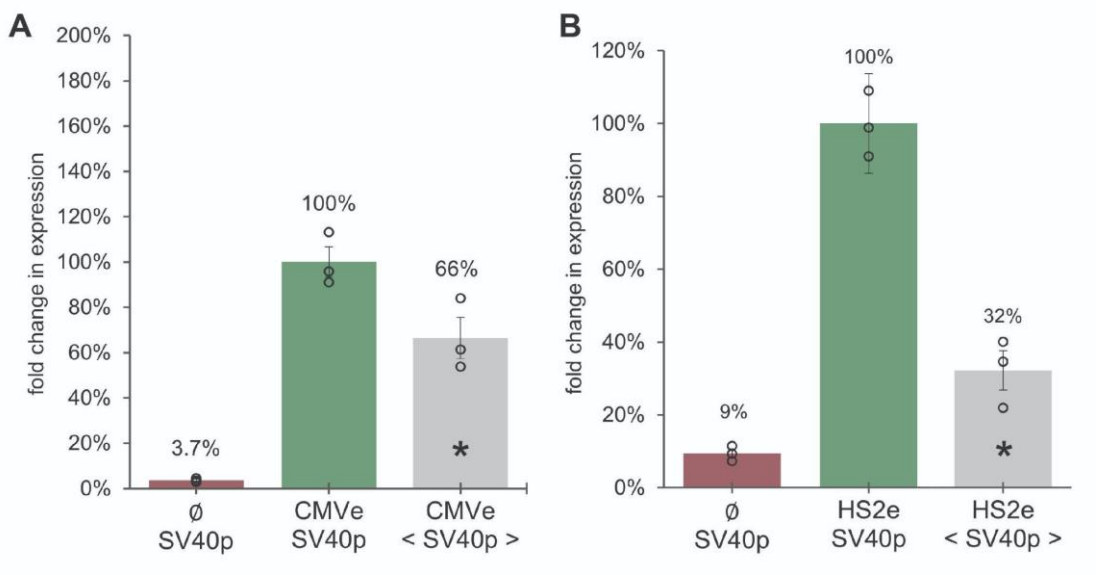


Figure 2.14 Enhancer-specific activity of the cHS4^2 enhancer blocker
Values shown as percent normalized to **CMVe-SV40p** construct (**a**) or **HS2e-SV40p** construct (**b**) with no enhancer blockers (green). **Ø SV40p** is promoter-only control (red). Each **>** or **<** represents two tandem same-orientation copies of the cHS4 enhancer blocker (cHS4^2) in forward (>) or reverse (<). Three biological replicates (open circles), error bars show standard error, t-test * = $p<0.05$, $p\geq0.05$ = n.s. vs e-p control (green).

Another interesting result from **Figure 2.13** is the 34% increase in activity when F2/3^4 is placed in the reverse orientation 3' of the luciferase gene (CMVe SV40p << in **Figure 2.13)**. This phenomenon has not been previously reported in any of the studies of this element or cHS4. The same F2/3^4 element decreased expression when placed 3' by Recillas-Targa et al. They only tested it in the forward orientation, not reverse, however, and their construct used a different promoter, enhancer and plasmid backbone and with the elements in a different order (p - gene - e - F2/3^4) [106]. Given its position directly 3' of the luciferase's polyA tail, it is likely that this increase in expression is due to either some stabilizing effect on the 3'UTR of the transcript, or to secondary structure resulting from the 4x repetitive region or some other sequence, which facilitates termination of transcription by physically preventing RNA Pol II from progressing.

I next tested the effect of all possible orientation and pair-position combinations of F2/3^4 using the three insertion sites (same plasmid as before shown in **Figure 2.12**). While the Felsenfeld group tested a few of these combinations, this is the first data

thoroughly testing all possible orientation-position pair combinations of the F2/3[4] enhancer blocker element using three positionings. Results are shown in **Figure 2.15** and **Figure 2.16.**

A number of different conclusions can be drawn from these data. First, any pair of F2/3[4] flanking either the upstream cassette (CMVe) or downstream cassette (SV40p) is as or more effective than a single F2/3[4] as expected **(Figure 2.15)**. The most effective combination was the CMVe >> SV40p >> plasmid, which almost completely (86% reduction) blocked enhancer activity. However, even with eight total copies of the F2/3 CTCF binding sites, there is still some residual activity above the promoter-only construct (10% above promoter-only) in this condition. Interestingly, flanking the SV40p-luciferase cassette was more effective overall than flanking the upstream cassette with CMVe, even using the same relative orientations (p=0.037 t-test comparing the averages of the two sets).

Another observation is that differences in activity are not driven by the orientation of the central enhancer blocker (between CMVe and SV40p), consistent with the single-F2/3[4] data. In **Figure 2.13** orientation-specific effects occur when the enhancer blocker is positioned outside and not between the enhancer and promoter. Neither are they driven by whether the F2/3[4] are in the *same* orientation. Variable reports of orientation-specificity in enhancer blockers have been published, including in *Drosophila [216]* and human cells [111]. Recillas-Targa et al. reported no orientation-specific effect for the cHS4 250bp element, but another group did for the 1.2kb cHS4 element [217]. While there are orientation-specific differences in paired F2/3[4] enhancer blockers in my data, it appears that they impact the degree of blocking, not completely ablating it, and are relatively small compared to positional effects.

The rest of the position-orientation combinations are shown in **Figure 2.16.** None of these combinations should prevent CMVe-SV40p communication, as the CMVe and SV40p do not have an enhancer blocker between them, and except for one condition, this is the case.

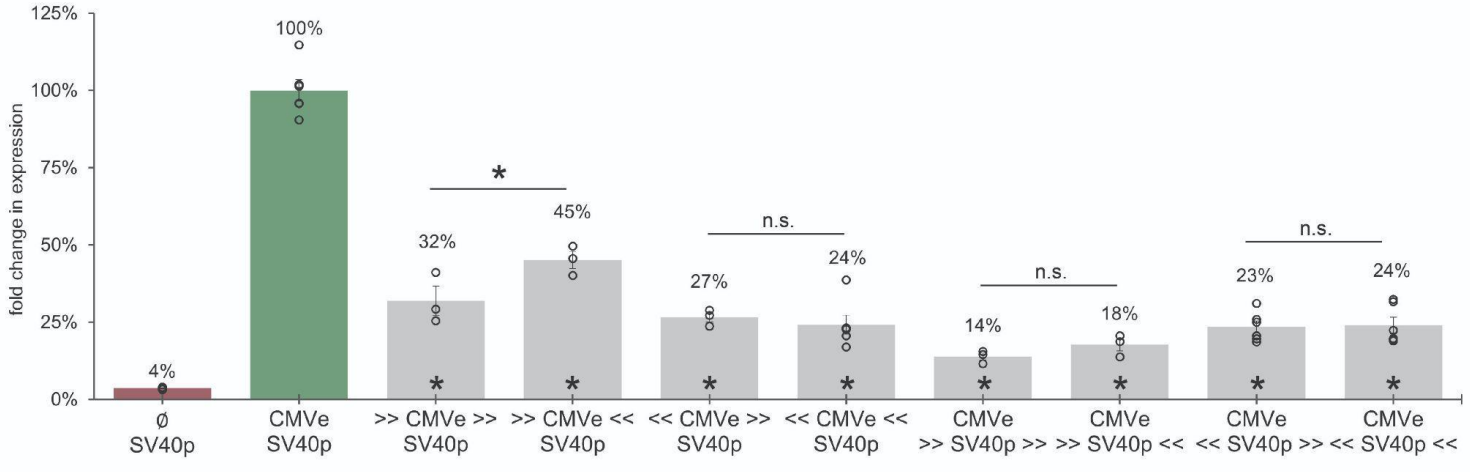


Figure 2.15 Impact of cassette-flanking F2/3[4] enhancer blockers on CMVe-SV40p activity
Values shown as % normalized to **CMVe-SV40p** construct with no enhancer blockers (green). **Ø SV40p** is promoter-only control (red). Three or more biological replicates (open circles), error bars show standard error, t-test * = p<0.05, p≥0.05 = n.s. relative to CMVe-SV40p.

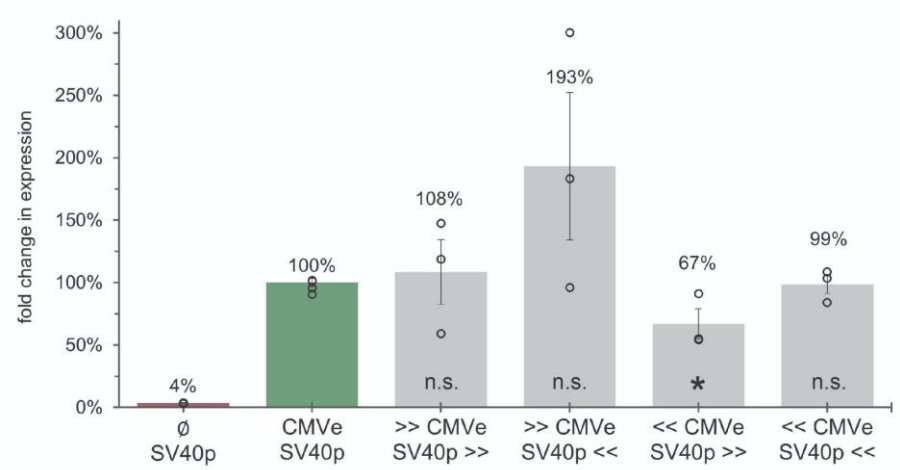


Figure 2.16 Impact of outside-flanking F2/3[4] enhancer blockers on CMVe-SV40p activity
Values shown as percents normalized to **CMVe-SV40p** construct with no enhancer blockers (green). **Ø-SV40p** is promoter-only control (red). Three biological replicates (open circles), error bars show standard error, t-test * = p<0.05, p≥0.05 = n.s. relative to CMVe-SV40p (at base of grey bars).

The precise mechanism(s) of enhancer blocker activity have not yet been fully elucidated. In vertebrates, CTCF is the primary enhancer blocking/insulating factor so many discussions of enhancer blocker mechanisms center on its activities. The cHS4[2] and F2/3[4] elements tested here represent classic examples of CTCF-based enhancer blocking activity. There are currently two primary models of CTCF enhancer blocker activity that both are consistent with different aspects of the above presented data and other studies [113].

The first model postulates some type of 'processive signal' [113] that travels from the enhancer to the promoter, similar to early 'tracking' models of enhancer activity [114]. This processive signal could be an active chromatin state spread from the enhancer to the promoter. An intervening enhancer blocker would create a nucleosome-depleted region, preventing the spread of active chromatin. This model is supported by a study on the cHS4 enhancer blocker in plasmids, where researchers showed that a cHS4 site formed a CTCF-dependent nucleosome-depleted region which interrupted spread of an H3 and H4 acetylation domain and "inhibited transfer of RNA polymerase from the enhancer to the gene" [176,218]. This processive signal model is helpful in explaining the effect of a single intervening enhancer blocker on e-p communication, like in **Figure 2.13.** This could also be one explanation for the results in the flanking paired-F2/3[4] data **(Figure 2.15 and 2.17)**, where the second enhancer blocker is simply preventing the enhancer from bypassing the intervening enhancer blocker by sending this processive chromatin/RNA polymerase signal the other way around the plasmid, reducing activity even further. However it does not explain the orientation- or position-dependence of the strength of effect, or residual activity that remains unblocked.

The second is the looping model, which is the dominant model in the field currently. In this model pairs of CTCF-bound sites bind to each other, resulting in the formation of a DNA loop between the sites, bringing elements within the loop closer, but physically excluding/separating them from contact with elements outside the loop [115]. This model is also consistent with the paired F2/3[4] data, where placing an enhancer blocker on either side of the promoter or the enhancer leads to formation of a double loop in the plasmid, with each element on a separate side **(Figure 2.15)**. This is also consistent with three of the conditions in **Figure 2.16**, where placing the enhancer

blockers such that a loop is formed containing both the CMVe and SV40p does not result in enhancer blocking.

A looping model might also explain the orientation- and position-specific effects, if a certain combination of enhancer blocker distance and orientation is optimal for loop formation. This could perhaps explain the stronger enhancer-blocking effect of flanking the downstream cassette, which due to its larger size (2.1kb), places the two flanking enhancer blockers roughly on opposite sides of the plasmid, where over the upstream cassette (1.6kb) would then form a loop containing about a third of the plasmid. Interestingly, in a genomic context, CTCF site convergence/divergence seems to be a much more important factor in whether a loop forms or not, than in plasmids shown here [219]. For a given position there is no significant difference in effect whether the sites are convergent or divergent.

Other notable aspects of the data from **Figures 2.14-2.17** also agree with observations of CTCF activity. First, CTCF strength is dose-dependent based on CTCF copy number [220]. A single F2/3^4, which has 4 copies, is able to block 59% of activity **(Figure 2.13)**, whereas the two copies in a cHS4^2 are unable to reduce activity (data not shown). However, consistent with some observations in literature showing synergistic effects of multiple weak CTCF-binding sites [220], these increases in enhancer blocking activity do not seem to be entirely linear (additive). **Figure 2.17** shows the predicted enhancer blocking effect for a pair of F2/3^4 sites (white bars), based on the summed observed effects of the F2/3^4 sites of the corresponding positions and orientations that make up the pair, from **Figure 2.13.** These values can be compared to the observed values from **Figure 2.15 and 2.17** (gray bars). For five of these positions, the effect of the two F2/3^4 is additive (marked with a '+' in **Figure 2.17**), for others there appears to be a synergistic effect upon adding the second element.

One of the most clearly identifiable trends in these results is stronger enhancer blocking when the F2/3^4 flanked the downstream cassette vs the upstream cassette. Combining the two plasmid models and the additivity data suggests that looping is the main driver of these differences in activity. Perhaps topological physical constraints of plasmid coiling make a looping state more stable when the enhancer blockers are

placed across from each other, flanking the downstream cassette, where they are roughly on opposite sides of the plasmid as mentioned before. Looping could lead to a stronger effect than that of the individual elements (synergy). When flanking the upstream cassette, looping could be less favorable due to these same constraints. This might also account for the impact of element orientation being larger, accounting for the difference between the first and second two conditions shown in **Figure 2.17**. In cases where looping is not favorable, the two enhancer blockers are still present and have activities according to their independent abilities to block processive chromatin/RNA Pol II signaling, which are additive, as looping is not occurring and the elements are not interacting.

Another observed effect is enhancer-dependent enhancer blocking, similar to *gypsy* in *Drosophila [215]*. A recent study by Hong et al. of three enhancer blockers genome-wide showed enhancer-dependent enhancer-blocking activity, based on the sets of transcription factors bound by that enhancer [109]. In my data, a set of flanking cHS4^2 are twice as effective against the HS2 enhancer as the CMV enhancer **(Figure 2.14)**. Finally, the two main anomalies in the data are the increase in expression in the CMVe-SV40p<< construct (37% increase - **Figure 2.13**) and the decrease in expression in the <<CMVe-SV40p>> construct (33% decrease - **Figure 2.16**). This may result from a combination of factors. First, CTCF has a variety of activities other than enhancer blocking/insulating, that depend on context and binding partners [121]. Second, CTCF-based enhancer blocking activity is highly sequence-context dependent in the human genome [109,220,221].

All of these results cumulatively demonstrate the complexity of intra-plasmid interactions between cis-regulatory elements, as well as the utility of plasmids for studying these interactions through controlled, multiplexed manipulations. Plasmid-based constructs will continue to be an important tool for interrogating enhancer blocker activity and for screening of novel enhancer blocker sequences. Using plasmids to better understand their activity and dependencies also allows us to improve our understanding of the best ways to leverage them as tools for insulating other plasmid-based constructs such as those used for gene therapies or more complex plasmid-based assays.

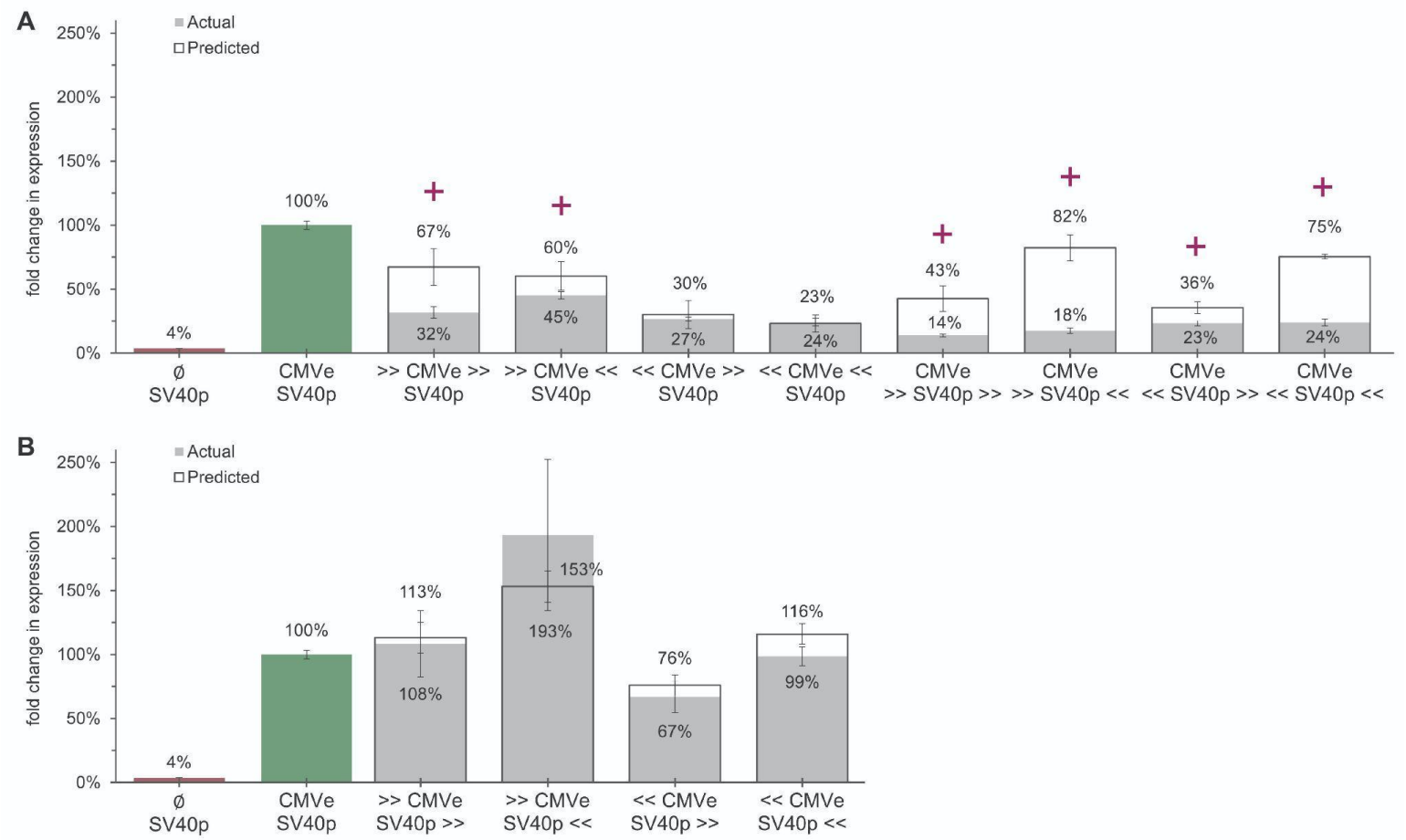


Figure 2.17 Predicted vs observed activity values for flanking F2/3[4] elements
Actual values shown as grey bars, predicted values shown as white bars, for **a.** up- or down-stream cassette-flanking F2/3[4] pairs or **b.** outside-flanking F2/3[4] pairs. Magenta + indicates actual vs predicted values are significantly different (potential synergy). Predicted values calculated as: *(1 - % observed decrease from F2/3[4] A - % observed decrease from F2/3[4] B)* using data from Figure 2.14 for individual F2/3[4]. Observed data (gray) error bars are standard error for (three or more) biological replicates. Predicted value error bars are calculated as: *square root[(std. error F2/3[4] A)^2 + (std. error F2/3[4] B)^2]*.

2.4.7 The Impact of Transfection on Cell Function

The final consideration for plasmid-based reporter assays which I will discuss here relates not to the plasmid components, but to cellular responses to the transfection process and plasmid uptake itself. The process of introducing foreign DNA to a cell line or organism cannot be considered entirely neutral.

In the case of almost any transfection method for cell cultures, cells must be collected, spun, counted, and distributed onto plates. It is challenging to maintain a cell culture at 37°C at all times during this process, and so cells are necessarily perturbed. Additionally, during electroporation, electric current is used to porate cell membranes [163]. Analysis of transfected cells must make the assumption that the perturbations necessary to introduce DNA either do not elicit significant biological changes in response, or that if these do occur, they are temporary and do not impact the cell at the time of readout (24 hours to weeks after). While these assumptions may not be entirely correct, there is no option to avoid the impact of cell manipulation. So interpretation must take into account that readouts from cell lines may be impacted, however if so, this impact is present across all studies using transfected cell lines, and does not mean that useful information cannot be obtained.

If the substrate used is viral, an inflammatory response by the cell is possible [163]. However with a viral integrating substrate, readouts can often be taken at a much later time point, after this effect may have subsided and DNA is stably integrated into the cellular genome. Similarly, presence of plasmid DNA in the cytoplasm after transfection can trigger an innate immune response [185]. The cellular transcription changes needed to activate factors involved in this response mean that reporter assay readout might not be taking place in a cell in its basal state. Muerdter et al. report that a majority of the most commonly used cell lines have high expression of genes related to the antiviral interferon response pathway [222]. While they reported that this did alter enhancer assay results in one cell line (HeLa), they were able to mitigate this effect using inhibitors of this pathway [185]. They confirmed that some cell lines do not activate this response after transfection, including K562 cells. For any reporter assay using interferon-responsive cell lines, researchers may want to consider replicating their method using inhibitors to prevent this effect during transfection.

Finally, for experiments using multiple plasmids, such as co-transfection controls, promoter interference and resource competition can be an issue. A paper by Di Blasi et al. nicely details examples of competition for cellular machinery between co-transfected plasmids and provides suggestions for addressing this issue [223]. However in at least one case in my data **(Figure 2.10)**, co-transfection of two plasmids containing the CMV enhancer did not result in an overall decrease in expression, but actually an increase (relative to the appropriate co-transfection control plasmid with no extra CMV enhancer), indicating that this effect is perhaps not universal or may not occur in some experiments.

The brief overviews presented here are in no way a comprehensive review of all the possible physiological and confounding effects that can result from transfection of ectopic DNA into a cellular system. I have presented a few of the main concerns to illustrate the importance of an awareness of the entirety of the model system used for reporter assays and the potential for confounding or biasing effects due to the biological system being used, rather than the elements being tested.

2.5 Conclusion

The importance of plasmid backbone sequences' (sequence context) contribution to function, the ability of multiple CREs to interact across a plasmid, and the importance of orientation/positioning of enhancer blockers, reflect many aspects of known CRE function in genomic contexts. While this added complexity requires a closer attention to the details of plasmid design, this also supports its strength as a model of genomic CRE activity. Understanding the specific biological and disease relevance of a particular element in its native context is perhaps best done using other methods, then complementing it with reporter assay data. However, for interrogating mechanisms, plasmids are a powerful tool due to the ease and control with which we are able to manipulate them. They have been used elegantly for decades to help dissect the basic mechanisms of CRE elements. However their simplicity can be deceptive, and understanding the particulars of any reporter assay design is a crucial step in generating informative, reproducible results. While some of these aspects of assay

design have been apparent for decades, others are more recently emerging or being re-discovered.

2.6 Methods

2.6.1 Plasmids

The Renilla co-transfection control plasmid used is a 3705bp pRL-SV40P containing *Renilla reniformis* luciferase under control of the SV40 enhancer and SV40 promoter (and a same-orientation AmpR and ori). Except for **Figure 2.8** and **Figure 2.10**, test plasmids used the same backbone containing ori, AmpR for selection. All SV40 promoter, CMV enhancer, and Gateway sites, and Firefly luciferase genes are the same sequence, rearranged by restriction-digest and ligation or Gateway recombination (attR -> attB). Firefly luciferase and SV40 promoter originated from a pGL3 vector.

2.6.2 Cell Culture

All transfections listed in this chapter were completed in the human myelogenous leukemia cell line K562 (CCL-243^{TM}, ATCC). Cells were grown at 37^{o}C and 5% CO_2. Cells were cultured in RPMI-1640 complete media made with RPMI-1640+L-glutamate media (ThermoFisher, 11875093) containing 10% heat-inactivated FBS (ThermoFisher, 10437028) and 1x antibiotic/antimycotic (ThermoFisher, 15240112). Cells were passed every 2-3 days depending on confluence and were discarded after passage 10.

2.6.3 Transfections

All transfections of K562 used electroporation with a NEPA21 Electroporator (Nepagene). Cells were checked for minimum 75% viability and 50% confluence prior to electroporation. Cells were collected by centrifugation at room temperature (RT) at 100xg for 10min. Supernatant was removed and cell pellets resuspended in 15mL per initial 50mL conical of cell culture of RT Opti-MEM Reduced Serum Medium (ThermoFisher, 31985062). 500uL was removed and set aside for a cell count and viability checks. Remaining cells were centrifuged again at 100xg for 10min at RT and the 500uL aliquot counted during this time. After spin, supernatant was again removed

and cells were resuspended with fresh Opti-MEM to a final concentration of 1x10^6 cells/90uL, measured for accuracy by pipette.

For each condition biological replicate, 99uL of cells was added to 1.5mL microcentrifuge tubes containing 11uL pre-aliquoted DNA (see 2.6.3 below). Cell-DNA mixtures were mixed by pipette, then 100uL of mixture (90uL cells, 10uL DNA) was added to each 2mM cuvette (x3 for biological replicates) (Bulldog Bio,12358-346). Cuvettes were electroporated using poring pulse: 275V, 5ms length, 50ms interval, 1 pulse, 10% D rate, + polarity. Transfer pulse: 20V, 50ms pulse length, 50ms pulse interval, 5 pulses, 40% D rate, +/- polarity. Immediately following electroporation of each cuvette, 900uL RPMI 1640 complete media pre-warmed to 37^{o}C. Cells were transferred to 24-well tissue culture plates by pipette and incubated (as listed above in Cell Culture) for 48 hours.

2.6.4 Preparation of DNA for Electroporation

All data in this chapter were generated using Firefly luciferase expression. DNA amount per cuvette was set at 1.5ug of a 5416bp pGL3 plasmid, per 1x10^6 cells, and all other plasmids were transfected in molar-equivalent amounts. All were co-transfected with a pRL plasmid containing Renilla luciferase at a 1:25 molar ratio. DNA concentrations were measured by Qubit 43 Fluorometer (Invitrogen) using the dsDNA BR Assay Kit (Invitrogen, Q32850). All DNA to be used in a transfection was measured at the same time, using the same Qubit dye & buffer mastermix to account for measurement fluctuation due to mastermix preparation and room temperature. DNA mixes for each condition and replicate were made in a 1.5mL microcentrifuge tube containing Firefly DNA and Renilla DNA in molecular-grade water, at a 1.1x scale to account for pipetting error. These were made prior to electroporation to minimize cell time at RT.

2.6.5 Readout of Luciferase Signal

Luciferase readouts used Promega's Dual-Glo Luciferase Assay System (Promega, E2920). Readout was done using a GloMax-Multi+ Detection System plate reader (Promega, E7081). At 48 hours (+/- 4 hours) post-electroporation, all cells from

each well were collected by pipette into individual 1.5mL microcentrifuge tubes. Cells were centrifuged at 500xg for 5 minutes. Supernatant was removed by either pipette or vacuum except for ~50uL to avoid disrupting the cell pellet. 450uL RT 1x PBS (Phosphate-buffered saline, Invitrogen, 10010023) was added and cells resuspended by vortexing. Immediately before loading onto luciferase readout plates, cells were vortexed again to ensure even suspension. For each biological replicate, 3 wells of a white, flat-bottomed 96-well plate (VWR, 82050-726) were each loaded with 50uL of cell-PBS suspension. After a full plate was loaded with sample, 50uL of Dual-Glo Luciferase reagent was added and mixed by multichannel pipette. The plate was incubated for a minimum of 10 minutes prior to readout (maximum 30min). Plates were read for Firefly signal at 10 reads per well, removed, and 50uL of Stop & Glo reagent added and mixed by multichannel pipette. Plates were again incubated for a minimum of 10 minutes, and Renilla signal read at 10 reads per well.

2.6.6 Analysis of Luciferase Expression Data

The 10 reads taken per well were averaged separately for Firefly and Renilla signal to get well values. Background signal was calculated by averaging reads across 6 wells (2 biological replicates) of untransfected cell controls. Background Firefly signal was subtracted from each Firefly technical replicate to get background-adjusted values. This was repeated for Renilla signal. Adjusted Firefly technical replicates were then divided by the corresponding adjusted Renilla value for that well. A low-expression control condition was chosen for value normalization (typically a promoter-only plasmid), which varied depending on the experiment. Background adjusted F/R tech rep values were individually divided by this control (average of all of its F/R biological replicate values). This gave a final fold-change value.

Normalized F/R = [(F - F cell background)/(R - R cell background)]

Fold change = (normalized F/R for tech rep) / (bio rep average of normalized F/R for control condition)

Technical replicates were then averaged to get biological replicate values. Biological replicates were averaged and final fold change values plotted on graphs. Where multiple reads for a condition were taken across different plates or days, or where

values are represented as % instead of fold change, values were normalized to correct for variation and in order to be able to compare replicates. Normalization was done by dividing all values by the 'high' control readout for that plate - the same high-expressing control plasmid used on every plate within an experiment (for example CMVe-SV40p).

2.6.7 Statistical Testing

Error of biological replicates is given as the standard error for all graphs [standard deviation(bio reps)/square root(# bio reps)]. T-testing was run comparing biological replicates using a one-sided, heteroscedastic model and significance indicated for that comparison by * if the test showed $p<0.05$. Unless otherwise indicated, on bar charts showing fold changes in expression: error bars represent standard error of three or more biological replicates, circles show individual biological replicate values, and the number above each bar is the average value of the three biological replicates. Where there were replicates from multiple plates/days of the same condition, and normalization to the high-expressing control was done, t-testing was run pre- and post-normalization and if either one of these t-tests gave ≥ 0.05, the conditions were listed as not significant.

2.7 Notes and Acknowledgements

Jessica Switzenberg helped with construction and design ideas for the plasmids. Thanks to Dr. John Moran and Dr. Shigeki Iwase who generously allowed me to use their equipment for electroporation and readout. Thanks to Greg Farnum for construction of the tricky 4x F2/3 repetitive enhancer blocker element using Emma Assembly [161]. The design of the original high-throughput assay that I used for exploration of assay design elements in this chapter was originally conceived by Dr. Alan Boyle and built by Jessica Switzenberg. The new iterations were a collaborative effort between myself, Torrin McDonald, Sierra Nishizaki-Sweiso and Jessica Switzenberg.

CHAPTER III

Multiplexed Long-Read Plasmid Validation and Analysis Using OnRamp

3.1 Abstract

Recombinant plasmid vectors are versatile tools that have facilitated discoveries in molecular biology, genetics, proteomics, and many other fields. As the enzymatic and bacterial processes used to create recombinant DNA can introduce errors, sequence validation is an essential step in plasmid assembly. Sanger sequencing is the current standard for plasmid validation, however this method is limited by an inability to sequence through complex secondary structure, and lacks scalability when applied to full-plasmid sequencing of multiple plasmids due to read-length limits. While next-generation sequencing (NGS) does provide full-plasmid sequencing at scale, it is impractical and costly when utilized outside of library-scale validation. Here we present OnRamp (Oxford nanopore-based Rapid Analysis of Multiplexed Plasmids), an alternative method for routine plasmid validation which combines the advantages of NGS's full-plasmid coverage and scalability with Sanger's affordability and accessibility by leveraging nanopore's novel long-read sequencing technology. We include customized wet-lab protocols for plasmid preparation along with a pipeline designed for analysis of read data obtained using these protocols. This analysis pipeline is built into the OnRamp web app, which generates alignments between actual and predicted plasmid sequences, quality scores, and read-level views. OnRamp is designed to be broadly accessible to researchers regardless of programming experience in order to facilitate more widespread adoption of long-read sequencing for routine plasmid validation. Here we describe the OnRamp protocols and pipeline, and demonstrate our ability to obtain full sequences from pooled plasmids while detecting sequence variation even in regions of high secondary structure at less than half the cost of equivalent Sanger sequencing.

3.2 Introduction

Cloning of recombinant DNA into plasmid vectors is a fundamental tool of molecular biology and central to many discoveries in genetics for decades, including the first sequencing of the human genome [224]. It continues to underpin modern-day research in genomics, protein expression and purification [225], transcriptional regulation [123], and gene therapies [226]. However the standard for plasmid sequence validation, an important step in cloning due to the error-prone nature of recombinant assembly [227]; [228], is still Sanger sequencing, a PCR-based method invented in 1977 [229].

Sanger sequencing uses a PCR-amplification based approach to obtain base-pair resolution of DNA sequence in stretches of up to 900bp [229]. Despite being an important tool for simple, low-throughput sequence validations, Sanger also has a number of limitations. These include the need to synthesize target-specific primers, inaccuracy in long mononucleotide stretches [230], difficulty sequencing through regions with strong secondary structure (such as repetitive elements), and a limit of about 900bp sequence output per run [231]. While the 900bp limit can be addressed by tiling multiple sequencing runs across the same plasmid, this requires synthesis of multiple primers and quickly becomes expensive and laborious when applied to multiple transformants. As a result, typical Sanger validation protocols involve sequencing only the portion of a plasmid that was most recently modified or that contains protein-coding sequence, and routine validation of full plasmid sequences after cloning is not standard practice. However, we are increasingly recognizing the potential impacts that plasmid backbone sequences, structural elements, and bacterial sequences, which are not commonly sequence-validated and can vary widely between plasmids, have on regulation and activity of other plasmid components. Most plasmids contain multiple elements which contribute to function [232]; [233], including bacterial sequences [185], which may accumulate undetected errors as a result of spot-check sequencing approaches and can impact downstream function.

High-throughput next-generation sequencing (NGS) does allow for simultaneous sequencing of large numbers of plasmids and provides full plasmid sequences [234]. However, NGS is cost-prohibitive outside of large-scale approaches, and sample

pooling coordination, indexing compatibility issues, equipment cost, and turnaround time are major barriers to its widespread adoption and make it unsuited for routine plasmid validation. Additionally, NGS does not allow for detection of variation outside of unique regions of plasmids in libraries due to the inability to uniquely map short reads to an individual plasmid [235].

With the advent of Oxford Nanopore Technologies' (ONT) benchtop long-read sequencing platform, a new option has become available for plasmid validation. Nanopore sequencing platforms can generate reads on the order of megabases and have been employed in a variety of applications, including resolving previously intractable complex structural variation , whole genome sequencing, targeted enrichment sequencing, clinical diagnostics, RNA sequencing, and metagenomics [236]; [237]; [238]; [239]; [240]. Nanopore's read length allows for validation of entire plasmid sequences in a multiplexed format, unlike Sanger, and the low cost of the platform relative to NGS makes it more accessible. While some groups have applied nanopore to the task of plasmid sequencing, they use transposase- and barcode-based *de novo* assembly approaches [241]; [242]; [243]. Importantly, the approaches used in these studies all require bioinformatic expertise in order to properly analyze data and interpret results from libraries prepared using these methods. In order to take steps toward more widespread adoption of full-plasmid sequencing using nanopore, accessibility is crucial to address. Researchers who do routine plasmid validation are likely to be bench scientists. Therefore, in order to improve accessibility and utility to a broader population, it is important to have protocols and analysis tools that not only allow for rapid, easy analysis of nanopore plasmid data, but also provide analysis outputs which make interpretation as easy as it is for the current validation standard, Sanger sequencing.

Here we present OnRamp (Oxford Nanopore-based Rapid Analysis of Multiplexed Plasmids), a tool that leverages ONT's long-read technology to obtain full sequences of pooled plasmids. OnRamp addresses the need for an approach that is simpler and more cost-effective than NGS, while providing full plasmid sequences at medium-throughput scale in a rapid, amplification- and barcode-free manner at under $1.25 per kb, less than half the cost of equivalent Sanger sequencing. OnRamp

comprises both custom protocols and an analysis pipeline for ONT long-read pooled plasmid data, available through the OnRamp webapp (https://onramp.boylelab.org/). OnRamp uses a reference-based approach which allows for the viewing of reference-consensus alignments, alignment quality scoring for rapid identification of correct vs incorrect plasmids, and a view of individual read alignments for interpretation of base call confidence and detection of sub-population level variation, all through the OnRamp web application, making interpretation of sequencing results accessible and simple. We describe here custom plasmid preparation protocols for use with OnRamp, testing of the OnRamp pipeline using simulated read data, and demonstrate detection of variation using real plasmid data across plasmid pools containing both dissimilar and highly similar (clonal) plasmid sequences.

3.3 Results

3.3.1 OnRamp Protocols and Pipeline

The OnRamp protocols use ONT's nanopore sequencing platform, which requires ligation of DNA ends with specialized adapters used to facilitate sequencing. Our method is unique in that it leverages full-length plasmid reads for assembly and does not require barcodes for multiplexed runs, which allows for rapid and simple sample preparations. Here we provide two methods for plasmid pool preparation for OnRamp runs based on the adapter ligation method: transposase-based or restriction digest-based **(Figure 3.1a)**. The first uses ONT's Rapid Sequencing Kit which utilizes a transposase to randomly fragment equimolar pooled plasmid DNA and simultaneously ligate ONT's specialized sequencing adapters, to provide compatibility with typical nanopore protocols. In the second, plasmids are linearized with a single-cutter restriction enzyme (RE), which allows for control of both the number and locations of cuts within the plasmid, increasing the likelihood of obtaining full-length plasmid reads compared to the transposase-based approach, and which is used for preparation of plasmid pools containing clonal or highly similar plasmid copies. Restriction-digested plasmids are pooled in equimolar amounts after digestion, end-repaired and mono-adenylated, then adapters are added to plasmid ends using ONT's ligation kit.

Following adapter ligation, plasmid libraries are loaded onto primed Flongle flow cells and deeply sequenced to base-pair resolution over 16-24 hours using ONT's MinION sequencing platform with Flongle adaptor **(Figure 3.1b)** with single reads spanning entire plasmids. Basecalled read files generated by a nanopore run are then submitted to the OnRamp web app along with plasmid reference sequence files for analysis. The pipeline run by the web app aligns reads using medaka (github.com/nanoporetech/medaka) to generate a consensus sequence for each plasmid by aligning reads to user-provided references **(Figure 3.1c)**, then consensuses are aligned to their matched references using EMBOSS Needle [244] to generate optimal global pairwise alignments **(Figure 3.1d)**.

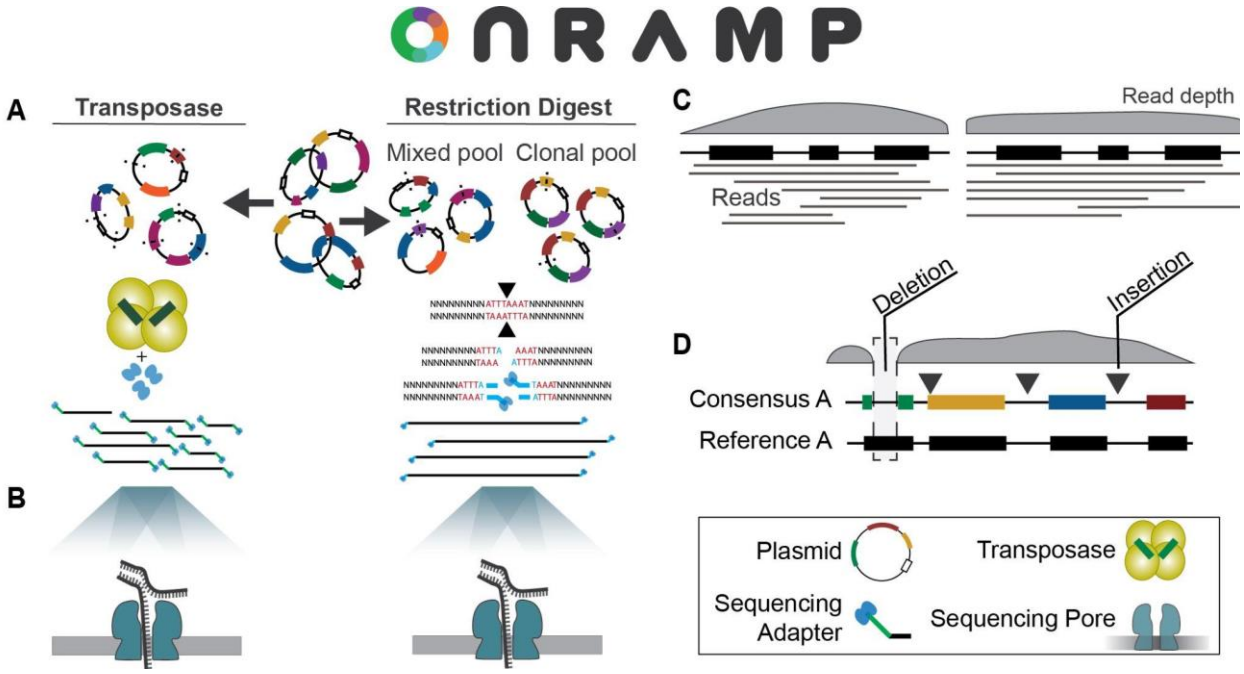


Figure 3.1 OnRamp protocol and pipeline
Pooled plasmids have Nanopore adapters added by transposase or by digestion & ligation (**a**) and then are sequenced (**b**). Basecalled reads are provided to the OnRamp webapp, which generates consensus sequences (**c**). Consensus sequences are then aligned to user-provided references to identify variation (**d**).

After submitting a run on the OnRamp web app **(Fig 3.2a)**, users are given outputs which include: a sequence-level alignment between reference and consensus files showing any insertions, deletions or base substitutions **(Figure 3.2b)**, a quality score based on number and length of insertions or deletions (gaps), or base substitutions in the consensus relative to the reference **(Figure 3.2c),** and an Integrated

Genome Viewer (IGV) [245] view showing read alignments used to generate the consensus **(Figure 3.2d).**

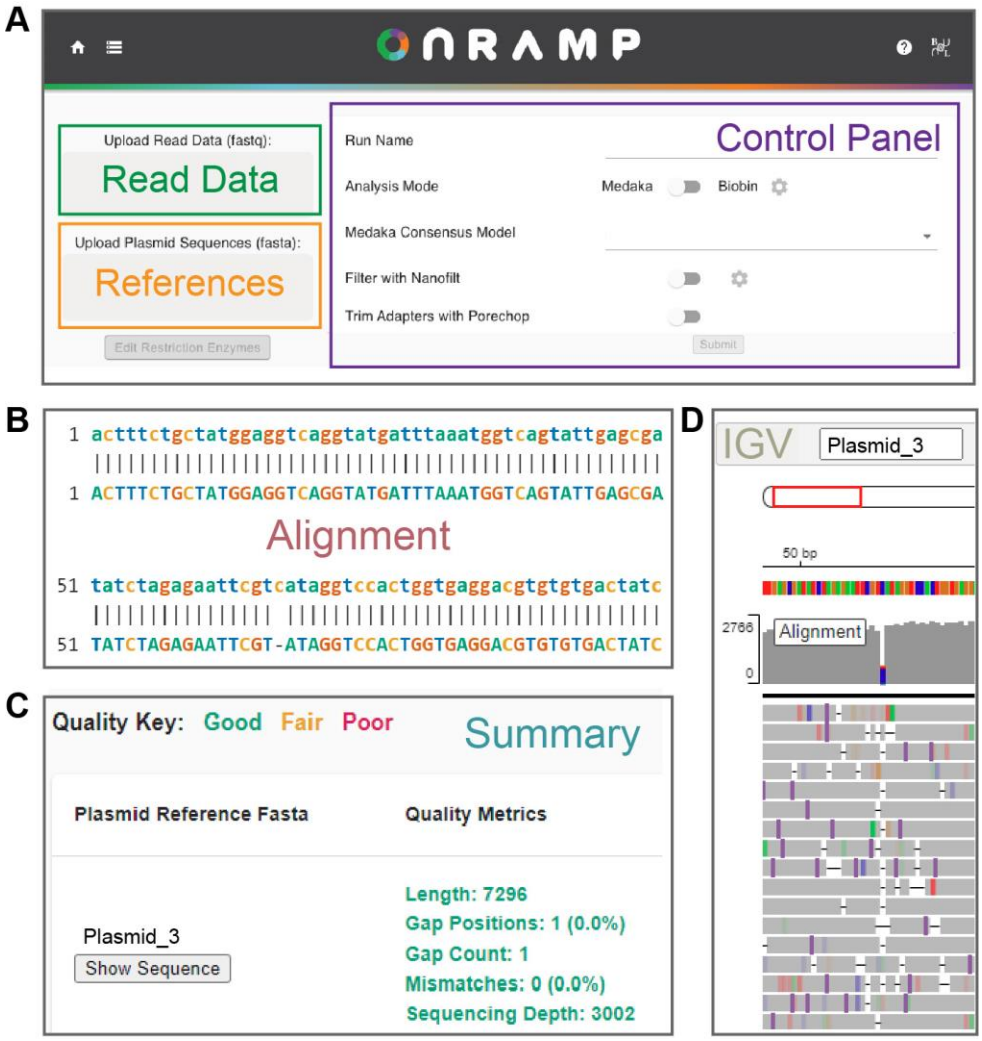


Figure 3.2 OnRamp webapp and analysis display
a. Image of submission page where users submit read data and plasmid reference files and choose analysis settings. **b-d**. Output generated from example data, including sequence alignments (**b**), alignment quality metrics (**c**), and IGV viewer panel showing individual reads (**d**).

3.3.2 OnRamp Detects Base-Pair Level Variation in Simulated Datasets

To assess the ability of our OnRamp pipeline to accurately detect sequence variation occurring in plasmids from a mixed plasmid pool, we first constructed simulated read data using NanoSim [246], a tool designed to simulate nanopore reads. Read libraries were constructed for 30 dissimilar plasmids (average length 4.4kb) of known sequence, simulated to be prepared using the ONT transposase rapid adapter

kit, giving randomly distributed read start sites. NanoSim generated 29,984 reads, with an average of 967 reads per plasmid, which were pooled and then mapped back to their respective references using OnRamp in medaka mode.

In medaka mode, OnRamp uses the medaka consensus tool to generate polished consensus sequences by simultaneously mapping all reads against all references to generate best alignments, using reference sequences in place of a draft genome assembly. OnRamp mapped on average 614 reads to each plasmid, which were used to create 30 consensus sequences. Across these 30 sequences, a total of three errors (2 missing single bases at the start of one consensus due to lack of depth, and a 1bp gap at a homopolymer run in another plasmid) were observed upon alignment back to their reference sequences (1 error per 10 plasmids). Since no errors were expected given these reads originated from known sequences, we tested what level of read coverage would eliminate these gaps. We repeated consensus construction using 500%, 100%, 50% or 10% of the 29,984 reads and measured gaps in the resulting alignments **(Figure 3.3a).** Alignment accuracy varied with read coverage as expected, with more coverage giving increasing accuracy. Consensus errors consisted primarily of gaps at homopolymers and missing sequence at consensus ends due to unequal coverage across the alignments. Additionally, increased errors in calling homopolymers is a known limitation of ONT data [247].

Next we used this simulated dataset to test OnRamp's ability to detect indels. A simulated read pool generated from a reference plasmid containing a 100, 10, or 1bp insertion or deletion was added to the 30-plasmid read pool. We used OnRamp to generate polished consensuses as above and results showed that insertions and deletions of 100bp, 10bp and 1bp were all correctly identified even at 100 reads per plasmid **(Figure 3.3b-e)**. Read count did not impact ability to detect mutations, but rather affected whether additional variation occurred elsewhere in the consensus (points above the dotted lines in **Figure 3.3d and e**) as a result of lack of coverage, especially at map ends and homopolymers.

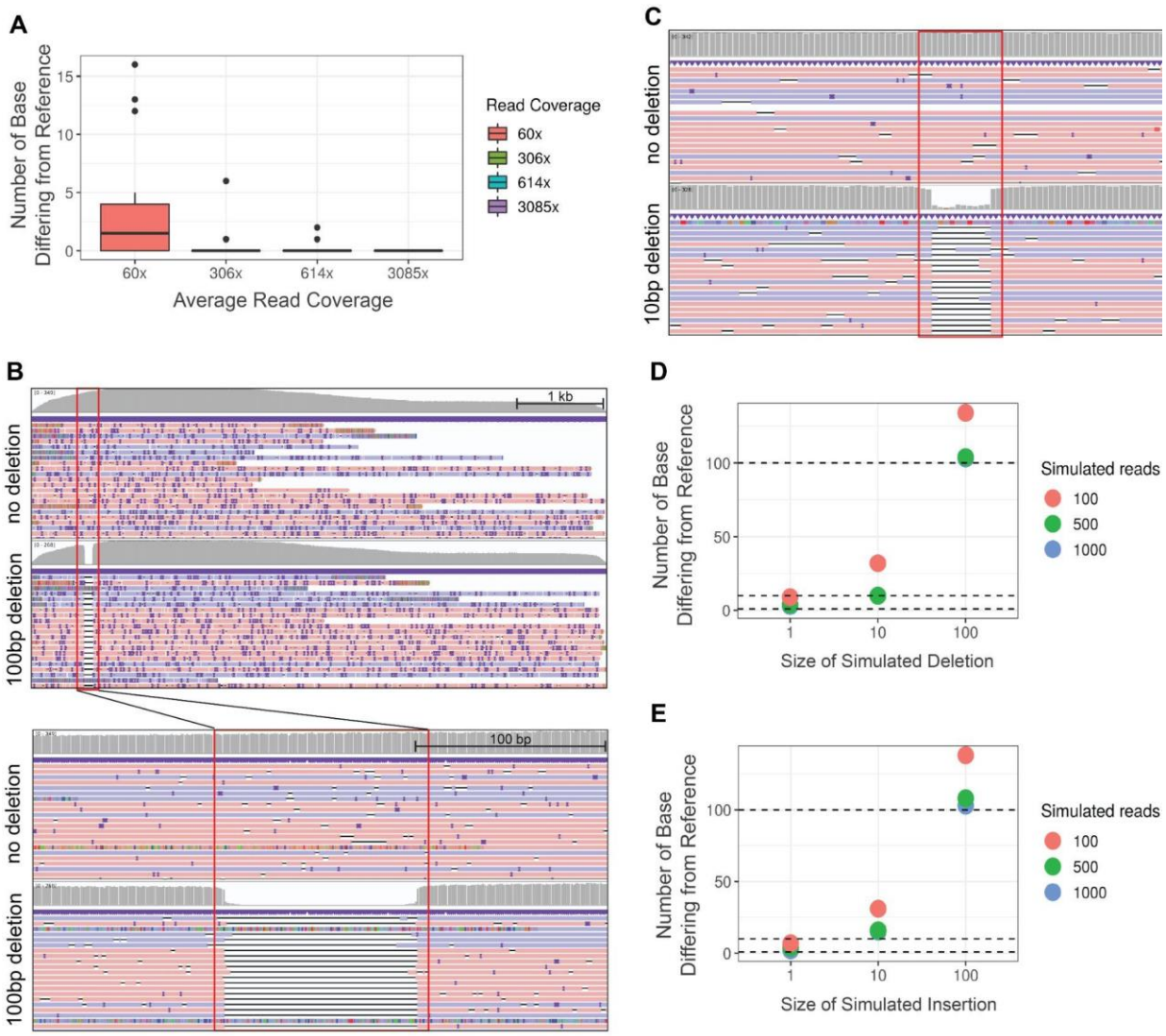


Figure 3.3 Detecting insertions and deletions in a plasmid pool using a simulated read library

a. Consensus accuracy vs read coverage - number of gaps in consensus sequence (before indels) for simulated read depth experiments at various read coverage. **b and c.** IGV view of read pileups for reads with vs reads without a 100 bp deletion (**b**) and a 10 bp deletion (**c**). Deletions highlighted by red boxes. Gray top row shows read depth at each position. Below, minus-strand reads (purple) and plus-strand reads (red) with inserts (dark purple) and deletions (black). **d and e.** Number of base-pair differences between reference and consensus files for each simulation condition at different read depths. Dotted lines indicate expected number of differences due to simulated deletion (**d**) or insertion (**e**).

3.3.3 OnRamp Correctly Assigns Reads to Highly Similar Plasmids

We next used simulated data to test OnRamp's ability to correctly assign reads originating from a pool of plasmid sequences with high sequence identity without using barcoded sequencing adapters. We created an average of 971 reads for each of 16 plasmid references differing only by 24bp, 12bp, or 6bp-long unique regions and used NanoSim to construct simulated read pools. OnRamp was then used to analyze reads and references using biobin mode. In biobin mode, OnRamp scans all provided reference sequences for unique sequence to use for distinguishing the references, then aligns each read to these unique regions to obtain an alignment score. Two tunable alignment scores, context and fine map, are used to assign each read (see methods, **Figure 3.4 a and b**). Each read that meets the scoring criteria is binned to this reference and then OnRamp generates a consensus for each plasmid individually from its assigned bin of reads using medaka's consensus tool. Using biobin mode, fewer than 6% of reads were assigned to the incorrect reference **(Figure 3.4c)** and OnRamp was able to generate consensus sequences for pools containing plasmids that differed only by the 12bp or 24bp markers. For the 6bp marker, consensus sequences contained many more gaps due to a low number of assigned reads. We also ran this test using OnRamp's medaka mode **(Figure 3.4d)**, however this is not recommended as medaka mode uses non-uniquely assigned reads in consensus generation and can lead to read mis-assignment. We suggest using OnRamp's biobin mode for correctly mapping highly similar plasmid pools where there is at least 24bp of unique sequence to differentiate the plasmids. For highly similar plasmid pools with less than ~24bp unique sequence (for example plasmids that are clonal copies), a simple alternative to the plasmid preparation protocol is provided below which works for any amount of similarity.

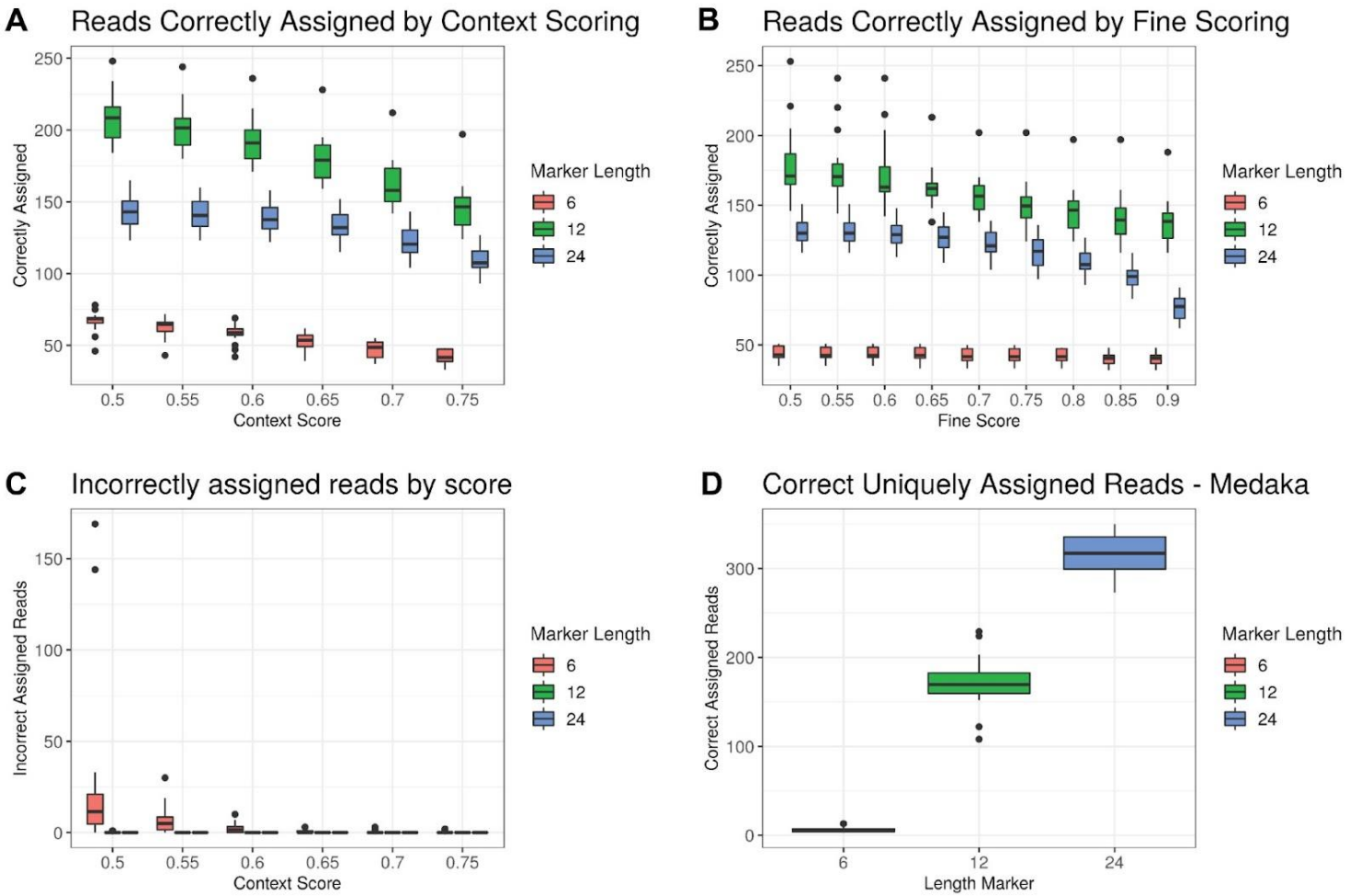


Figure 3.4 Number of correctly assigned reads in different modes for 30 simulated plasmids containing 6bp, 12bp, or 24bp unique regions

Reads aligned using biobin mode (**a, b**) with a fixed fine score, comparing read counts at different context scores (**a**) or with a fixed context score, comparing read counts at different fine scores (b). The number of reads incorrectly assigned using different context scores (**c**). Count of uniquely mapping, correctly assigned reads using medaka mode (**d**).

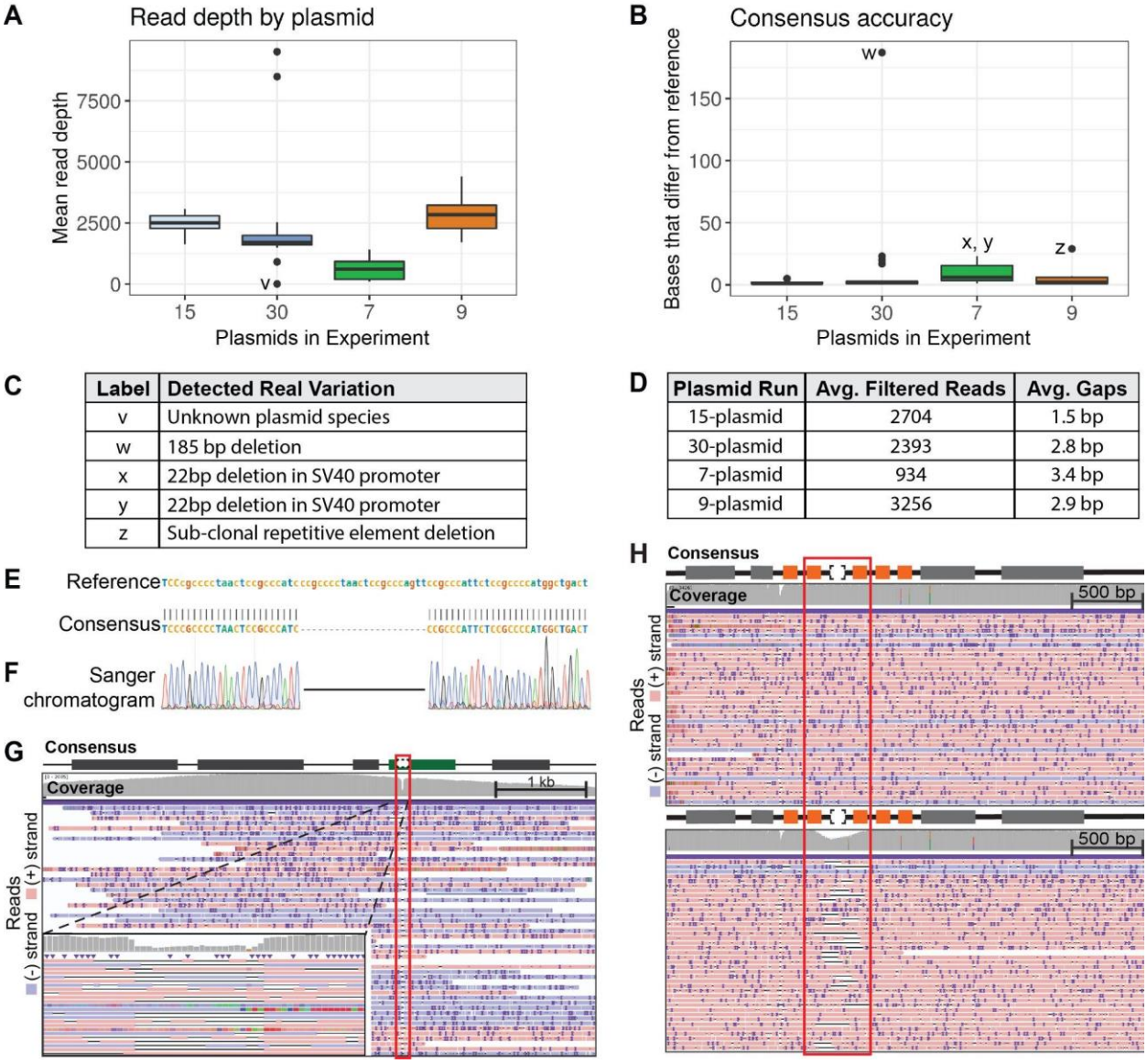
3.3.4 Nanopore Plasmid Sequencing Reveals Mutations in Real Plasmid Data

To evaluate the performance of OnRamp with sequencing of real plasmids, we ran four separate plasmid pools containing plasmids of a variety of sequences, similarity levels, and sizes, using both the transposase and restriction based protocols, nanopore sequenced them, and analyzed them using OnRamp's pipeline. A 7-plasmid pool was prepared using the transposase from ONT's Rapid Sequencing Kit. This experiment generated 6539 reads which passed guppy's quality filtering, an average of 934 uniquely assigned reads per plasmid **(Figure 3.5a)** and a consensus accuracy of 3.4 gaps per plasmid on average **(Figure 3.5 b, d)**, as measured by per-base differences in consensus vs reference.

The high read coverage and read length generated allowed us to distinguish reads and generate consensus sequences from three highly similar plasmids in this run that differed only by two 4bp sequences. Additionally, real sequence variation was detected in this run (not included in the per-plasmid gap average). A 22bp deletion, too small for detection by diagnostic digest & gel electrophoresis, was identified in the SV40 promoter of two plasmids **(Figure 3.5 c, e)** and validated by Sanger **(Figure 3.5 e-g)**. This deletion occurred outside of a region manipulated by molecular cloning and would not normally have been checked and caught by Sanger sequencing.

The 9-, 15-, and 30-plasmid pools were prepared using the restriction digest & ligation method. In the 9-plasmid pool, plasmids had over 1000 reads per plasmid, with on average 3256 quality filtered reads per plasmid, and an average of 2.9 gaps per plasmid consensus **(Figure 3.5 a-d)**. As in the simulated data, some of these gaps were from homopolymer errors and likely related to known issues with correctly calling homopolymers in ONT data (see Discussion). In this run, we were able to sequence through 4x and 6x 40bp repeats in 6 of the plasmids that were previously intractable to Sanger sequencing due to high secondary structure. The high read coverage obtained on these runs allows us to identify sub-clonal populations in plasmid sequences, which can occur as a result of plasmid recombination during bacterial growth. We detected a sub-population level deletion of one of these repeats in a plasmid from this run using OnRamp **(Figure 3.5 c, h)**. This high coverage is reflected even in the 30-plasmid pool, which averaged 2393 quality filtered reads per plasmid, and minimum of 900 reads per

plasmid, generating base-pair resolution for all but one plasmid, which had previously failed a diagnostic restriction digest check. This run had 2.8 gaps per plasmid on average, excluding a 185bp deletion detected in one plasmid **(Figure 3.5 b, c)**.



Label	Detected Real Variation
v	Unknown plasmid species
w	185 bp deletion
x	22bp deletion in SV40 promoter
y	22bp deletion in SV40 promoter
z	Sub-clonal repetitive element deletion

Plasmid Run	Avg. Filtered Reads	Avg. Gaps
15-plasmid	2704	1.5 bp
30-plasmid	2393	2.8 bp
7-plasmid	934	3.4 bp
9-plasmid	3256	2.9 bp

Figure 3.5 Real plasmid sequencing experiment characteristics and variant detection
a. Per-plasmid read depth across pooled sequencing runs. **b.** Per-plasmid count of bases in consensus sequence that differ from reference (gaps). **c.** Table describing variation contributing to outliers (labeled points v-z in panels a and b). **d.** Table summarizing read and gap data for experiments shown in a and b (gap counts do not include variants listed in **c**). **e.** OnRamp alignment results showing a 22bp deletion in consensus sequence relative to reference **f.** Sanger sequencing traces showing validation of deletion in (**e**). Traces were continuous across the gap, separated to show deletion (line). **g.** IGV browser view of individual nanopore reads mapping to deletion (red outline) from (**e**) in an SV40 promoter (green box). Left inset: zoomed view. Horizontal black lines are deletions **h.** IGV view of reads mapping to a clone without (top) or a clone with (bottom) a sub-clonal repetitive element (orange boxes) deletion (red outline). IGV: Black lines are deletions; dark purple marks are insertions.

3.3.5 Validating Plasmid Sequences in Pooled Plasmid Clones

The 9- and 15-plasmid runs described above all contained some plasmids that were clonal copies of each other. Normally, as a result of plasmid pooling, reads originating from different clones of the same plasmid (or highly similar plasmids) would all map to the same reference, making differentiation of the read source impossible without barcoding. However, we were able to successfully leverage nanopore's long read length in a simple alternative restriction-based protocol to differentiate reads originating from identical clones in the same pool without the need for barcoding. For plasmid libraries containing multiple plasmid clones or highly similar plasmids (≤24 bp difference), each clone is cut with a different unique restriction enzyme from its matched partners prior to pooling **(Figure 3.6a)**. During analysis, a copy of the same plasmid reference sequence is provided for each clone, except with the linear sequence origin set at the digest site used for that clone (termed 'rotated' reference). While each cut clone contains the same total sequence, the alternate digest sites create linear fragments (reads) that map precisely to their matched 'cut' reference sequence, but poorly to the same sequence reference 'cut' at any other site (**Figure 3.6a**). This approach is feasible due to the long-read nature of nanopore sequencing, where the majority of reads span an entire plasmid.

We validated this approach experimentally using three ~6.5kb plasmids **(1,2 and 3 in Figure 3.6b)** which are identical except for a ~500bp region. Five clones of each of the three plasmids **(a-e in Figure 3.6b)** were digested using different restriction enzymes for clones within a set (with the closest cut sites 579bp apart) pooled, prepared, and sequenced. An average of 2704 reads uniquely mapped when using rotated references **(Figure 3.6b),** compared to 7 reads uniquely mapping to non-rotated references, indicating that using different cut locations with clones is sufficient to create reads that align uniquely to their matched rotated reference. The 9-plasmid run contained 3 sets of clones and one unique plasmid **(Figure 3.6c).** The sub-population level deletion of a repetitive element discussed above **(Figure 3.6g)** was detected in a plasmid that was part of a set of three clones in this run.

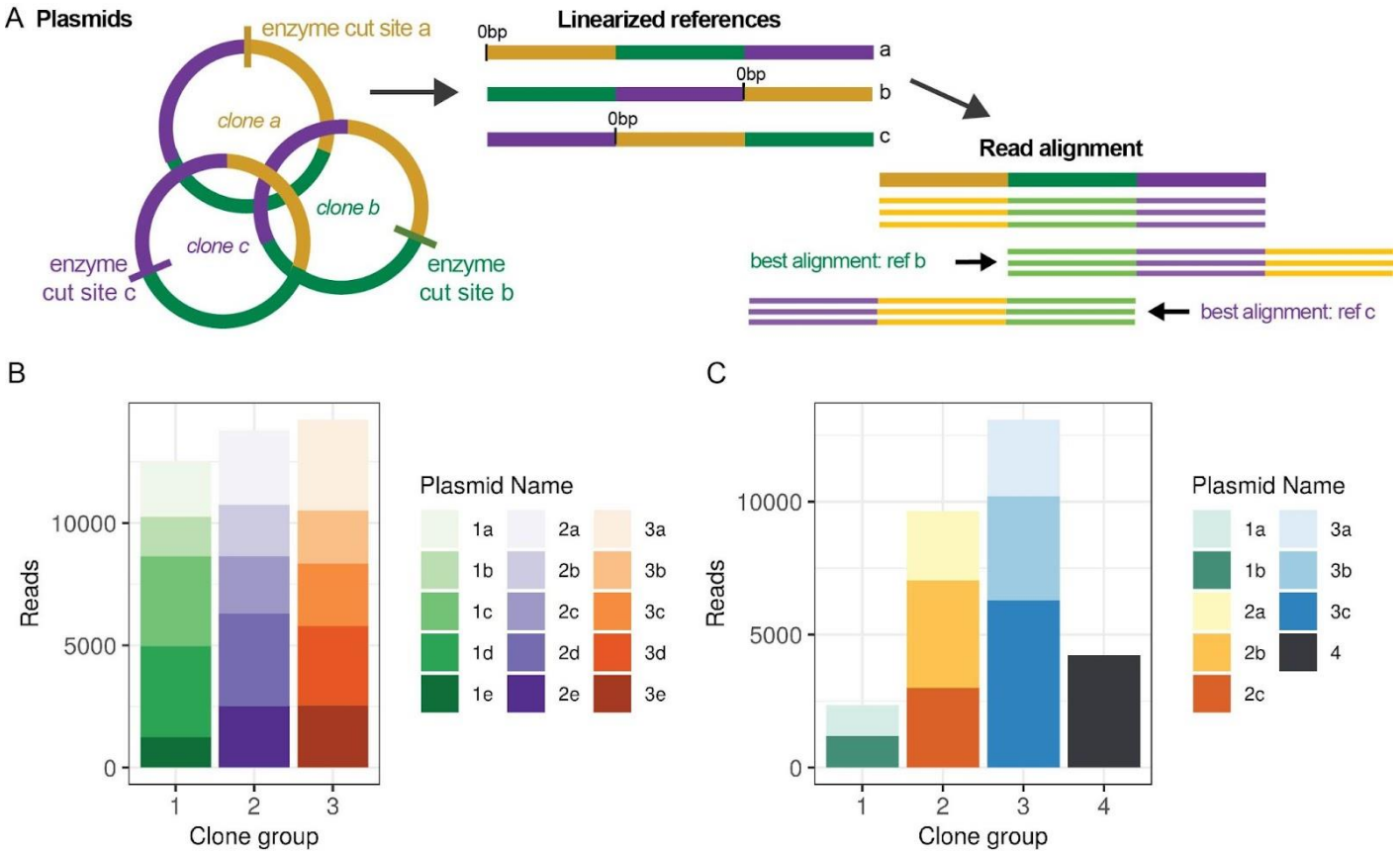


Figure 3.6 Restriction-digest barcoding for highly similar and clonal plasmids
a. Diagram of restriction cut-site method for unique read mapping of clonal plasmids using 'rotated' references. **b.** Number of reads mapping uniquely to each plasmid in a 15-plasmid clonal test pool. **c.** Number of reads mapping uniquely to each reference in a 9-plasmid mixed clonal run.

3.4 Discussion

3.4.1 Advantages of Long-Read Plasmid Sequencing and OnRamp

Assessing recombinant plasmid sequence fidelity is an integral part of any molecular cloning workflow. While Sanger sequencing is an elegant, cost-effective method for low-throughput plasmid validation, it can be inadequate for multi- whole-plasmid sequencing and handles regions with complex secondary structure poorly. As a result, routine validation of full plasmid sequences after cloning is not standard practice outside fields like gene therapy, where rigorous validation is required. However, we are increasingly recognising the potential impacts that plasmid backbone sequences, structural elements, and bacterial sequences, which are not commonly sequence-validated and can vary widely between plasmids, have on regulation and activity of other plasmid components. This has implications especially for the replicability of results across plasmid-based studies (and between plasmids within the same study) of gene regulation and expression if mutations resulting from cloning or bacterial recombination, which can impact function, go undetected.

As an alternative, high-throughput next generation sequencing's run cost, equipment cost and sample coordination requirements make it inefficient for standard plasmid validation workflows outside of large plasmid libraries. Additionally, NGS requires amplification, and due to its short-read nature, cannot identify and correctly assign mutations outside unique regions in highly similar plasmid pools. With the introduction of Oxford Nanopore Technologies' sequencing platforms, sequencing of many plasmids in their entirety at high read depth is now possible.

The advent of benchtop long-read sequencing with nanopore technology provides an important step toward obtaining full-plasmid sequences by covering entire plasmid sequences with a single read. However, to take steps toward more widespread adoption of full-plasmid sequencing using nanopore as standard practice, accessibility is crucial to address. A primary aspect of accessibility is a lab's level of programming expertise. OnRamp addresses this by eliminating the need for users to have knowledge of coding when analyzing nanopore plasmid sequencing data. It is also tailored to be easily interpretable, using a custom quality score system to flag plasmid consensus

sequences based on fidelity to their reference, and providing multiple readouts in formats familiar to molecular biologists and compatible with existing plasmid annotation tools.

OnRamp's use of reference-based assembly also facilitates its accessibility. While some techniques have been published for recombinant plasmid verification using ONT, they rely on transposase barcoded libraries and *de novo* assembly to validate plasmid libraries [241]; [242] or are not feasible for sequencing many plasmids simultaneously at comparable costs to sanger [243]. Additionally, these approaches all require some degree of programming for use of their analysis pipelines. Tools are needed for analysis of nanopore-based plasmid sequencing data which are accessible to a broad spectrum of researchers, without the need for training in bioinformatics, and which facilitate interpretation of results, in order for full-plasmid nanopore sequencing to become more widespread as an option for validation.

Using a reference-based system allows us to provide a consensus-reference alignment and alignment quality score to users to facilitate easy identification of mutations and simplify validation. It also allows us to integrate with IGV to provide a read-level view, which is important as this can provide additional information on plasmid sub-population level variation or allow for human interpretation of borderline error calls, similar to checking Sanger trace files for overlapping peaks.

Additionally, ONT's compact benchtop sequencing platform is much more affordable than most NGS sequencing platforms and allows for in-lab sequencing with results available as soon as next-day, without the need to ship samples to a core or company. OnRamp provides a rapid (0.5-2 hours for preparation, 16-36 hours for results) and cost-effective approach for medium-throughput plasmid sequencing. Using the reagents and protocols described here, we were able to fully sequence 30 plasmids at $1.25 per kb, significantly less than the cost of using Sanger sequencing to obtain equivalent data. Additionally, the OnRamp web app facilitates analysis and data interpretation in a manner that is accessible to labs without the need for extensive bioinformatics support.

3.4.2 Summary of Results

Testing OnRamp using simulated read libraries demonstrated its ability to correctly assign sequencing reads to reference sequences and construct consensus sequences even with highly similar plasmids (only 24bp difference). Testing OnRamp on real plasmid runs showed that OnRamp provides high sequence read depth across plasmid pools, generating consensus sequences spanning entire plasmid lengths at base pair resolution (even on Flongle flow cells with as few as 20 pores). The read depth we obtained (over 900 reads per plasmid for a 30-plasmid pool) using even ONT's lowest-capacity flow cell (the Flongle), allows for high confidence in base-level calls in consensus sequences despite nanopore's 10% error rate [248].

The high coverage we obtained also allows for detection of sub-population level variation even in a region with complex secondary structure and high clonal similarity **(Figure 3.5H)**. These mixed populations will not be represented in consensus files, however they can be detected by viewing the read alignments in IGV, which are provided as part of OnRamp's output. This also allows users to interrogate underlying read data to determine confidence in consensus sequence base and indel calls should they wish. This is similar to Sanger sequencing results, where sequence files do not show sub-population structure, but trace files might. Additionally, this mutation (a deletion of one of a 6x repetitive element set) likely occurred as a result of bacterial recombination after cloning, underscoring the importance of obtaining full-plasmid sequencing data rather than running spot-check validations using Sanger. These experiments also revealed a 22bp mutation in a functional non-coding plasmid element (SV40 promoter) which was previously undetected by diagnostic restriction digests, showcasing the ability of the tool to determine uncharacterized structural and sequence variation.

3.4.3 Limitations

A limitation of Sanger sequencing is the tendency for indels to occur after homopolymer sequences (sequences with repeats of the same base) [249]. While this type of error was also detected in our simulated and real plasmid data, consistent with reports of these errors in ONT sequencing data [247], ONT have worked to address this

issue. They have improved homopolymer sensitivity with their new R10 pore chemistry, which should reduce the rates of errors in homopolymers up to 10bp in length [250].

Using OnRamp's medaka mode to generate consensus sequences, we were able to rapidly validate our plasmids based on alignments to reference sequences. Some limitations of this approach arise as a result of medaka being a reference-based approach, as opposed to an assembly-based method. OnRamp uses a reference-based system for analysis, as it is a tool designed specifically for routine plasmid validation. While this precludes use of OnRamp for *de novo* assembly of unknown plasmid sequences (an uncommon case in routine screening), there already exist well-designed tools for *de novo* assembly of nanopore data [241]; [242]; [243]. For instance, while we were able to detect most variation in our constructs, consensus sequences for plasmids with very large indels (>1000bp) or where large portions of the plasmid have inserted backwards relative to the reference, could not be generated. However, these large rearrangements should be easily detected by complementary diagnostic restriction digest tests, which are often a routine step in cloning protocols. Using the alternate biobin mode, choosing unique regions in the reference is essential to binning reads. Indels in the unique portion of the reference can lead to incorrectly binned reads or failure to generate a consensus. An alternative method is to use the clonal restriction-based method we described to separate reads from highly similar or even identical plasmids.

3.4.4 Future Directions

We designed OnRamp specifically to make reference-based full-plasmid sequence validation rapid, affordable, and widely accessible to a variety of labs in order to facilitate standardization of routine full-plasmid validation. Other potential applications include improving understanding of the rates of bacterial recombination-based errors, especially in repetitive sequences, during plasmid production. These errors have been observed in our lab, even within *Escherichia coli* (*E. coli)* strains designed for direct repeat stability. Quantifying recombination frequencies and sequence-dependencies across bacterial strains using engineered plasmid models of highly repetitive regions would be facilitated by OnRamp's reference-based alignments and the restriction-based

protocol. This provides high numbers of full-plasmid fragments which would be needed to quantify infrequent sub-population-level events. This could allow for the design of plasmids that retain repetitive sites while using less recombination-prone variants.

The ability to sequence reliably and affordably through repetitive regions using long-read technology will also facilitate our ability to model, clone, and study the functional impacts of repetitive sequences in plasmids, where sequences are easier to manipulate on a per-base and high-throughput scale, due to our ability to reliably sequence-validate these constructs. This could have the potential to rapidly advance our understanding of the genomic role of repetitive sequences.

Finally, should the availability of an accessible reference-based tool for nanopore-based plasmid sequencing facilitate full-plasmid validation across many labs, studies might be made possible to improve the reproducibility of research on regulatory element activity. Given the variety of different plasmid backbones used for reporter assays, by completing large-scale analysis of plasmid backbone content vs element activity, it might be possible to determine which sequences contribute to variability across studies of the same elements, and provide a model for plasmid designs which improve reproducibility.

In summary, OnRamp offers rapid, medium-throughput full-plasmid sequencing without secondary structure limitations or the need for primers. It provides more affordability and simplicity than NGS, and with our streamlined web application, makes analysis and interpretation of results accessible and straightforward.

3.5 Methods

3.5.1 Vector Construction and Maintenance

Plasmids were constructed using either EMMA [161] or gateway- or restriction-based cloning methods. The EMMA toolkit was a gift from Yizhi Cai (Addgene kit # 1000000119). Various parts from the toolkit were used for construction of the vectors, and mCherry was cloned from pHR-SFFV-KRAB-dCas9-P2A-mCherry to become a usable part. pHR-SFFV-KRAB-dCas9-P2A-mCherry was a gift from Jonathan Weissman (Addgene plasmid # 60954 ; http://n2t.net/addgene:60954 ;

RRID:Addgene_60954) [251]. Expression vectors were grown in either Stbl3 or DH5α chemically competent *E. coli* strains.

3.5.2 Transposase-Based Plasmid Preparation

For transposase-based preparation, plasmids were treated using the Rapid sequencing kit and following ONT's protocol (ONT, SQK-RAD004). Pooled plasmid DNA is brought to 7.5uL using H2O and combined with 2.5uL FRA, and incubated 30°C for 1 minute and then at 80°C for 1 minute then put on ice. 1uL of RAP is added and mixed by flicking, then spun down and incubated for 5 minutes at room temperature. DNA is loaded onto a primed flow cell.

3.5.3 Plasmid Pool Linearization by Restriction Digest & End-Repair

Plasmid DNA was isolated using the QIAprep Spin Miniprep Kit following the manufacturer's protocol (QIAGEN, 27104) and eluted in water. Plasmids were linearized by restriction digest using a unique cut site, with times, temperatures and reaction volumes varied for other enzymes according to NEB recommendations. Example pooled restriction digest: NEB Buffer 3.1 (NEB,B7203S) was added to 1X and the final volume was adjusted with nuclease free water to 200uL. SwaI (NEB, R0604L) was added according to the total amount of DNA present for linearization (minimum 10 Units enzyme per 1 ug DNA), and the sample was digested at 25°C for 30 minutes. Plasmid pools were generated prior to digest if all contained the same unique restriction site, or after digest for plasmid pools where each plasmid required a different restriction enzyme. For plasmids where different restriction enzymes are used on each plasmid, heat-inactivation of each enzyme (following manufacturer instructions) or if not possible, column cleanup (Qiaquick PCR purification kit, QIAGEN, 28104) to remove enzyme was done and is a crucial step prior to pooling to prevent cross-cutting of other plasmids in the pool after combination by still-active enzymes.

Digested plasmids were diluted and pooled into a single 1.5mL microcentrifuge tube using the following rules to calculate desired amount of each plasmid: 1. using an equimolar amount of each plasmid, 2. a maximum of 1000ng total plasmid for the entire pool 3. using at least 10ng of each plasmid and 4. a total of 50uL volume. Amount of

each plasmid in a pool ranged from 15ng-100ng across experiments in this paper. If any digests generated 3' or 5' overhanging bases, pooled plasmids were end-repaired using 1uL (5U) DNA Polymerase I Klenow Fragment (NEB M0210S) with 33 μM each dNTP and 1x NEB CutSmart buffer per 1000ng DNA pool, with incubation for 15 minutes at 25°C, and heat inactivation for 20 minutes at 75°C. Following digestion and end repair, A-tailing was completed using 1uL of 10mM dATP and Taq DNA polymerase (NEB, M0273S) per 50uL of sample with incubation at 75°C for 15 minutes.

3.5.4 ONT Adaptor Ligation

For restriction-prepared enzymes, following DNA linearization, end-repair and A-tailing, ONT's ligation sequencing kit was used (ONT, SQK-LSK109) to add adaptors. One half volume of ligation buffer (4X T4 ligase buffer, 60% PEG 8000), 5uL of T4 DNA ligase (NEB, M0202M), and 2.5uL of AMX (ONT, SQK-LSK109) was added to the plasmid mixture then incubated on a tube rotator at room temperature for 10 minutes. One volume of 1X Tris-EDTA buffer (pH 7.5; Invitrogen, 15567027) and 0.3X room temperature SPRI beads (Beckman Coulter, B23317) were added for selection of >2 kb fragments. The sample-SPRI bead mix was incubated on a tube rotator for 10 minutes on the bench at room temperature. The SPRI beads were washed twice with 100uL of Long Fragment Buffer (LFB; ONT, SQK-LSK109) and the sample was eluted in 9uL of Elution Buffer (EB; ONT, SQK-LSK109).

3.5.5 Nanopore Sequencing

Flongle flow cells were loaded into minION sequencers containing Flongle adaptors from ONT. Flow cells were primed for the sequencing runs following ONT's standard protocol, using flow cell priming buffers provided by ONT. Briefly, flow cells are quality-control checked for a usable number of active pores (~ 0.5-1 pores per plasmid was used here as the minimum). Flow cell was washed with FB then SQB buffer mixed 1:1 with water. DNA prepared from previous steps is mixed with SQB and LB immediately prior to loading following ONT's protocols.

3.5.6 Simulated Reads

NanoSim was used to construct pooled plasmid read libraries. First, a model was created using 81,070 reads (N50=6,003bp) from a previous plasmid sequencing experiment, and the 30 plasmid sequences (average length = 4,318.7bp) were used as the reference genome and input in the characterization set. This model was then used to simulate reads from other plasmid references and from references constructed with 1, 10, 100, and 1000bp deletions and insertions of random sequence as well as plasmids with 6, 12, 24bp unique regions.

3.5.7 Bioinformatics Pipeline

Basecalling was completed using Guppy (Oxford Nanopore Technologies, 4.5.2) using the dna_r9.4.1_450bps_hac.cfg configuration and passing reads (Q >= 7) were filtered using Guppy or NanoFilt [252]. Adapters were trimmed using Porechop (https://github.com/rrwick/Porechop). The OnRamp webapp allows users to use Porechop and NanoFilt to trim reads and filter by their chosen q score and read length. Reference sequences were generated using SnapGene (https://www.snapgene.com/). The reads and references were then used as input for OnRamp during pipeline testing and development. The OnRamp pipeline and web tool are then run in either medaka, or binning mode, as detailed below.

The medaka mode uses ONT's medaka (https://github.com/nanoporetech/medaka, version 1.4.4) to create consensus sequences and should be used for mixed plasmid pools or for clonal pools prepared by restriction enzyme digest (detailed under 'Validating plasmid sequences in pooled plasmid clones'). The medaka consensus module was utilized to generate consensus sequences from read pileups using the '-g' flag to stop filling in gaps with draft/reference sequence during consensus stitching.

The binning mode is used for very highly similar sequences, such as those with a small unique identifier. The biobin module mode of plasmid sequencing was used to bin reads based on unique sequences in the provided references. The biobin mode/module searches the reference sequences for unique sequences longer than 3bp and a set is constructed for each plasmid reference. Each input read was then aligned to these

regions using Biopython pairwise aligner with alignment parameters match: 3, mismatch: -6, open_gap: -10, extend: -5. Reads were first aligned to an extended portion of the plasmid containing 20bp flanking the unique region and assessed using the ‘context score’. For reads that passed this threshold, the aligned portion was then aligned and scored against the exact unique region and high scoring reads (fine score > 80) were assigned to the plasmids. Each of the resulting bins was then passed to medaka for consensus polishing.

The resulting alignments are then filtered (MAPQ >= 10) for visualization using the Integrative Genomics Viewer (IGV) [245]. Final pairwise alignments were constructed between the reference and consensus sequences generated by medaka using EMBOSS needle (EMBOSS:6.6.0.0).

3.6 Data Accessibility

The OnRamp is available through a web app at https://onramp.boylelab.org/. The command line version and pipeline used for the application are available at https://github.com/Boyle-Lab/bulk_plasmid_seq_web and https://github.com/crmumm/bulkPlasmidSeq. All plasmid read data generated in this study has been submitted to the Zenodo database under accession number PRJNA123456.

3.7 Notes and Acknowledgements

This work is available as a preprint manuscript on *bioRxiv*, initially posted on March 15th, 2022: Mumm, C. & **Drexel (Englund), M**., McDonald, T.L., Diehl, A.G., Switzenberg, J.A., Boyle, A.P. On-Ramp: a tool for rapid, multiplexed validation of plasmids using nanopore sequencing. https://www.biorxiv.org/content/10.1101/2022.03.15.484480v2.

M.D. and T.M. were supported by NIH Training Grant Michigan Predoctoral Training in Genetics (T32GM007544) C.M was supported by University of Michigan Genome Science Training Program (T32HG000040). Thanks to Torrin McDonald for jump-starting the project and establishing Nanopore sequencing in the wet lab and welcoming me onboard. Thanks to Camille Mumm for performing data analysis, and

then patiently explaining the bioinformatics side of the project to me so that I could write accurately about it. Adam Diehl has been responsible for regular improvements and updates to the OnRamp website. Thanks to our reviewers who helped us improve the manuscript.

A.P.B., M.E, C.M, and T.M. conceived the project. C.M and A.G.D. developed OnRamp pipeline and web app. C.M performed data analysis. All authors guided the experiment design and data analysis strategy (C.M conceived and performed simulated experiments, T.M conceived and performed transposase *in vitro* experiments, M.D. conceived and performed *in vitro* clonal plasmid sequencing experiment, and J.S and M.E. ran the other restriction-based *in vitro* sequencing experiments). A.P.B. supervised the experiments, analysis, and data interpretation. M.E., C.M, and T.M. wrote the manuscript and all authors contributed edits and revisions. All authors read and approved the final manuscript.

CHAPTER IV

Characterization of *Cis*-Regulatory Activity in the Regulatory Domain Containing *PRDM1* and *ATG5* in Human Cells

4.1 Abstract

Studies characterizing single genomic domains in depth have been instrumental in illuminating the complex, multi-layer regulatory interactions that occur in a gene's native genomic context. While high-throughput regulatory assays can provide statistical support to trends in activity and mechanisms, they necessarily focus on a single isolated aspect of regulatory function at a time, where low-throughput assays provide higher fidelity and their scale allows for the testing of multiple different functions. However, both low- and high-throughput regulatory assays in humans have largely focused on the functions of promoters and enhancers until very recently. Given our increasing understanding of the prevalence and importance of silencers and enhancer blockers in the human genome, studies are needed which incorporate negative regulatory element functions into our model of the dynamics of gene regulation. Here I use a region of the human genome containing the genes *PRDM1* and *ATG5*, which have important, but differing roles in cellular biology and function, to investigate combinatorial regulatory interactions at the level of regulatory domains, genes and elements across two different cell lines. I test putative *cis*-regulatory elements in the region in an unbiased manner for positive or negative regulatory activity and contextualize this activity to genomic function using histone modification and transcription-factor binding data. I present a model for the differential regulation of the two genes where cell-type specific expression of one gene is modulated by a large number of weak enhancers, and the ubiquitous expression of the other is controlled by a small number of strong enhancers. I support the mechanisms of tissue-specific silencing of *PRDM1* as being related to heterochromatin spreading due to loss of an enhancer blocker, rather than active silencer activity, and I reveal position-dependent patterns of CRE additivity and synergy

related to HS2 enhancer function. Finally, I characterize the internal regulatory logic of the strongest enhancer in the region, DHS 16, a candidate *ATG5* enhancer.

4.2 Background

4.2.1 Low-Throughput Multi-Element Studies Elucidate Complex Regulatory Dynamics

The majority of *cis*-regulatory element reporter assays, especially in humans, have focused primarily on mapping and characterizing a single class of CRE, usually enhancers or promoters [94]. Yet silencers and enhancer blockers are both known to play important roles in transcription regulation [34], are distributed throughout the human genome [85], and contribute to cell-type specific regulatory dynamics [83]. A gene's regulatory state is a result of the contributions and interactions of multiple *cis*-regulatory elements combined. Thus, in order to generate a complete and accurate model of a gene's regulation, testing for the possibility of both positive and negative regulatory elements controlling expression is essential.

Additionally, *cis*-regulatory element research is increasingly focusing on high-throughput studies using massively parallel reporter assays (MPRA) [122,123,253]. These studies allow for the characterization of hundreds to thousands of elements simultaneously, which provides a number of advantages. For example, the cultivation of many putative elements allows for determination of statistically significant trends and similarities across each element type, which can contribute to a better understanding of mechanism [120]. They also allow for rapid characterization of elements across many cell and tissue types, supporting studies of differences in tissue-specific CRE behavior and evolutionary trends [254,255]. However, high-throughput approaches also have significant disadvantages. These assays often have higher false-positive or false-negative rates and are more variable in replicability of effect strength than their low-throughput counterparts, depending on assay design or readout [91]. The primary concern however, is that while high-throughput assays are excellent at testing many elements at once, they do not generally provide context for the elements' activities. Without follow-up studies to relate a CRE element's MPRA activity to a specific genomic context, tissue type and gene, related transcription factors, histone marks, and/or to

coordinated activity of other elements in the region, namely without the addition of functional context, the utility of MPRA data is limited primarily to generalized genome-wide pattern or mechanistic discovery.

Low-throughput approaches allow for greater in-depth focus on a smaller number of regulatory elements and provide higher-fidelity results. Studies characterizing single genomic regions using low-throughput approaches have been powerful tools as models for many aspects of CRE activity, mechanism, and interaction in a genomic context. Examples include the β-globin locus [256], the *H19/IGF2* region [257], and the *SHH* locus in *Drosophila* [258]. Once these models were established, successive studies were able to build upon initial discoveries to create working models of genomic regulatory activity and its complex dynamics. They are also better suited to the study of multiple element types due to the cost and effort limitations of building and analyzing data from high-throughput libraries for even a single element type. Low-throughput, multi-element studies, whether the elements are of the same or different classes, are both well-suited and crucial to answering a number of outstanding questions about the combinatorial dynamics of gene regulation, discussed below.

4.2.2 CREs Coordinate to Determine Expression Patterns within a Cell and Across Cell Types

Multi-element interaction and coordination is a well-established phenomenon among positive regulatory elements. A single enhancer can regulate multiple genes, promoters can regulate other promoters [259], and a promoter can be regulated by multiple enhancers [260]. Within the set of enhancers regulating a single gene, some may be redundant and drive similar expression patterns as others in the set [134,261]. In other cases, the activity of multiple enhancers may be additive, or they may behave synergistically and combine to produce expression greater than the sum of their individual enhancing activities [262,263]. Predicting which elements will behave redundantly, which additively, and when, is not yet possible as the precise mechanisms underlying these behaviors are not fully understood.

When adding negative regulatory activity, the complexity and number of potential models for gene regulation increases. At the level of a single cell type and state, do

silencers and enhancers act additively to produce moderate levels of expression through simultaneous action on the same gene, as seems to be the case in episomal studies [89], or are silencer and enhancer activation mutually exclusive in the genome? During differentiation or in response to signaling, are genes inactivated by the loss of enhancer activity, the gain of silencer activity or both? A gene that is expressed in one cell state or type could be repressed in another, not only through inactivation of enhancers, but by activation of a silencer, or spreading of repressive chromatin marks through the inactivation of an insulator. Which of these mechanisms is used and what determines in which situations one mechanism is used or another is still unknown. These questions are additionally complicated by the preponderance of evidence for the existence of dual enhancer-silencer CREs [83,84]. This suggests that in some cases, conversion of an already active element through binding of a different class of transcription factors, not inactivation/activation of a different element, could be sufficient.

4.2.3 Differential Regulation of Genes within the Same Regulatory Domain

To study the set of CRE interactions that regulate a gene, it is important to determine which elements comprise that set. This is accomplished using a number of different complementary approaches, including functional methods such as detection of CRE-gene co-accessibility [71], co-expression of eRNA and mRNA [72], genomic deletions paired with measures of expression changes, and localization methods such as CRE-gene proximity or presence within the same chromatin loop or insulator boundaries, corresponding with a topologically associating domain (TAD) [143]. TADs are regions with a higher level of physical contact between elements inside the region than with areas sequences outside the domain, as defined by interaction frequency measured by high-throughput chromatin capture (Hi-C) methods [264]. TAD boundaries are often enriched for CTCF and cohesin binding sites, supporting cohesin-mediated looping as the mechanism for their formation, which increases intra-TAD element proximity and decreases inter-TAD element contact [143,265].

TADs are important for defining a gene or set of genes' regulatory domain, as in many cases they correlate with boundaries of histone modifications [266,267], indicating they may represent the mechanism by which genomic structure and function are related

[268]. TADs also overlap with domains of transcriptional activity [146], so it is to be expected that genes within the same TAD would have similar expression levels or patterns. Genome-wide expression profiling across differentiating cells did in fact demonstrate expression correlation for genes within the same TAD [266,269]. Additionally, in enhancer trap assays, reporters that inserted within the same domain usually had similar expression patterns [270]. Yet this is not always the case, as in a number of studies, TAD deletion or rearrangement had small effects, no effect, or gene-specific effects on expression within the TAD [271,272]. What mechanisms cause these differential patterns across genes when a TAD structure is disrupted, and in cases where genes are normally differentially expressed within a TAD, what mechanisms regulate this are important questions. Answering these questions also contributes to a better understanding of both the role of TADs and CTCF boundaries and the interaction of different regulatory processes in determining gene expression. Are these effects mediated by enhancer-promoter distance or specificity [113], or by the presence of additional enhancer blocker sites within a TAD? Or does this represent a potential role for silencers with gene-specific and cell-type-specific activity, which allows for TAD-level activation of multiple genes simultaneously across most cell types, except for those in which the silencer is active to down-regulate one particular gene? Additional low-throughput single-domain multi-element studies are needed to address these questions.

4.2.4 Transcription Factors Coordinate to Determine Expression Patterns of a Single CRE

Within a single element, some of the same principles of modularity, synergy, redundancy and combinatorial activity can apply. The necessity of multiple transcription factors (TFs) for activation allows for modulation of an element's activity using a more complex regulatory grammar, in order to precisely control timing, cell type and signal response-related expression. The grammar of a CRE is defined by the constraints of number, type, spacing, and orientation of the transcription factor binding sites that comprise the CRE [55,273]. Similar to regulatory units made of many CREs, there are examples that support a number of different models of how transcription factor regulatory logic works within a single element. In the billboard model, TFs bind relatively

independently and behave additively, with few constraints on their relative positioning [274]. The TF collective model incorporates the possibility of more interdependence, and included both TF-DNA and TF-TF binding [275]. Finally, ‘enhanceosome’ enhancers have the most strict limitations, where specific TF order, spacing, and motifs are all necessary for proper function [50]. There are examples of enhancers which correspond to each of these models, however it has been suggested that the most accurate model for most enhancers reflects various combinations of these models to varying degrees [51]. Identifying which CREs fall into which classes based on these models could help bridge the gap in our understanding of the connection between enhancer sequence and function. Additionally, the models discussed above are based primarily on enhancer behavior. More examples need to be explored to determine whether the same principles apply to silencer activity. These models might also help distinguish CREs that behave as silencers exclusively from those that are dual enhancer-silencers, by determining whether the two classes have different TF binding logic.

4.2.4 Integration of Regulatory Information Across Multiple Levels

In order to study the mechanisms of combinatorial element activity as they relate to expression dynamics, a model is needed that combines these multiple layers and levels of regulation to characterize the interactions and dependencies at both the domain, gene, and element levels in the same context. Here I present a study of *cis*-regulatory activity for a single regulatory domain in the human genome, in which I test for both positive and negative regulatory activity in an unbiased manner. I will present a model of the region which encompasses element activities, classes, positional dependencies, and additive potential. By studying these activities in two well-characterized human cell lines from two different tissue types, I am able to add multiple additional layers of information on these regions from past studies. Using these available datasets I characterize transcription factor binding, histone modifications, conservation and chromatin accessibility, to improve the depth of the model. I identify tissue-specific patterns of enhancer activity, characterize associations between this activity and chromatin state across the cell lines to build a model for gene regulation,

discuss variation in positional and synergistic effects of the DHS, and show how combinatorial interactions and context-dependence apply at all levels of regulation, including at the level of the regulatory domain, the individual element, and the transcription factors that drive activity.

4.3 Approach

4.3.1 Cell Lines

In order to be able to characterize not just individual regulatory elements, but changes in regulatory states of CREs as a result of differential regulation, I chose two cell lines which represent different tissue lineages for this study. K562 is a suspension lymphoblast cell line isolated from a female myelogenous leukemia patient. They are highly undifferentiated erythroleukemia cells which can be differentiated into erythrocytic, granulocytic and monocytic progenitors [276]. HepG2 is an epithelial liver cell line [277] isolated from a hepatocellular carcinoma of a male patient [278]. Importantly, both these cell lines are classified as ENCODE Tier 1 & 2 cell lines respectively, meaning that they were chosen by the ENCODE consortium to be prioritized for use in studies and so there exist many datasets which have already been generated for these cell lines. These include gene expression, transcription factor binding, histone modification, DNA accessibility, chromatin contact datasets and more. As a result, many additional layers of characterization were already available for the elements after they were classified through my regulatory assay testing. Additionally, these cell lines are both easily transfectable (K562 through electroporation and HepG2 via lipid-mediated transfection).

4.3.2 Regulatory Domain

The region of chromosome 6q21 in the human genome encompassing the genes *PRDM1* and *ATG5* (chr6:106,525,000-106,790,000) has multiple characteristics that make it an ideal model for the study of regulatory dynamics within a TAD and across cell lines. First, this 265kb region, encompassing both genes and +2kb upstream +15kb downstream of the gene ends, is contained within a single 1Mb TAD, as defined by HiC chromatin contact data, taken from Rao et. al 2014 [265], shown in **Figure 4.1**. In the

Figure, TADs are visualized as triangles of darker color, representing more frequent interactions between the regions under the triangle than between those regions and the adjacent genomic regions (outlined in black). The TAD encompasses a gene desert, containing only these two genes, which increases the likelihood that most of the CREs in the region regulate the genes in question. The genes are also 77kb apart, minimizing complexity compared to regions with multiple overlapping genes, and allowing me to interrogate whether or not proximity is a factor in which CREs regulate which gene. Reflecting past studies of TAD conservation [143], the larger TAD is conserved (present in multiple cell lines across diverse tissues, data not shown), but in K562 there is a sub-TAD that encompasses the two genes (light blue triangle in **Figure 4.1**) which is less conserved and lost in HepG2. This sub-TAD corresponds to the 265kb region that I refer to as the *PRDM1-ATG5* regulatory domain.

Additionally, indicators of regulatory function show variation between the two cell lines and across the region within a cell line. chromHMM states (a tool which uses a Hidden Markov model for computationally predicting biological regulatory states based on histone modifications and ChIP data [279], show (multi-colored bars in **Figure 4.1**) more potential functional enhancer (yellow and orange) and transcribed (green) sites in K562 and more repressed (light/dark grey) regions across the region in HepG2 **(Figure 4.1)**. Using these, one can see that the regions containing *ATG5* are very active in both cell lines, but the active region which encompasses *PRDM1* in K562 is lost in HepG2 (black boxes). chromHMM states are not sufficient for confirmation of activity but do suggest broader trends which indicate interesting regulatory dynamics for this region. These dynamics also correspond to differences in gene expression and chromatin accessibility **(see Figures 4.2 and 4.6e)**.

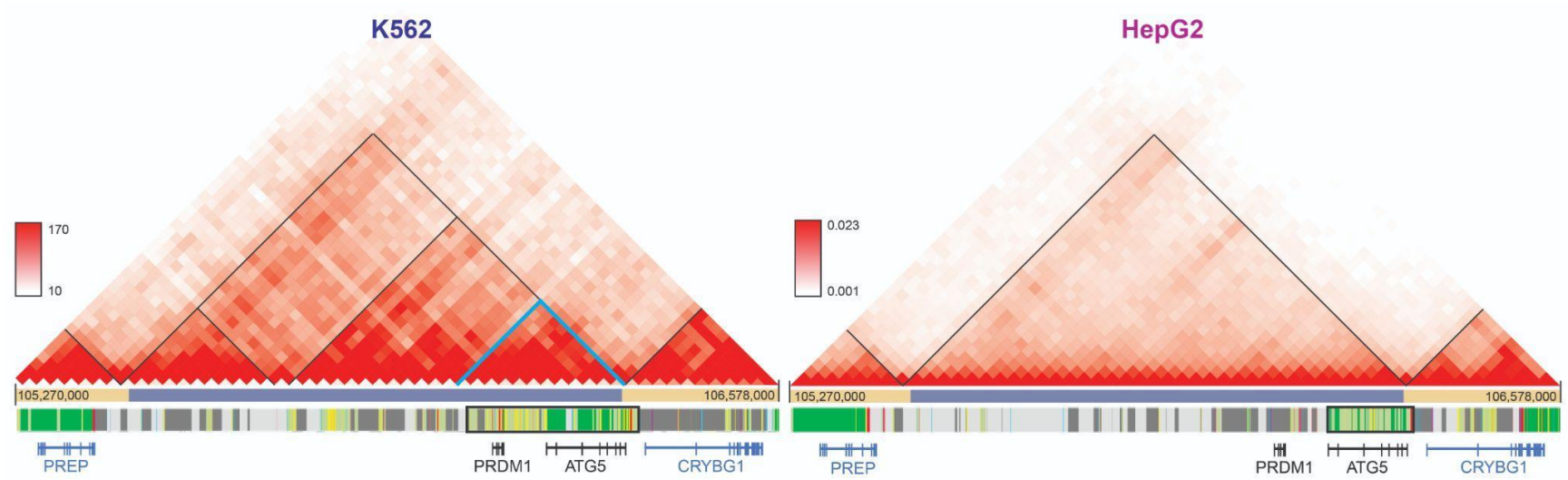


Figure 4.1 Overview of *PRDM1-ATG5* genomic regulatory structure

HiC maps generated using the 3DGenome viewer (http://3dgenome.fsm.northwestern.edu/), using datasets from Rao et. al 2014 for K562 (left) and HepG2 (right) for chr6:105270000-106578000 in hg19. TAD calls for the region from the Rao data are shown by blue/yellow bars under the HiC maps, and relative gene positions are at the bottom. chromHMM calls for this region (ENCODE Genome Segments track in UCSC browser) are the multi-colored bars just above gene indicators, shown for respective cell lines. Extent of *PRDM1/ATG5* active regions are indicated by black boxes. Signal key for HiC on left taken from 3DGenome viewer, darker red square indicates higher contact frequency.

4.3.3 *PRDM1* and *ATG5*

This region is also interesting for its dynamic expression patterns. *PRDM1* and *ATG5*, are both contained within the same TAD, but are differentially expressed within each cell line, and across the two cell lines **(Figure 4.2a, b).** This potentially allows us to address both the question of what regulatory mechanisms are responsible for generating differential expression levels within a TAD, and the question of what regulatory element dynamics are responsible for the differences in expression, and what the interplay or overlap is between the mechanisms driving intra-domain and inter-cell expression differences.

The gene ***PRDM1*** (Positive Regulatory Domain I-Binding Factor 1), the human homologue of mouse *BLIMP1* [280], is named for its discovery as a factor bound to a regulatory domain in the beta-interferon *IFNB1* promoter [281]. *PRDM1* is a DNA-binding transcriptional repressor containing a PR/SET domain and five zinc fingers which mediate DNA binding and recruit histone methyltransferases and deacetylases [282]. *PRDM1* is a member of the PRDM family of genes, which have a number of different roles in animal development [283]. *PRDM1* has a variety of functions across a range of tissues [284]. It is expressed in early heart development [285], and drives patterns of gene expression in epidermal (skin and hair) formation [286]. Studies in mouse show it is crucial to proper primordial germ cell specification [287] [288] and in zebrafish it is involved in neural crest specification [289].

Outside development, *PRDM1* plays a crucial role in innate and adaptive immune cell fate. *PRDM1* regulates cell function, maturation and proliferation in natural killer (NK) cells [290] [291] and macrophages [292]. Its role in regulating B- and T- cell function and differentiation has been well-characterized and is an area of ongoing interest [293–295]. *PRDM1* is also important for maintenance of resident NK- and T-cell populations in non-lymphoid tissues such as liver, kidney, skin and gut [296]. As expected from its various roles in immune cell development and regulation and its function as a negative regulator of p53 transcription [297], misregulation of *PRDM1* has been linked to poor outcomes in cancers such as lung cancer [298], in B-cell, NK-cell and T-cell lymphomas [299,300], and in pancreatic duct adenocarcinoma [301].

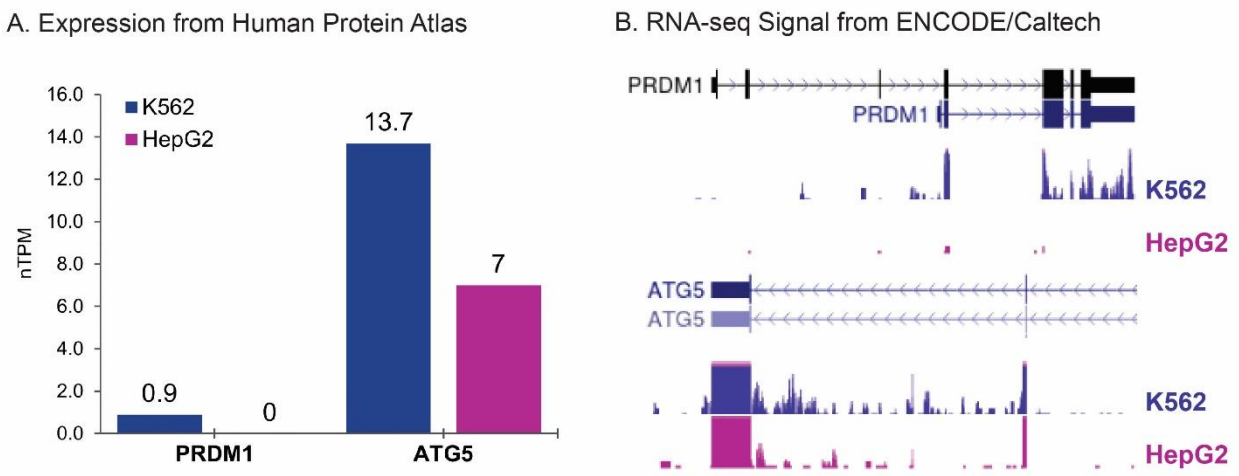


Figure 4.2 *PRDM1* and *ATG5* expression in K562 and HepG2 cells
a. Expression for *PRDM1* and *ATG5* in K562 (blue) and HepG2 (magenta) taken from the Human Protein Atlas database. Values represent TPM (transcripts per million) normalized by HPA for comparison across tissues (nTMP). **b.** RNA-seq paired 200bp signal for *PRDM1* and *ATG5* polyA-total cell RNA for HepG2 and K562, shown under their respective gene structure diagrams (only 3' end of *ATG5* shown). Visualized using UCSC genome browser. RNA signal peak heights are displayed on the same scale for *PRDM1* and *ATG5*.

ATG5 (Autophagy-Related 5) is a member of the family of autophagy-related genes [302]. Autophagy is a highly conserved cellular process involving the degradation and recycling of organelles and proteins which occurs as part of normal physiological processes, or in cases of cellular starvation or infection [303], and ATG5 is a key protein in this pathway [304] [305]. ATG5 binds with ATG12 to form a complex essential for creation of an autophagic membrane [306]. As a result, *ATG5* is an essential gene for cell function [307]. Mice which are *ATG5*-null only survive a day after birth [308], and *ATG5*-null mouse embryos fail to develop past the 4- to 8-cell stages (null oocytes fertilized with sperm wildtype for *ATG5* can develop) [309]. Only one genetically inherited pathogenic mutation has been reported for any of the ATG genes in humans - in *ATG5*. The mutation resulted in a hypomorphic allele which reduced its binding affinity for ATG12, and was associated with autosomal recessive spinocerebellar ataxia-25 [310]. The role of autophagy and *ATG5* in neural function is also supported by studies in *Drosophila*, where loss of *ATG5* causes motility defects and severe ataxia [310], and in mouse, where *ATG5* deficiency in neurons creates motor function deficits and neurodegeneration [311].

ATG5 is also involved in apoptosis through a cleaved form of the protein [312], and in regulation of lipid storage in hepatocytes [313]. Like *PRDM1*, it plays a role in

immune cell function in multiple ways, and is important for B- and T-cell development and proliferation [314] [315] and for MHC II presentation in dendritic cells [316]. More directly, the autophagic system itself can act as a part of the innate immune system [317]. Also similar to *PRDM1*, *ATG5* is implicated in a number of cancers, however its role is more complicated. Normal autophagy can act as a cell death mechanism, or allow cancer cells through increased autophagic activity to survive [318,319]. *ATG5* is downregulated in colorectal cancer [320], and has been studied in both K562 cells [321] [322] and HepG2 cells [323], in models of myelogenous leukemia and hepatocellular carcinoma, in both cases to study the role of autophagic processes in drug resistance/susceptibility.

Genome-wide association studies (GWAS) have associated SNPs in the *PRDM1-ATG5* region (between the two genes) with susceptibility to systemic lupus erythematosus (rs6568431, rs742108) [324] and rheumatoid arthritis (rs548234) [325]. Both of these diseases are auto-immune in nature, and as discussed, both *ATG5* and *PRDM1* have been linked to immune function.

ATG5 is ubiquitously expressed at similar levels across many tissue types, consistent with its role in basic cellular process, but *PRDM1* has a wider range of expression values and higher tissue-specificity, consistent with its roles in immune function and in development of specific tissue types. Given the importance of both of these genes for cellular function, it seems that mechanisms must be present in their genomic space which ensure precise, and separable, spatiotemporal control of expression in spite of their co-localization within a single domain of chromatin accessibility. Using these two genes as a model allows for interrogation of what those mechanisms are or are not.

4.3.4 Element Choice

The goal for this study was to test for the presence of all possible CRE element types in as unbiased a manner as possible, so as to include the potential for discovery of silencers and enhancer blockers as well as enhancers. While some combinations of histone marks and transcription factor binding can be predictive of enhancer activity, these trends are not always correctly predictive and may bias studies to focus only on

enhancers that follow ‘model’ characteristics, excluding interesting edge cases. For silencers, there is not as of yet a predictive set of chromatin marks or factors available despite attempts to find such predictive factors across a number of high-throughput studies [84,85]. Studies of enhancer blockers often focus on CTCF-binding sites but may in the process be selecting against the detection of other enhancer blocker types [108].

In order to choose regions with as little bias for element type as possible, I used only chromatin accessibility, across at least one of a number of cell lines (including but not exclusive to K562 and HepG2) as my metric for candidate region selection. Accessibility is an indicator of function, as active CREs are bound by various transcription factors which mediate their activity, and to bind DNA, these factors must displace nucleosomes to make binding sites accessible [97]. This results in regions of accessible chromatin which can be mapped using assays which leverage the inaccessibility of nuclease-wrapped chromatin to enzyme digestion or transposase activity. In DNase-seq, accessible DNA is digestible with DNaseI (DNaseI hypersensitive sites or DHS), while nucleosome-wrapped DNA is protected. Protected (non-cut) DNA fragments can be purified, sequenced, and mapped, then read pileup signal inverted to show where DNase was able to cut, creating peaks of signal [326]. I visualized existing DNase-seq datasets from ENCODE (Duke) for a number of cell lines using the UCSC Genome Browser to locate regions of chromatin accessibility within this region that fell outside gene promoters or exonic sequence.

Twenty regions in total were chosen for testing, encompassing 16 elements, with three split into multiple fragments where the region had alternate adjacent accessibility peaks across the cell lines. The size of the DHS regions varied (from about 500bp to 1100bp) as they were chosen to include at least the sequence underlying DNase-seq signal peak calls (shown as bars above peaks in figures, indicating regions with the most signal) and often a larger region to encompass the full accessible regions, which varied in size. **Figure 4.6c and 4.7b** show the genomic locations of the cloned regions, which were isolated from the genome using PCR (polymerase chain reaction) and cloned into an assay panel **(Figure 4.3)**. The regions tested include a number of, but not all of, the peaks in K562 and HepG2 in this region, as some regions were resistant

to cloning even after multiple attempts. Regions were chosen from the region stretching from just upstream of *PRDM1* to within the 3' end of *ATG5*. A number of regions were also included which are not in accessible chromatin in K562 or HepG2 but which were accessible in a number of other cell lines. The tested regions are referred from here on forward as DHS 1-34.

4.3.5 Assay Design

Each one of the DHS elements isolated by PCR was cloned into three different plasmid based CRE reporter assays, to be tested for enhancer, silencer, and enhancer blocker activity. This assay panel **(shown in Figure 4.3)** was built using a firefly luciferase gene for readout, an SV40 promoter (SV40p) [327], and for the silencer and enhancer blocker assays, the HS2 enhancer (HS2e), an erythroid-specific enhancer from the beta-globin locus (Tuan et al. 1989).

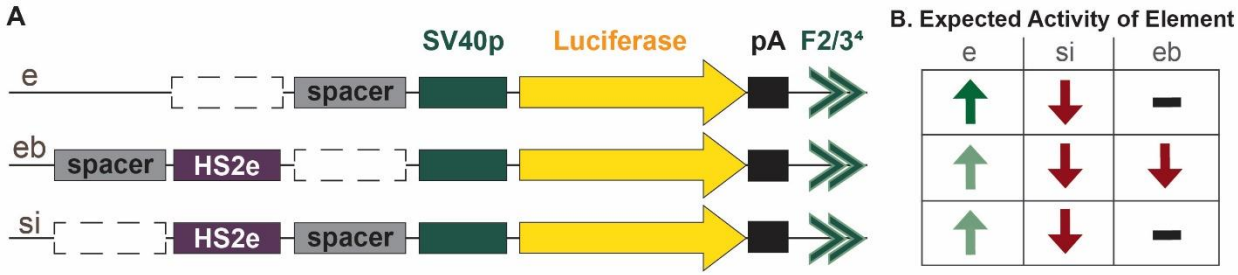


Figure 4.3 Enhancer, silencer, and enhancer blocker assay panel
a. Components of each assay e-enhancer, eb-enhancer blocker, si-silencer. Test element (or control r33 neutral sequence) insertion site shown as dashed box. Spacer is 750bp neutral sequence. **b.** Expected impact of inserting each element type (columns) in each assay (rows), relative to control expression (up, down or no change).

In this panel, DHS are classified as enhancers if they increase activity in the enhancer assay, silencers if they decrease expression in the enhancer or silencer assay alone or both the silencer and enhancer blocker assays, and enhancer blockers only if they decrease expression in the enhancer blocker position but *not* in the silencer position **(Figure 4.3b)**.

The assays also included a number of modifications beyond typical regulatory assay designs intended to control for potential confounding effects discussed in Chapter 2. In order to prevent HS2e upstream or looping bypass of any tested enhancer blockers, a fixed F2/3^4 enhancer blocker element (four copies of the core CTCF-binding

footprints from cHS4 which have eb but not insulating activity) was added 3' of the luciferase's polyA tail. This means that any DHS inserted into the enhancer blocker position would lead to the HS2e being flanked with enhancer blockers, generating a stronger effect.

Additionally, given the evidence for the potential impact of enhancer-promoter spacing on expression, and the wide range in length of my tested products, a 750bp (average size of my DHS regions) 'neutral' spacer sequence was inserted between the enhancer and promoter in the silencer assay. This element was tested functionally to establish that it did not alter expression upon addition to the assays in the panel (data not shown). This spacer is of importance primarily for enhancer blocker testing. Without the spacer, the HS2 enhancer is 50bp away from the SV40 promoter. Any element inserted in the enhancer blocker position would move the HS2e at least 500bp further away from the promoter (expression dropoff in my test and Davis et al. [137] showed a dropoff in expression starting around 150-250bp). This could result in a decrease in expression that mimics enhancer blocker activity as it would not be an effect present when the element was tested in the upstream position in the silencer assay, but would be unrelated to enhancer blocking activity. In order to be able to better compare expression and element activities across the assay panel, the fixed F2/3[4] enhancer blocker and the spacer sequence were added across all three assays, such that all assays have the exact same sequence content as much as possible.

The dashed boxes that show test element insertion positions are receiver sites for Gateway cloning. As demonstrated in Chapter 2, base plasmids with these sites do not make good controls. For all three assays, a control neutral sequence, the 750bp random r33 sequence tested in Chapter 2, was cloned via Gateway reactions into the test site. In all following figures and experiments, the 'e' 'si' and 'eb' controls used are these base assays with this r33 sequence inserted. For si and eb assays, the e control represents the 'low' expression control and is shown in red (SV40p levels of expression), and the respective eb/si control plasmids are the 'high' expression controls and are shown in green (SV40p and HS2e expression). DHS regions were similarly first cloned into pEntr1a Gateway donor vectors then inserted into all three assays by Gateway. See section 4.6.1 in Methods for details.

4.4 Results

4.4.1 Assay Panel Validation Using Positive Controls

I first tested my assay panel using known functional regulatory elements to establish that each assay can detect regulatory activity as expected and establish the ranges of expression of the assay within a relevant biological context. **Figure 4.4** shows the results of these tests in K562 cells. Two enhancers were tested in the enhancer assay: CMV (human cytomegalovirus), a ubiquitous strong enhancer [200], and HS2. The T39 silencer element, characterized in K562 by [81], shows 40% silencing activity in the silencer plasmid, but no change when placed downstream of the enhancer in the eb plasmid. This is consistent with previous reports of position-dependent silencer activity [89]. The F2/3^4 enhancer blocker shows strong activity reduction (69%) in the eb, but not the si assay, as expected.

4.4.2 Enhancer Assay: DHS Activity in K562 and HepG2 Cells

Candidate DHS (locations shown in **Figure 4.6c)** were first tested in the enhancer assay in K562 and HepG2 cells **(Figure 4.5 a & b)**. Interestingly, all tested regions increased expression to some degree in K562 cells except for DHS 9.2, and none showed silencing activity. In HepG2, only ten regions increased expression, and two showed potential silencing activity (DHS 11, 30 - patterned boxes). One reason for 19/20 regions in K562 showing increased expression could be some undetermined, non-specific effect of inserting sequences causing increased expression independent of enhancer function. If this is the case, this effect does not seem to be the same between the two cell lines. Regardless, in order to account for this possibility, we consider only regions which have expression ≥ 200% as putative enhancers, shown as grey bars in **Figure 4.5**. Using this cutoff, 14 regions act as weak enhancers in K562, and 6 in HepG2. The strongest enhancers for each set (above 300% in K562 and above 200% in HepG2) include DHS 1, 3, 10, and 16 in both cell lines, but the HepG2 set additionally includes DHS 19 & 25.1. DHS 19 and 25.1 are also the only regions to have higher expression in HepG2 than K562 and only DHS 13, 15.1 and 16 show the same expression levels across both assays.

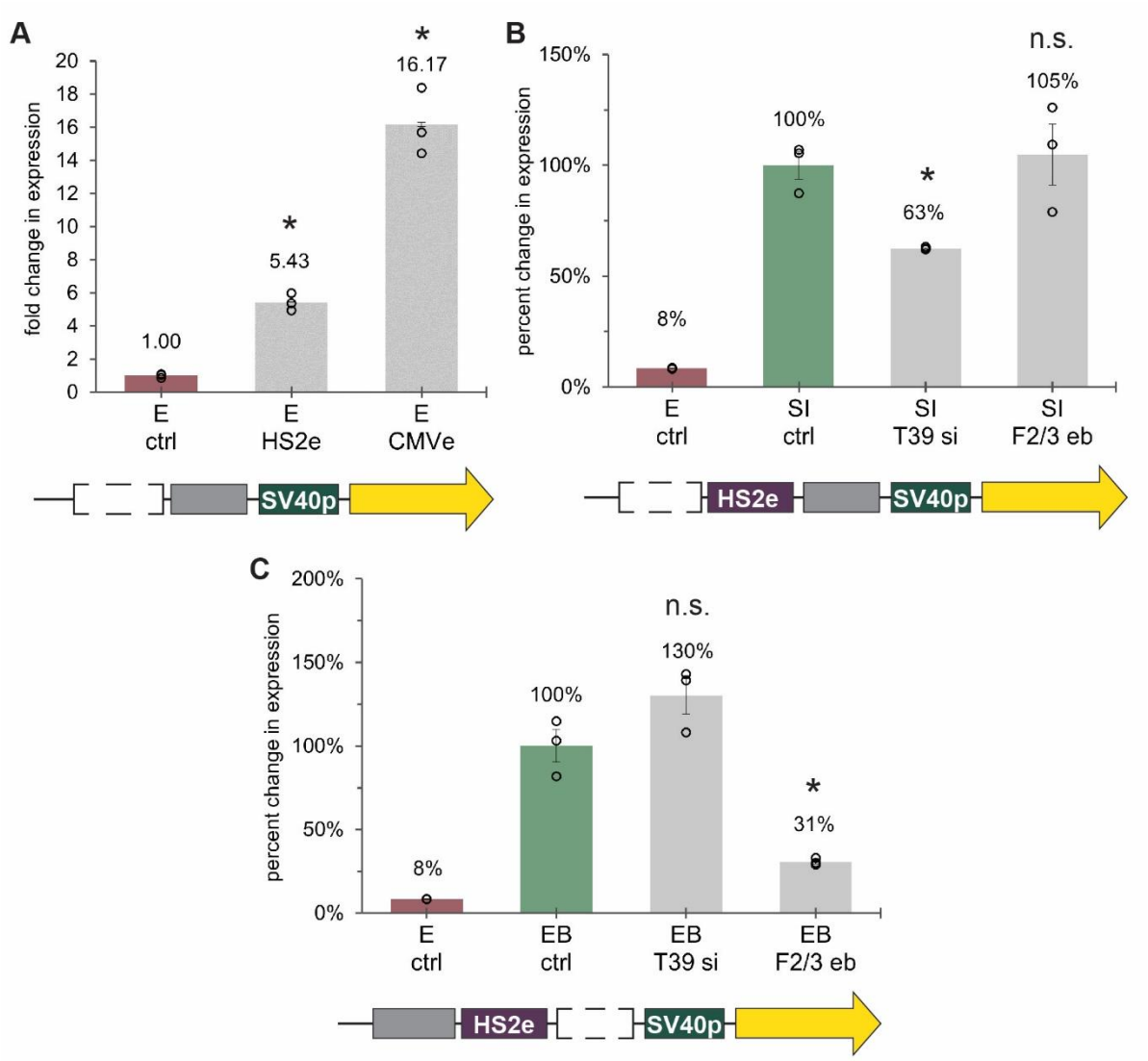


Figure 4.4 Validation of assay panel using positive controls

Fold or percent change in expression vs control plasmids for E, SI, and EB assays. Diagram of assay used shown below diagrams, with insert site shown as dashed box and fixed spacer fragment as grey box, luciferase in yellow. Assay plasmid identity on top line of X-axis labels in capitals, inserted element shown on second line, with element class in *lowercase. **a.** Fold change in luciferase activity over E assay control shown in red (neutral r33 sequence in test site-SV40p) of HS2 and CMV enhancers inserted in test site. **b.** Percent change in expression relative to SI control plasmid (green) of T39 silencer and F2/3[4] eb controls. Same E control from panel a in red. **c.** Percent change in expression relative to EB control plasmid (green) of T39 silencer and F2/3[4] eb controls red E control same as in a and b. T-test vs respective controls $p<0.05$ indicated by *. 3 biological replicates shown as open circles, error bars represent +/- standard error.

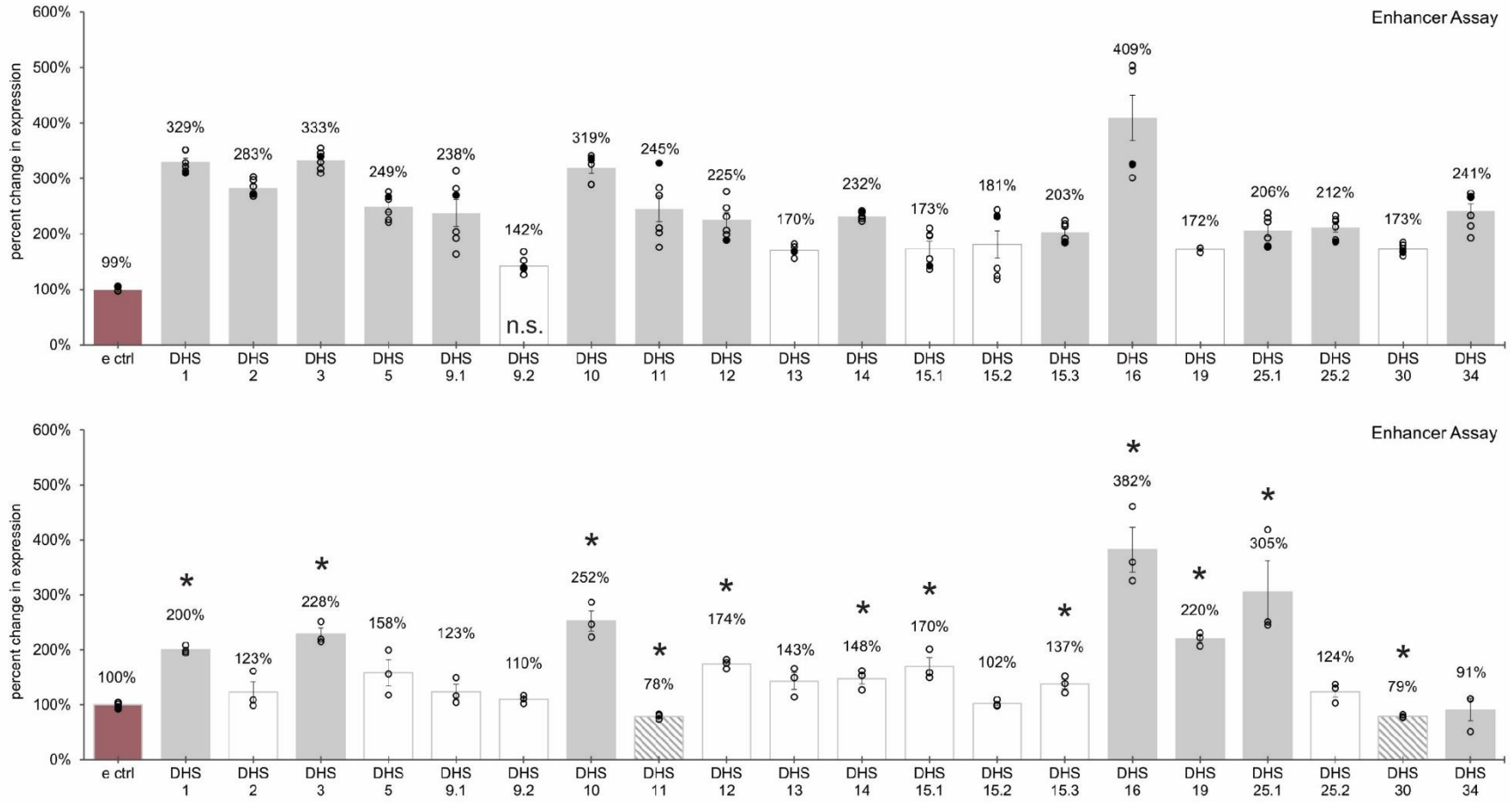


Figure 4.5 DHS activity in enhancer assay for K562 and HepG2

Top panel - K562, bottom - HepG2. Red bar is enhancer assay control, set as 100% expression (SV40p). Expression ≥ 200% shown as grey bar, 100%-200% white, and <100% as hashed diagonal lines. Top panel, all DHS t-test $p<0.05$ for expression above control, except DHS 9.2 (n.s.). Bottom panel, $p<0.05$ indicated by *. Biological replicates (3 or more) shown as open circles, error bars represent +/- standard error.

Both DHS 11 and DHS 30 are potential examples of dual enhancer-silencer elements, showing increased expression in K562 (245% and 173%) and statistically significant activity below SV40p levels in HepG2 (78% and 79%). However these regions are not in accessible chromatin in either cell line and so this activity is considered episomal specific, not reflecting their native regulatory role, where they are inactive.

4.4.3 Contextualizing Enhancer Assay Results

In order to put these results into a meaningful context, we looked at activity alongside chromatin data **(Figure 4.6)** and transcription factor binding **(Figure 4.7)** for these regions. The smaller number of enhancers, overall lower activity, and presence of silencers in HepG2 but not K562, matches what is expected from looking at chromatin accessibility and chromHMM predictions for the region these DHS are part of **(Figure 4.6)**. K562 has many more chromatin accessibility peaks in this region, and has histone modifications associated with activity (green, yellow, red on chromHMM tracks). In HepG2, there are only four distinct accessible regions, and histone modifications indicate largely repressed chromatin all the way up to *ATG5*.

Looking at how genomic data supports episomal results, we see that in K562, all 7 tested DHS with chromatin accessibility peaks (Figure 4.6) showed enhancer activity in the episomal assay, indicating consistency with episomal results. Of the 10 DHS not accessible in K562, 6 showed enhancer activity in the assay, but two of these have evidence of some TF binding, indicating that some of these results (4/20) are likely not relevant within a genomic context.

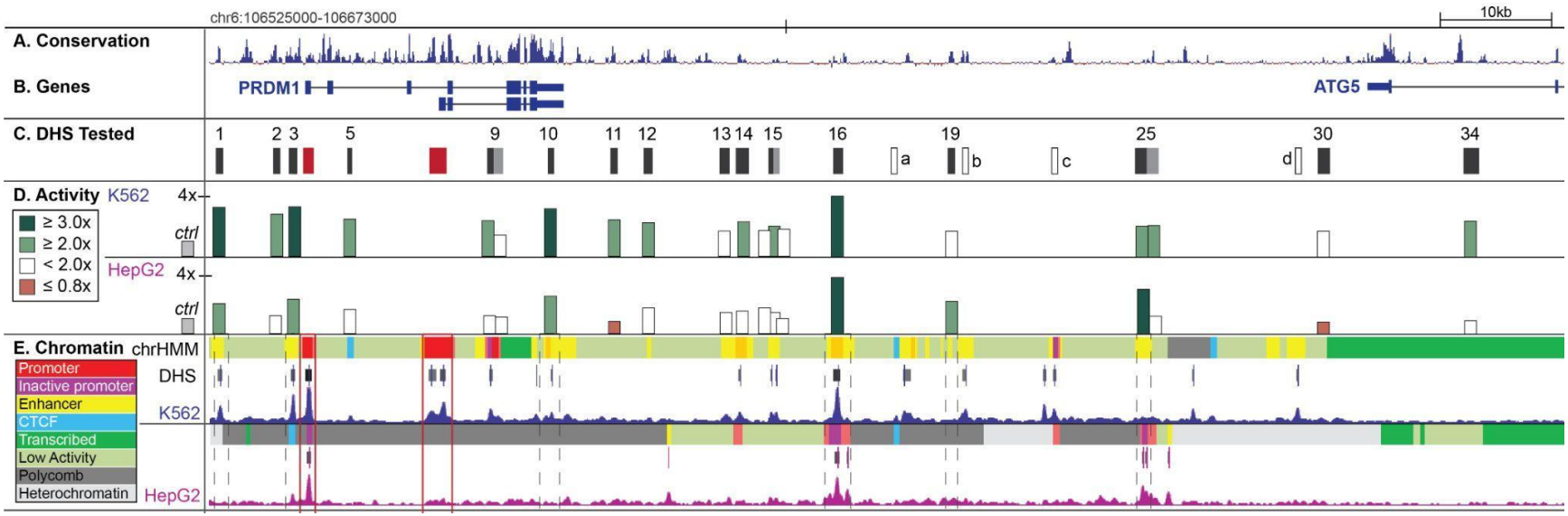


Figure 4.6 Map of DHS activity in chromatin context for K562 and HepG2

a. Basewise evolutionary conservation across 100 vertebrates by phyloP. Conserved sites shown in blue, evolving in red, peak height is |-log(p-value)| under null hypothesis of neutral evolution. **b.** *PRDM1* (two isoforms) and *ATG5* 3'UTR and last exon locations. Exons are blue boxes, introns are lines. **c.** Locations of DHS tested and their genomic positions. DHS 9, 15, 25 were tested as two halves, shown as black and grey boxes. Red boxes are *PRDM1* promoters. Open boxes a-d mark potentially functional sites that were not tested. **d.** Activity from Figure 4.5a-b for K562 and HepG2 mapped underneath respective DHS at the same scale. Peak heights are to scale. Green are enhancers, red are silencers. **e.** chromHMM genome segmentations, predictions based on histone modifications, CTCF, and RNA Pol II, are colored horizontal bars (see key for color codes). Open chromatin by DNaseI HS from ENCODE/Duke. Peaks shown in blue for K562 and magenta for HepG2. Peak calls for enriched DHS signal are black/grey boxes above peaks. Signal for both tracks uses the same scale and $p<0.01$ cutoff. *PRDM1-ATG5* genomic region with DHS sizes, marks and locations kept to scale. Red vertical boxes mark *PRDM1* promoters, dashed boxes highlight DHS active in HepG2. Peak views from UCSC genome browser, hg18.

TF binding data also supports enhancer activity for DHS. **Figure 4.7** shows binding sites from ChIP-seq datasets for K562 and HepG2. TFs listed represent any that are bound at three or more DHS in K562, or at any DHS in HepG2. Other than *PRDM1*'s promoters (red boxes), there are seven total high TF-occupancy sites in this region in K562 cells (there are half as many available TF datasets for HepG2 as K562, so there are fewer bound TFs but also less data overall). Of these DHS 1, 3, 14, 15.3 & 16 all behave as enhancers in the episomal assay (these regions also have supporting DHS peaks). Two other regions, marked as c and d, were not tested due to cloning limitations, but also have strong evidence of high TF binding and accessibility, supporting enhancer identity, as well as regions a and b.

Histone modifications in DHS also largely show consistency with assay activity (data not shown here, taken from ENCODE Broad histone datasets). Almost all of the same peaks with accessibility and high TF binding in K56 show enrichment of H3K4me1 over H3K4me3 signal, and p300 binding. DHS 16 has all of these characteristics, and also H2K27ac signal, all strong correlative markers for enhancer activity [328]. The two strongest enhancers in HepG2, DHS 16 and 25.1, are in open chromatin, and are the only regions in HepG2 with any H3K4me1 signal (with higher K4me1 vs me3) other than promoters. Additionally, DHS 11, which is a silencer in HepG2, shows enrichment for H3K27me3 signal in HepG2, one of the few marks associated with a class of silencers [87].

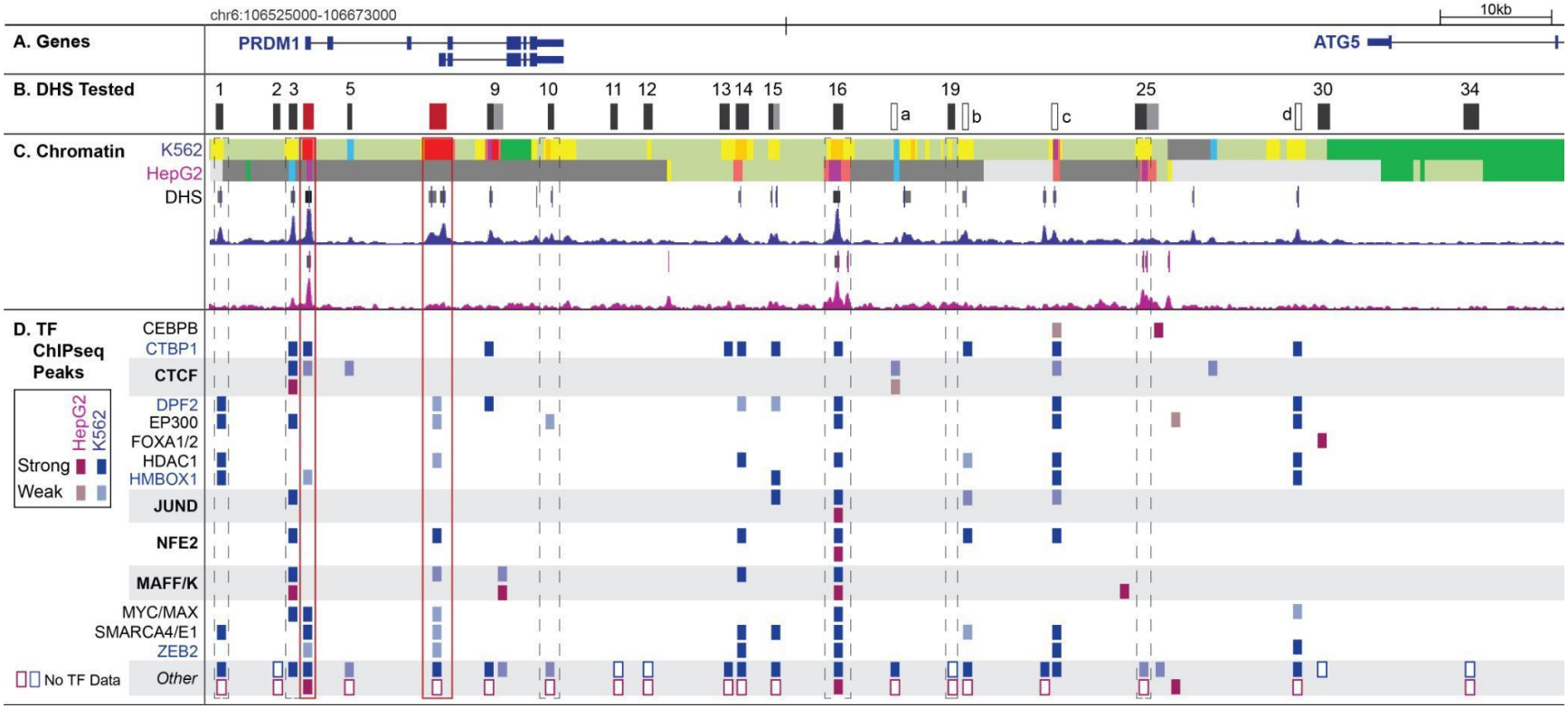


Figure 4.7 Map of DHS TF binding and chromatin context for K562 and HepG2

Red vertical boxes mark *PRDM1* promoters, dashed boxes highlight DHS active in HepG2. **a.** *PRDM1* (two isoforms) and *ATG5* 3'UTR and last exon locations. Exons are blue boxes, introns are lines. **b.** Locations of DHS tested and their genomic positions. DHS 9, 15, 25 were tested as two halves, shown as black and grey boxes. Red boxes are *PRDM1* promoters. Open boxes a-d mark potentially functional sites that were not tested. **c.** chromHMM genome segmentations, predictions based on histone modifications, CTCF, and RNA Pol II, are colored horizontal bars (see key for color codes). Open chromatin by DNaseI HS from ENCODE/Duke. Peaks shown in blue for K562 and magenta for HepG2. Peak calls for enriched DHS signal are black/grey boxes above peaks. **d.** Transcription factor binding ChIP-seq data from ENCODE 3. On the left side, TFs which bind one or more DHS in HepG2, or at least 3 DHS in K562 are listed (TF name in blue means there is no data for that TF in HepG2). A bar under a DHS location indicates a ChIP signal for that factor in that DHS (not to scale with ChIP peak size). Blue box means the TF binds in K562, magenta in HepG2, lighter colors indicate a low signal or peak in only one replicate. 'Other' indicates at least one TF not listed here bound at the site. Open boxes indicate no evidence of any TF binding for that cell type at that DHS, in ENCODE 3. Coordinates are for hg18.

4.4.4 Silencer and Enhancer Blocker Assay Results

All DHS were also tested in the silencer and enhancer blocker assays in K562 cells **(Figure 4.8).** None of the tested regions showed silencer activity in K562. This is consistent with enhancer results and the model for *PRDM1* activity in K562 cells. Neither did any of these elements show enhancer blocker activity. Predicted enhancer blockers DHS 3 and 5 both show enhancer activity only, despite binding the known human enhancer blocker factor CTCF. While DHS 11 and 30 showed potential silencer activity in the enhancer assay in HepG2, they were not tested in the si and eb assays in HepG2.

Despite the lack of NRE activity for these elements, two interesting effects were revealed by the use of the si and eb assays and the use of an assay panel rather than a single assay type - positional effects and synergistic effects. These effects both relate to the presence of the HS2 enhancer in the si and eb plasmids. As the majority of the DHS tested as enhancers in K562, in the si and eb assays which contain HS2e, the combinatorial effects two enhancers can be assessed.

Positional effects are detectable at both an assay- and DHS-specific level **(Figure 4.8)**. At the assay level, overall expression for DHS is higher in the silencer assay (DHS upstream of HS2e), and lower in the eb assay (DHS downstream of HS2e) - indicated as a blue + above the bar for that DHS. It seems that for most of these elements, the DHS has little contribution to expression when placed downstream of HS2e, but can impact expression when upstream of it. A potential explanation for this is discussed in **Section 4.5.2** below. At the DHS level, the strength of this positional effect varies. DHS 12 doubles expression equally in either position (up-199% / down-222%). Bigger positional differences seem to be driven not by changes in activity across the enhancer blocker assay (DHS-down), but by increases in activity in the silencer assay (DHS-up).

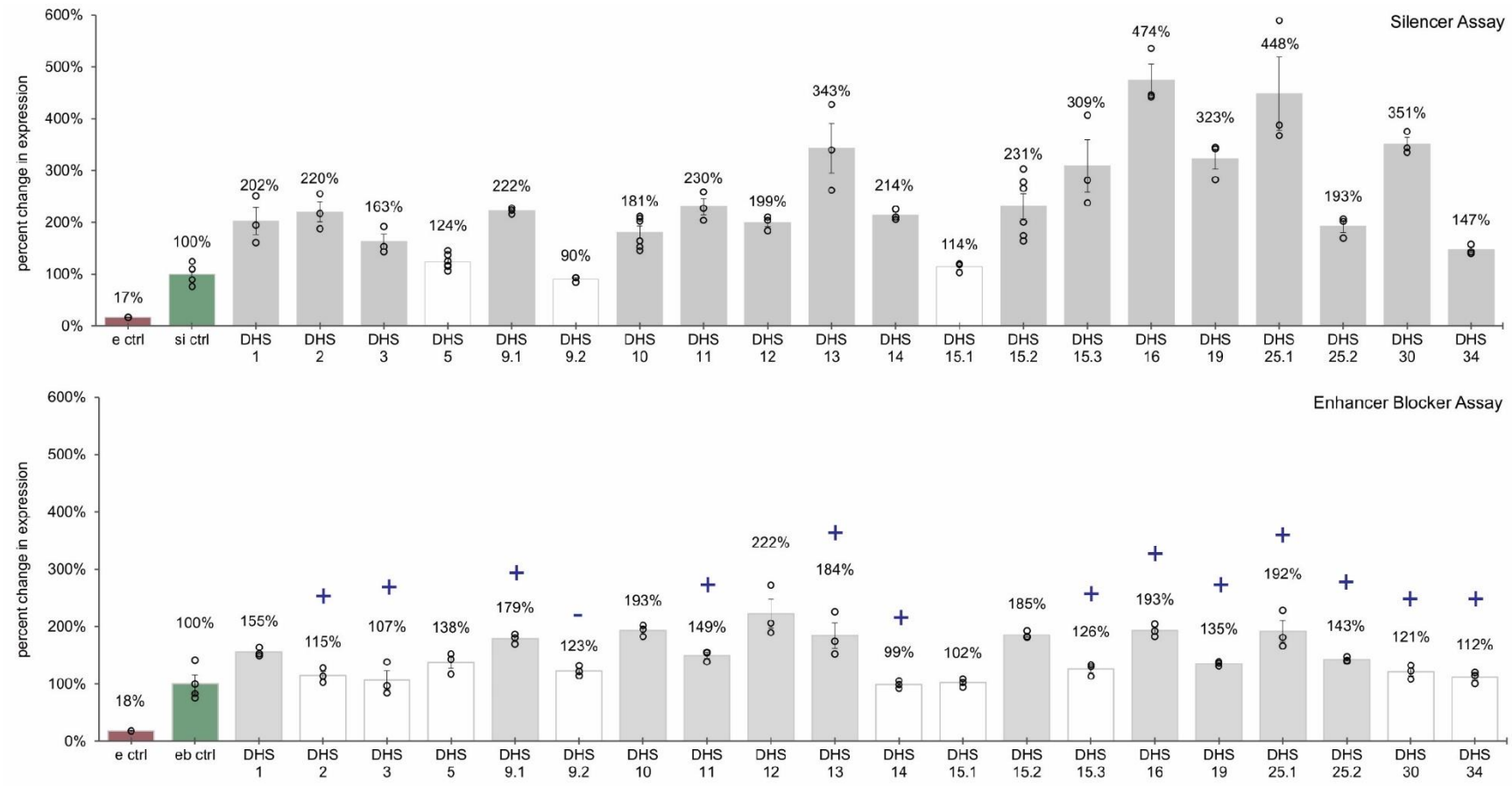


Figure 4.8 DHS activity in silencer and enhancer blocker assays

Top panel - silencer assay (upstream DHS), bottom - enhancer blocker assay (downstream DHS) results. Red bar is enhancer assay with control in test position (ctrl-SV40p), green bar is si or eb assay with control (ctrl-HS2e-SV40p or HS2e-ctrl-SV40p), set to 100%. For both assays, grey bar = t-test $p<0.05$ for expression above control, white is $p>0.05$. Biological replicates (3 or more) shown as open circles, error bars represent +/- standard error. Blue + shows significantly increased normalized expression values for the element in si vs eb assay (- is decreased si vs eb).

4.4.5 Additive and Synergistic Activity of Multiple DHS Enhancers

This positional effect is also connected to enhancer synergy, the second effect revealed through the use of these assay panels. Previous studies have characterized the effects of two combined enhancers on gene expression to be either additive (activity of both = sum of individual activities), sub-additive (activity of both is less than sum of parts), or synergistic (activity of both is more than sum of individual activities) [262]; [263]. In order to determine which of these effects may be present with the DHS-HS2e combinations in K562, I looked at fold change increases in expression across all three assays and compared them to the e control (SV40p only) set to 1x. **Figure 4.9** shows the results. DHS activity was calculated as DHS activity in the enhancer assay (DHS-SV40p) minus SV40p activity (1x, measured by the e control's activity) and is shown in **Figure 4.9** as red boxes. HS2e activity was calculated as activity in the si and eb control plasmids, 6x (HS2e-SV40p), minus the activity contributed by SV40p (1x) to give 5x. This 5x was added to the calculated DHS activity to give predicted additive DHS-HS2e activity, shown as grey boxes. This was plotted against observed data from SV40p-subtracted si assay (DHS-e-p or DHS up, in green) and eb assay (e-DHS-p or DHS down in purple) data. Where observed (green, purple) boxes are higher than predicted (grey), that represents synergy, and where they overlap, additive behavior.

These results show that there are both positional and synergistic effects of HS2e-DHS combined activity. 15 DHS showed synergy and 5 were additive with HS2e in the DHS-up position (green vs grey). In the DHS-down position, 6 showed mild synergy, 9 were roughly additive, and 5 did not increase expression above HS2e at all. All DHS that showed DHS-down synergy also showed DHS-up synergy, but not vice-versa.

Neither positionality nor synergy in either position correlated with DHS strength in the enhancer assay. Neither is there a clear trend across the DHS with the strongest genomic support for enhancer activity (DHS 1, 3, 14, 15.3 & 16). DHS 16 showed the strongest synergistic activity, with expression 3.3 times higher than predicted for additivity. Synergy/positionality in this model is likely driven by compatibility with the HS2 enhancer, not genomic function of a DHS, explaining why genomic markers may not be good predictors of combinatorial activity with HS2e (see **Discussion**). Additional testing is needed to determine what drives synergistic DHS-HS2e relationships.

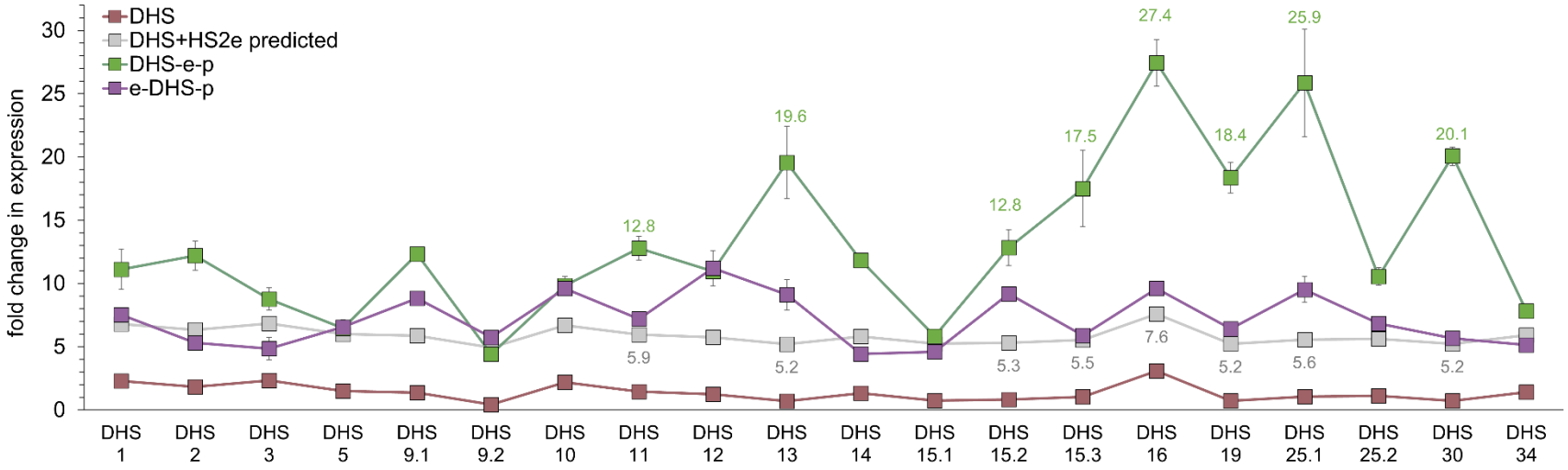


Figure 4.9 Predicted vs observed DHS-HS2e activities reveal additive and synergistic effects

Y-axis is fold change in luciferase activity relative to an SV40 promoter set at 1x. Boxes show values for each condition. Red - DHS activity alone, Grey - predicted DHS+HS2e activity, Green- observed activity for si assay (DHS-up) Purple - observed activity for eb assay (DHS-down). Predicted/observed values are listed above boxes for DHS where observed values are at least double predicted values. Lines between boxes do not represent any relationship between boxes but are an aid to visualization.

4.4.6 TFBS Deletion and Insertion Series in DHS 16

Across all the data discussed here, DHS 16 stands out as a strong model of many of the characteristics discussed. It showed the strongest activity in the enhancer assay (a 4-fold increase in luciferase activity) **(Figure 4.5)**, is expressed strongly and at has the same level in both K562 and HepG2, has the highest TF occupancy, is in a strong DHS peak in both cell lines, and has all three classic marks of enhancer identity - p300 binding, a high H3Kme1/me3 ratio, and H3K27ac. Additionally it shows a strong example of position-dependent synergy with HS2e. As the tested DHS 16 fragment was 900bp, I wished to dissect the region to determine which sequences (TF binding sites or TFBS) are necessary and sufficient for its activity. **Figure 4.10** shows a scale diagram of DHS 16 including the peak call from the DNase-seq accessibility data. It also shows the location and distribution of ChIP-seq peaks for TF binding in this region in K562, and TF footprint data (which uses TFBS motif and DNase-seq data to pinpoint TF binding sites [332]) as ChIP peak resolution is much larger than actual binding site size.

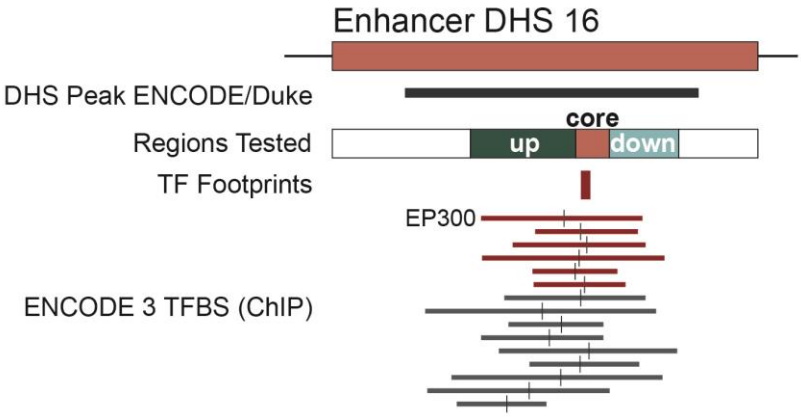


Figure 4.10 TF binding structure of DHS 16
DHS 16 in red, with three regions tested in deletion series below (green = up, red = core, aqua = down). DNase-seq data peak call for DHS 16 shown as a thick black line. Location in DHS 16 of ChIP-seq peaks for various TFs shown as horizontal lines with small vertical lines indicating ChIP peak position. TF with ChIP peaks in red are also supported by TF footprint data (small red box). All elements shown are in their correct positions and to scale.

Using these data, I determined that the majority of TF binding sites for this element fell within the 70bp red 'core' region shown below. However, given the potential complexity and sequence-context dependence of enhancer grammar, I decided to include testing of regions 220bp upstream and 146bp downstream of the core as well. I tested deletions of all three of these regions, and combination deletions, in the plasmid

containing DHS 16 in the enhancer assay context, shown in **Figure 4.11**, with diagrams showing which deletions were made and the resulting structure on the left and expression on the right. Baseline expression of the DHS 16-enhancer plasmid without deletions is shown in dark grey.

Results from this deletion series reflect a TF binding structure for DHS 16 that most closely matches the 'TF collective' model for enhancer structure. Deletion of the 'core' predicted region with the majority of TF binding reduces activity, however only by 38%, indicating that this region contributes, but alone is not driving the full enhancer activity. Deletion of the upstream region also results in a loss of activity (this region contains the entire ChIP peak for ZEB2 binding), but the downstream increase expression when it is lost. Finally, deleting the entire up-core-down fragment does not significantly change expression (there is perhaps some reduction masked by higher error for that condition), indicating a non-additive relationship between these sequences. These results support the importance of the entire DHS 16 sequence for driving the element's activity.

I next tested for sufficiency of the core/up/down regions to drive expression in an insertion series in my enhancer assay, shown in **Figure 4.12**. As a control for the effect of inserting a sequence, I inserted a random 70bp sequence taken from the r33 neutral control sequence used in these assays. This control had expression 15% over the assay backbone alone. Assuming this is representative of the effect of DNA insertion, the core TF binding region drives an additional 15% expression, and the up and down regions drive even more expression. Adding the (control-subtracted) up+down+core individual effects gives a predicted 83% of activity accounted for. Testing insertion of just the up-core-down region without flanking DHS sequence restores 60% of control-subtracted activity, much less than predicted by individual element effects, supporting a potential non-additive effect of the combination of elements, consistent with the results from the deletion series.

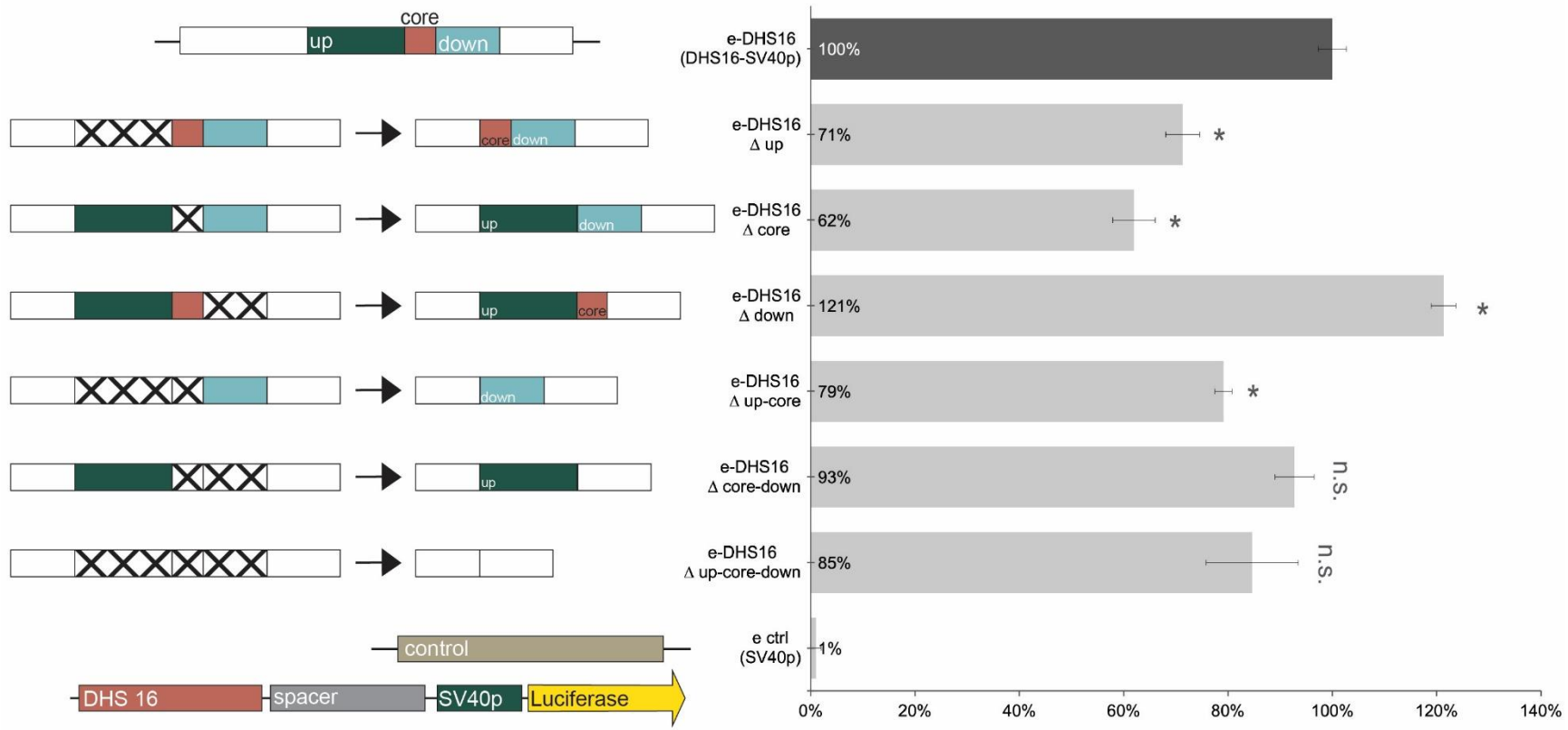


Figure 4.11 DHS 16 TFBS enhancer assay deletion series
Bottom shows base DHS 16 enhancer plasmid. **Left panel** shows regions of DHS 16 which were used for each condition (labeled up, core, and down). Deletions are shown as Xs, and resulting DHS 16 structure is shown to the right of the arrow. Expression values shown in **right panel**, normalized to full DHS 16 in enhancer assay (100%). t-test for expression vs e-DHS 16 control, $p<0.05$ indicated by *. Error bars represent +/- standard error.

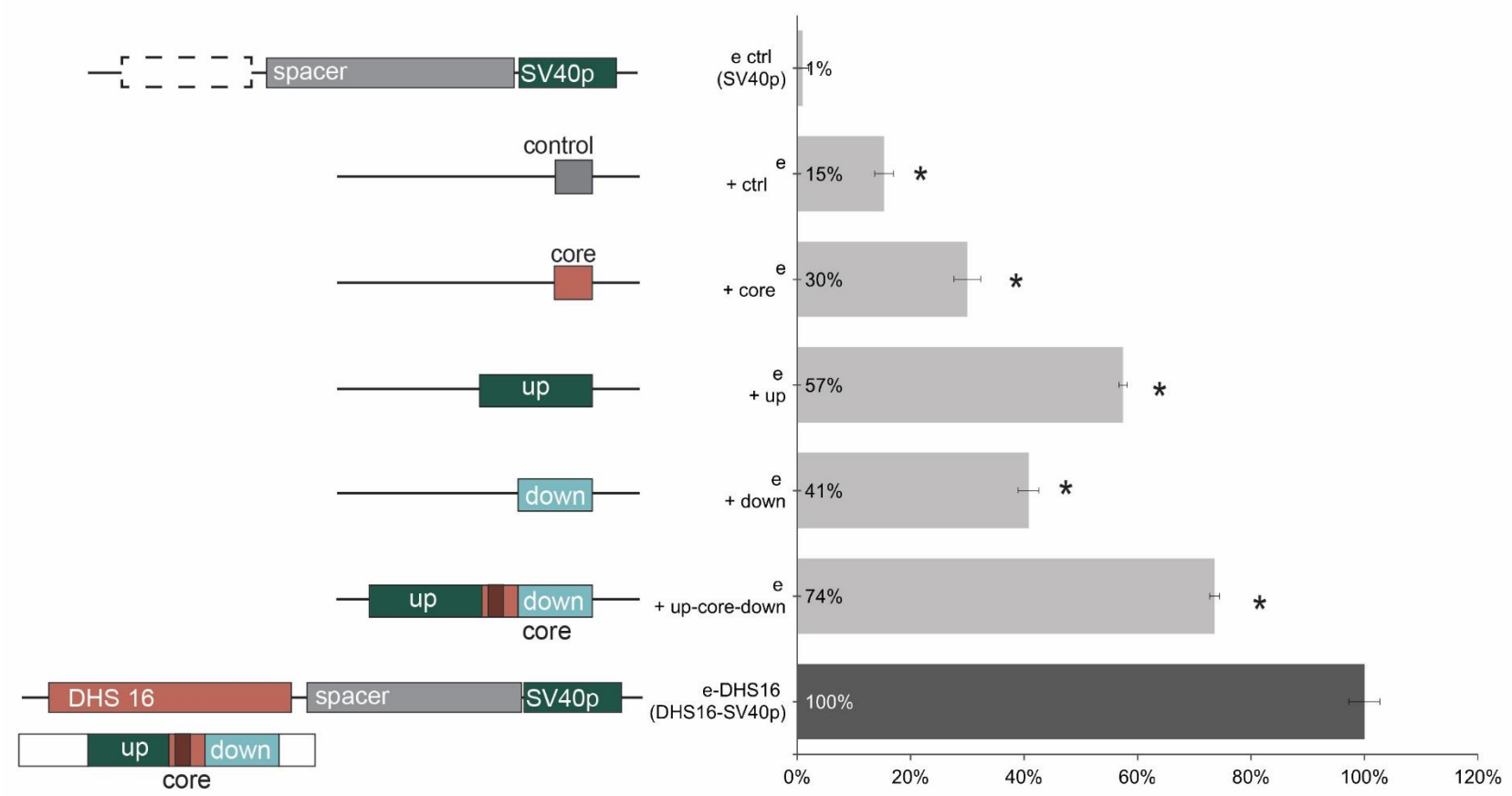


Figure 4.12 DHS 16 TFBS enhancer assay insertion series
Bottom shows base DHS 16-enhancer plasmid. **Left panel** shows regions of DHS 16 which were inserted for each condition, and their respective locations in DHS 16 in diagram at bottom left. Expression values shown on **right panel**, normalized to full DHS 16 in enhancer assay (100%). t-test for expression vs e-DHS 16 control, $p<0.05$ indicated by *. Error bars represent +/- standard error.

4.4.7 TFBS Silencer and Enhancer Blocker Deletion Series

In order to determine whether the presence of HS2e impacts the effects of deletions in DHS 16, and to isolate which regions of DHS 16 might be contributing to positional synergistic effects, I tested the same DHS 16 deletions as shown above in the silencer and enhancer blocker assays. **Figures 4.13** and **4.14** show the results of this test, completed in the DHS 16-HS2e silencer **(Figure 4.13)** and HS2e-DHS 16 enhancer blocker **(Figure 4.14)** assays. In order to focus on decreases in the expression contributed by DHS 16, the 'high' baseline in dark grey in these figures is the si/eb plasmid with DHS 16 (DHS16-HS2e-SV40p / HS2e-DHS16-SV40p) and the 'low' baseline (bar at bottom) is the expression driven from the si/eb control plasmid (ctrl-HS2e-SV40p / HS2e-ctrl-SV40p). Percent decreases are measured on this scale and not over the scale of total expression overall to isolate the percent of DHS 16 activity impacted.

Results from these two assays show opposing effects. In the silencer assay (DHS 16-up) where DHS 16 and HS2e showed strong synergy, no deletion or set of deletions significantly decreased expression of the overall construct. The exception is deletion of all three regions, which yielded only a 41% decrease in activity. In the enhancer blocker position, however, deletion of the core region depleted an amount of activity almost exactly equal to the entire effect of DHS 16, as was expected for the enhancer assay. This is consistent with a model where different sequences and factors contribute to DHS 16 synergy with HS2e than those that are responsible for driving its activity alone. A potential model for this effect is discussed in **Section 4.5.2** below.

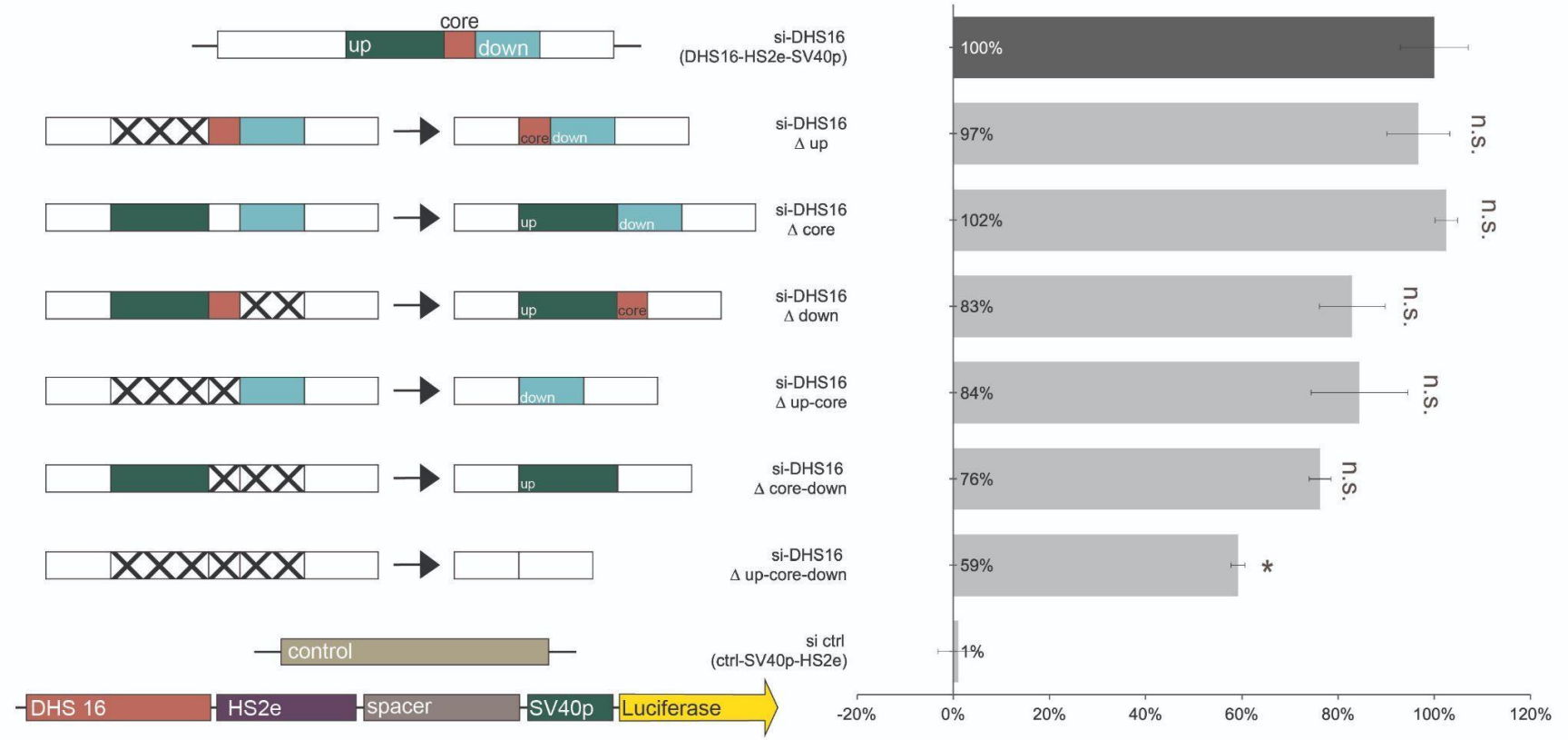


Figure 4.13 DHS 16 TFBS silencer assay deletion series
Bottom shows silencer-DHS 16 plasmid components. **Left panel** shows regions of DHS 16 which were used for each condition (labeled up, core, and down). Deletions are shown as Xs, and resulting DHS 16 structure is shown to the right of the arrow. Expression values shown in **right panel**, normalized to full DHS 16 in silencer assay (100%). t-test for expression vs e-DHS 16 control, $p<0.05$ indicated by *. Error bars represent +/- standard error.

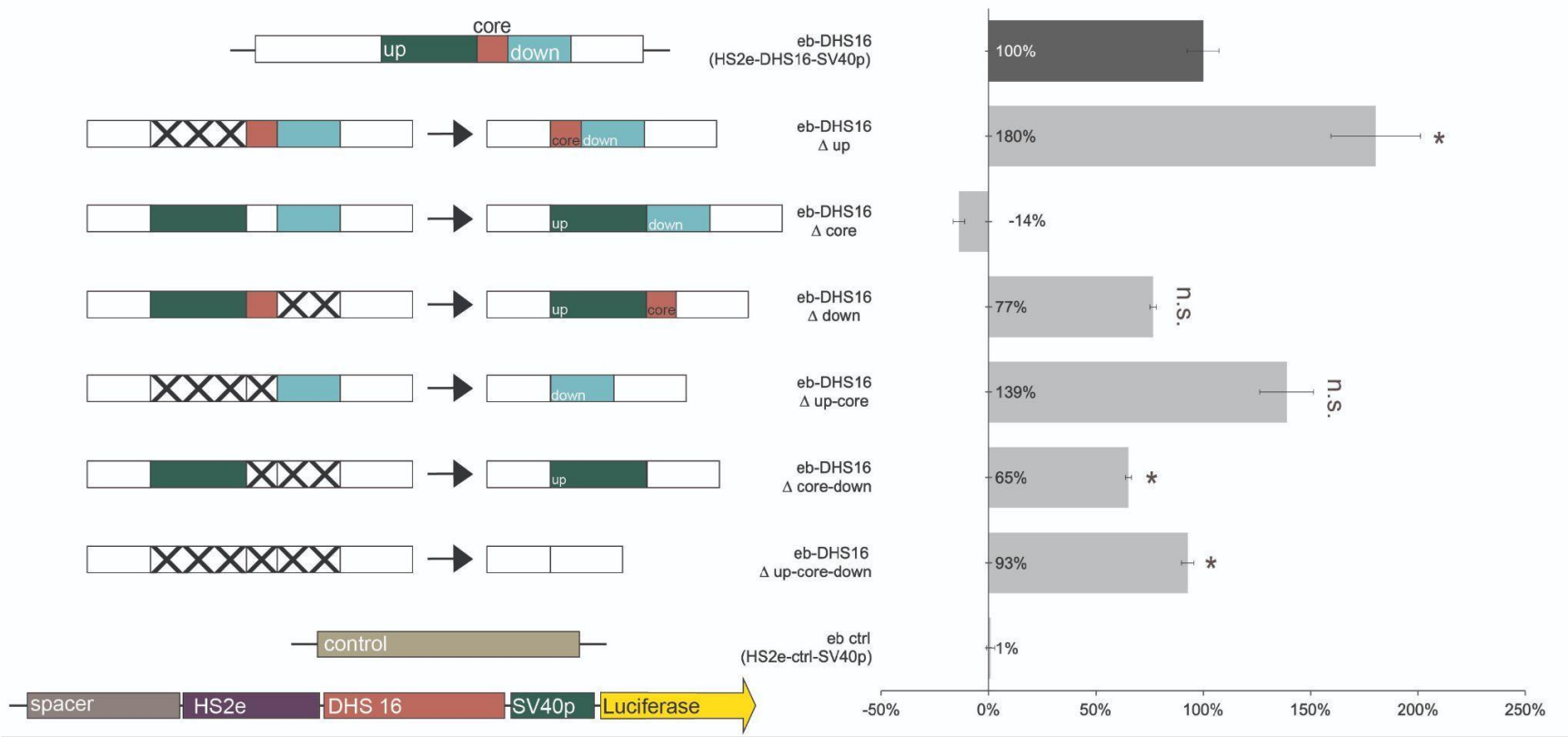


Figure 4.14 DHS 16 TFBS enhancer blocker assay deletion series
Bottom shows enhancer blocker-DHS 16 plasmid components. **Left panel** shows regions of DHS 16 which were used for each condition (labeled up, core, and down). Deletions are shown as Xs, and resulting DHS 16 structure is shown to the right of the arrow. Expression values shown in **right panel**, normalized to full DHS 16 in enhancer blocker assay (100%). t-test for expression vs e-DHS 16 control, $p<0.05$ indicated by *. Error bars represent +/- standard error.

4.4.8 Models for Enhancer Activity and Expression in the *PRDM1-ATG5* Region

There are a number of interesting models for this region suggested/supported by results from episomal testing and existing datasets. First, regulation of this region seems to be mediated by a large number of relatively weak enhancers in K562 cells. *PRDM1* is tissue-specific must be expressed at only specific developmental timepoints in some cell types - in B-cells, *PRDM1* needs to be expressed at a specific time point in differentiation and premature expression can disrupt function [329], [330]. A likely hypothesis is that having a more complex system of weak enhancers which must coordinate to drive expression allows more fine-tuned control of spatiotemporal expression patterns. It allows for a more complex grammar of regulation given the suite of TFs expressed in each cell type. This is also supported by the way TF binding is distributed across the DHS. The TFs listed in **Figure 4.7** are all bound in at least three DHS, and some bind at ten sites across the region, but the patterns of which DHS bind vary greatly for each TF. No two sites have identical sets of TFs bound, meaning they are all likely being regulated by slightly differing pathways.

Second, DHS 16 remains strongly active as an enhancer in HepG2, unlike other regions tested, despite being surrounded by large regions of repressed chromatin. This indicates an ongoing, important role for these enhancers in HepG2. I propose that DHS 16 is a likely candidate for an *ATG5*-specific enhancer. It may also regulate both genes in K562 and switch to enhancing just *ATG5* in HepG2. The strong enhancer activity and classical enhancer characteristics of DHS 16, including p300 binding, H3K4me1 & H3K27ac peaks, and high TF occupancy, make it a likely candidate for driving expression of *ATG5*, which is expressed at much higher levels than *PRDM1* in both lines, and is an essential gene. Conversely, the repression of all the *other* DHS enhancers in HepG2, where *PRDM1* is repressed, also supports their role as *PRDM1*-specific enhancers.

It is also highly likely that the upstream *PRDM1* promoter has a role in *ATG5*, as it, but not downstream *PRDM1* promoter, is strongly accessible in every single one of the 25 tissue-diverse cell lines checked, supporting a ubiquitous activity. That this promoter remains so active but does not drive PRDM1 expression, indicates it may

have an alternate role, perhaps as an enhancer, or some structural role in maintaining a DNA loop.

Third, these results begin to provide some insight into which mechanisms are and are not responsible for the differential regulation of *PRDM1* and *ATG5* in these two cell lines. Surprisingly, only two of the twenty elements tested showed potential silencer activity in HepG2 and none in K562, but neither of these elements is accessible in either cell line. Given that we have not tested every DHS site in this region, or the handful upstream of *PRDM1*, it is possible that there are silencers present which may still be discovered - we will consider this one possibility. However the expected pattern in the case that silencer CRE activity is responsible for the repression of *PRDM1* in HepG2, would be that a silencer region is in closed chromatin in K562 (inactive) and becomes accessible in HepG2. Looking at this region, including downstream through and past *ATG5*, there are only two genomic regions that follow that pattern: the small DHS peak to the right of DHS 12 (marked by a * in **Figure 4.7)**, and a much stronger peak downstream, between the second- and third-to-last exons of *ATG5* (not shown). The * region has no TF binding or histone modification support for activity. The upstream DHS seems to be an enhancer, as it has classic H3K4me1/3 ratios, p300 binding, high TF occupancy, and binding of FoxA1, a liver-specific activating TF.

So how does differential regulation occur? From an overview of the region and its chromatin state **(Figure 4.6e)**, it seems likely that it occurs primarily through the spreading of the region of heterochromatin, which is present in both K562 and HepG2 upstream of *PRDM1*, into *PRDM1* in HepG2 cells only. In HepG2, this heterochromatin domain, marked by a broad domain of histone K7me3 modifications and EZH2, continues up until about DHS 25.1, 17kb from the 3' end of *ATG5* (not shown, roughly corresponds with grey regions in chromHMM track for HepG2 in **Figure 4.6e**).

The likely cause of this chromatin spreading is the inactivation of an element upstream of *PRDM1* with chromatin barrier function in HepG2. DHS 3 is a candidate for this element, as it is CTCF-binding (ChIP-seq data), has reduced DHS signal and reduced CTCF signal in HepG2 compared to K562, and is located 1kb upstream of *PRDM1*'s upstream promoter. DHS 3's CTCF binding is conserved across a number of cell lines, and looking at patterns of chromatin state and CTCF binding across five cell

lines (K562, HepG2, HeLa, GM12878, and HUVEC), consistent correlations can be seen between activity of the region downstream of DHS 3 (indicated by DHS peaks and chromHMM segmentations) and the strength of CTCF binding at DHS 3 **(Figure 4.15)**. In the three cell lines where *PRDM1* is in active chromatin and which have more DHS accessibility in and downstream of *PRDM1* (K562, GM12878, HeLa), CTCF binding at DHS 3 is stronger. In HepG2 and HUVEC, where there is less accessibility and the region is in heterochromatin, CTCF binding is weaker. This supports a possible role for DHS 3 in acting as a chromatin barrier element which is inactivated, allowing spreading of heterochromatin downstream and repressing activity of *PRDM1*-specific enhancers. DHS 3 did not test as an enhancer blocker in my assay (discussed below), however it may have only chromatin barrier activity, which would not be detected in a plasmid-based assay.

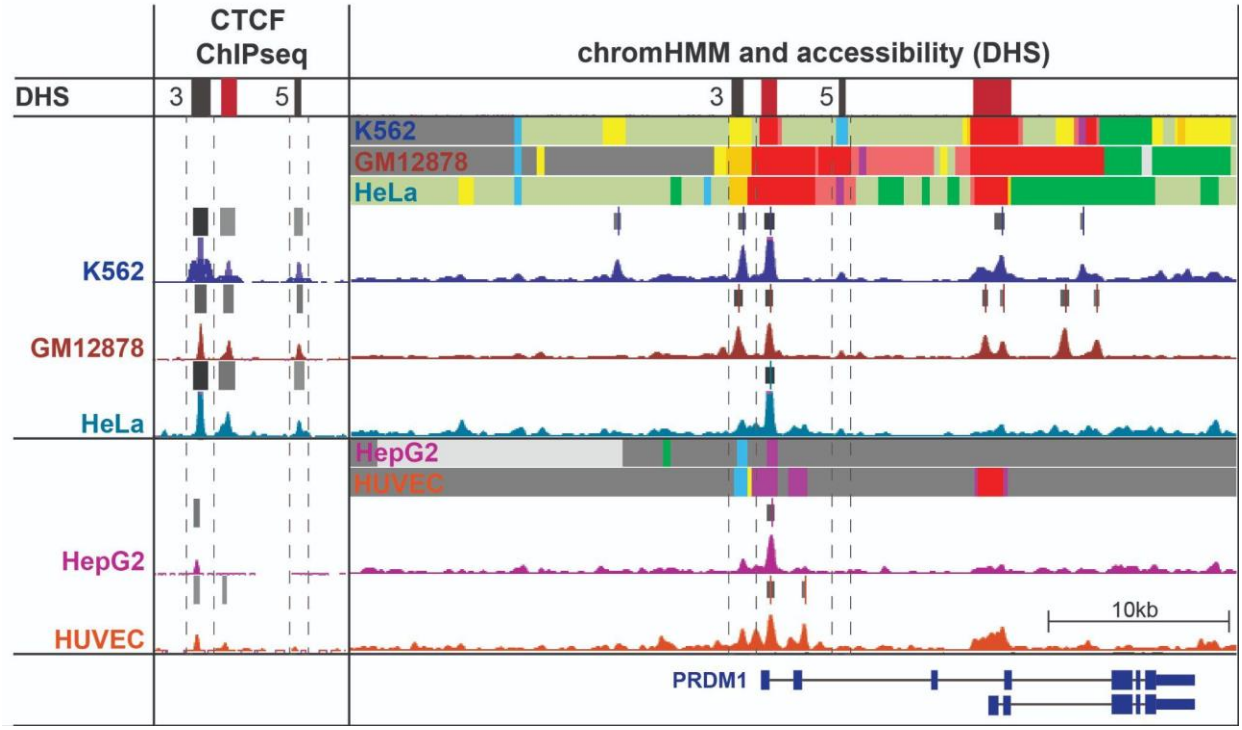


Figure 4.15 CTCF signal strength vs *PRDM1* region activity in five cell lines
All peak tracks are set to the same max height within each datatype set so that respective sizes show respective peak strength across cell lines. Names of five ENCODE Tier 1/2 cell lines on left, color-coded to match signal. Cell lines are separated by *PRDM1* region active (top three) or repressed (bottom two). **Left panel** shows CTCF ChIP-seq signal and peak calls for 5 cell lines at DHS 3 and 5 (dashed boxes). **Right panel** shows broader view of the same region, with chromHMM annotations (See **Figure 4.6** for color key) on top and DHS accessibility signal and peaks below for the region around DHS 3 and 5. *PRDM1* location shown at bottom for right panel.

This model must also encompass differential expression of *PRDM1* and *ATG5* - so heterochromatin spreading must be halted before reaching *ATG5* and shutting off an essential gene. Repression, marked by chromHMM predictions, EZH2 binding and H3K27me3, do in fact all drop off before the 3' end of *ATG5* just after or around DHS 25.1. This suggests a role for DHS 25.1 in acting to prevent this spreading. DHS 25.1 is the only tested DHS which is *inaccessible* in K562, and becomes active in HepG2. If it was important as an *ATG5* enhancer, it would more likely be active and accessible across all lines. Like DHS 16. Also unlike DHS 16, it has few marks of enhancer identity, despite testing with relatively strong enhancer activity - low TF occupancy, no p300, no H3K27ac. Instead it is marked by H3K4me1 and H3K27me3. Neither does it have CTCF binding as expected for an enhancer blocker, however. DHS 25.1 represents an element with possible non-CTCF insulator function with enhancing activity. Further testing is needed to support this role and characterize this interesting behavior. This model of heterochromatin spreading for the *PRDM1* region provides further support for DHS 16's importance as it escapes this repression.

Finally, what is responsible for the increased expression of *ATG5* in K562 cells over HepG2 **(Figure 4.2)**? A likely explanation is that the many weak DHS are tied to *PRDM1* through enhancer-promoter specificity and proximity, but that some of them, or the creation of a more permissible chromatin structure in that region by the activation of the many enhancers, also contributes to increased *ATG5* expression beyond the levels created by its dedicated ubiquitous enhancers, in K562 cells, and this excess expression is lost with their repression in HepG2, but ubiquitous enhancers maintained and *ATG5* expression is kept at necessary levels.

4.5 Discussion

Episomal reporter assays are powerful tools for isolation of a single element of effect and allow researchers to easily manipulate and control changes in sequence to determine function. Outside mechanistic studies however, it is important to relate these findings back to the larger context of the genome, where multiple layers of regulation are present at once and where combinatorial interactions are occurring at the level of the TF, the CRE, the gene, and the regulatory domain. Here I tested twenty sequences

from the *PRDM1-ATG5* domain for CRE activity in three different plasmid assays, using a design informed by an understanding of the nuances of plasmid assay design discussed in Chapter 2 **(Figure 4.3)**.

4.5.1 Episomal vs Native Context

By running this test in well-characterized and annotated cell types, I was able to link element activity back to a larger model of gene regulation in the genome by looking at patterns of activity and cell-type specificity relative to genomic patterns of location and chromatin state. This revealed that 16/20 of the DHS which acted as enhancers in K562 cells in an episomal context **(Figure 4.5)** had corresponding supporting histone mark or transcription factor binding data for the region **(Figures 4.6 & 4.7)**, showing that episomal data largely reflected evidence for native activity. This is consistent with a paper from Inoue et al., which showed correlation between genomic and episomal results for about 4/5 of elements [124]. The episomal results presented, in terms of both larger genomic activity trends for the region and specific element activity, are consistent with genomic evidence for activity in terms of strength and direction of effect. Where there are differences, the episomal assay seems biased toward ‘false’ positives, not false negatives, relevant to genomic function.

4.5.2 Additive and Synergistic Activity with HS2e

Using an assay panel rather than a single assay type revealed interesting patterns of positional and synergistic activity with relation to the HS2 enhancer **(Figure 4.9)**. This DHS-HS2e synergy and positionality data may be best understood with a focus on the biology of the HS2 enhancer. HS2 comes from a locus classically studied for its multiple-element structure and has been shown to have synergistic activity with HS3, another enhancer in the beta-globin locus [331]. As discussed in Chapter 1, while DNA looping is the predominant model for an enhancer-promoter contact mechanism, evidence does exist for a tracking mechanism in some enhancers. The HS2 enhancer happens to be an example of this; it has well-characterized tracking activity. It has been shown to drive transcription from its genomic locus toward the target promoter, regardless of positioning [63]. Disruption of this transcription elongation between HS2

and the promoter also blocks its enhancer activity, indicating it is essential to enhancer function [64].

Given this information, I propose that when the HS2 enhancer is upstream of a DHS, binding of transcription factors at the DHS would disrupt RNA Pol II transit to the promoter. This results in two possible outcomes. In the first, HS2e strongly driving RNA Pol II transcription through the DHS is sufficient to prevent binding of the DHS's transcription factors and it does not become active. This is consistent with the lack of significant increase in expression for DHS 2, 3, 14, 15.1 and 34 above HS2 levels **(Figure 4.8,** bottom, and purple boxes that fall below predicted additive levels in **Figure 4.9**). In the second case, both enhancers can activate, but only alternately, as activation of one prevents activity of the other. So they both alternate driving expression, giving additive activity as they are functioning necessarily independently. However when the DHS is upstream, there is no interference of TF binding at the DHS by RNA Pol II transcription, or vice-versa, and so both enhancers activate freely and synergy is possible. This synergy then varies based on the compatibility of the TFs binding at the DHS with those of HS2e. A more detailed breakdown of TF binding across the DHS might reveal patterns of TF binding correlations with HS2e synergy, or perhaps degree of DHS enhancer transcription plays a role.

One study found that in native genomic contexts, synergy is almost entirely limited to enhancers driving genes with tissue-specific expression [262]. They posited a model where additivity represents enhancer independence, where any one enhancer can sufficiently drive expression, providing redundancy. In contrast, for genes with tissue-specific expression patterns, cooperative binding provides synergy to drive full expression, but also requires multiple specific enhancers to be active, creating a mechanism for precise spatiotemporal control as discussed previously. While I saw no strong trend of this type in my data, DHS 9.1 and 25.1 are more tissue specific and show much more synergy with HS2e, than their 9.2 and 25.2 counterparts. The lack of correlation with enhancer strength observed in my data also differs from another study, in Drosophila, where stronger enhancers were shown to behave sub-additively [263]. By contrast, DHS 16, the strongest enhancer in my set, has the most synergy with HS2e. However trends are harder to predict using a sample size of tens rather than hundreds.

4.5.3 NRE Activity

One goal of this project initially was to discover and characterize novel silencer and enhancer blocker elements, with regards to their interactions with other elements and their role in differential gene regulation. While two DHS regions, DHS 11 and 30 did show silencing activity in my assay in HepG2 cells **(Figure 4.5)**, these regions are in inaccessible chromatin in HepG2 **(see Figure 4.6)** and so their episomal activity may not be reflective of their native role in gene regulation. However, separate from *PRDM1-ATG5* regulation, DHS 11 and 30 are useful additional examples of dual enhancer-silencer elements, a recently identified class of element that is relatively poorly understood [83]; [84]; [85]. DHS 11 and 30 showed 245% and 173% activity in K562 cells and 78% and 79% in HepG2, respectively. While this decrease is fairly moderate, weak silencer activity is consistent with previous work from the Hawkins lab. Their group used STARR-seq design combined with a strong SCP1 promoter to test for silencer elements in K562 cells [84]. A breakdown of their MPRA results shows that of the 3001 elements with $p<0.05$ for silencer activity (of 7323 tested across the human genome), 55% reduced activity only weakly, to between 75-90% of baseline activity. Only 275 elements, less than 10%, reduced activity 50% or more. This could indicate that weaker activity is the norm for the majority of silencer elements. Whether silencers might have weaker activity than enhancers on average, and if so why, and how it relates to their respective mechanisms of action, are important and as yet unanswered questions in this field.

Additionally, none of the DHS tested as enhancer blockers in K562 cells. Predictions for enhancer blocker activity in humans are based on CTCF binding, as this is the best known TF which mediates enhancer blocking/insulator activity in humans. However DHS 3 and 5, which both have ChIP-seq evidence of CTCF binding in K562 cells, behave instead as enhancers. This is not unprecedented, however, as CTCF is known to have a wide variety of activities [121], and as the assay panel here will not capture chromatin barrier (insulator) activity.

4.5.4 A Model for *PRDM1-ATG5* Regulation

While I did not discover NRE activity of accessible DHS elements in this region, the lack of such activity itself was useful in informing which mechanisms were *not* active in directing differential gene expression. In this region which contains a tissue-specific and an ubiquitously expressed gene, regulation is especially crucial. In my model, *PRDM1* is regulated by a large number of moderately active enhancers which are all contained within a region surrounding the gene, where *ATG5* is controlled by a smaller number of stronger enhancers. Repression of *PRDM1* in cell lines where it should not be expressed is managed by decreasing CTCF binding to allow for spreading of heterochromatin across *PRDM1* and its enhancers. Further work, to determine and validate the mechanisms or elements that both initiate this heterochromatin spreading and constrain it from crossing ATG5, is needed. It is possible that DHS 11 and 30 may have a transient role in establishing this heterochromatin state, where the DHS are not accessible in the HepG2 cell line, but were once active during a transition or differentiation to the final cell heterochromatin state and become inactive once the more permanent state is established. This is consistent with one of the proposed roles for silencers [34], however more research is needed to determine whether this is the case for some, or all silencers.

4.5.5 Deletion Series in DHS 16 Reveal Complex Interactions

I also used the generated assay panel to dissect the sequence of the strongest enhancer identified in this region, DHS 16. DHS 16 is a likely candidate regulator of ATG5, given its pattern of strong episomal activity and accessibility across both cell lines. It is unique in being the most strongly expressed in both cell lines in episomal tests **(Figure 4.5),** has H3K4me1 and H3K27ac histone modifications, and p300 binding and high transcription factor (TF) occupancy, all supportive of enhancer activity **(Figures 4.6 and 4.7)**. Additionally, it exemplifies the strong position-dependent synergistic effects observed for many DHS with HS2e **(Figure 4.9)**. Dissection of DHS 16 via deletion **(Figure 4.11)** and insertion **(Figure 4.12)** series of the core transcription factor binding site, and sequences up- and down-stream showed that this region has complex and non-additive interdependencies between its sequences. The results shown

in those two figures support the potential for functional activity across the entire DHS 16 element, not just the core TFBS region, as when that region was deleted activity was not completely ablated. DHS 16's overall activity seems to be generated through complex non-additive interactions of many TFs across the whole region, which produce an activity that dependent on all the sequences together for proper expression, as is consistent with the previously discussed 'TF collective' model for multi-TF binding in an enhancer.

Finally, the results of a similar deletion series in DHS 16 in the si and eb assay contexts clearly demonstrates the importance and power of modeling multi-element interactions in studies of CREs. The results from F**igures 4.13** and **4.14** show that the relative strength and roles of each component sequence in DHS 16 differ in the context of another enhancer (HS2e). Unlike in the DHS 16-only plasmid, deletion of the DHS 16 core sequences does fully reduce the activity of DHS 16 in the enhancer-blocker assay, and has no effect at all in the silencer assay. As these are both inconsistent with the e-DHS 16 results, it is likely that the reduction in activity is due to the impact of those sequences on DHS 16-HS2e synergy. In the enhancer blocker core-deletion case, it is possible that the drop in activity is coming from HS2e and not DHS 16. The combination of the up and down sequences when the core is deleted might create a novel TFBS which interferes with HS2e's function or creates secondary structure which interferes with RNA Pol II procession, as discussed above for the model of HS2 enhancer activity.

DHS 16 binds NF-E2, an erythroid-specific pioneer factor that is well-studied as an important factor in driving HS2 enhancer activity [333]. In fact, every named TF listed in Figure 4.7 as binding DHS 16 (all strongly) also binds HS2 strongly in its native locus (NFE2, p300, MAFF, JUND, ZEB2 etc.). One might expect DHS 16 and HS2e to compete for these factors, but it appears that in fact they behave either additively or synergistically, but not sub-additively in this context. A comparison of the TFs that bind differentially between the two enhancers might reveal what factor is responsible for the loss of this synergy when HS2e is the upstream enhancer. This might also help determine why DHS 16 seems to behave as a TF collective (an additive collection of individual units) in the downstream context of HS2e but not when by itself, and why its core parts seem to have no functional effect on expression when upstream.

If this region reflected the TF billboard model for TF binding in enhancers, activity of the regions would be additive and independent, however that is not the case, as the multi-element deletions do not show a loss of activity matching the summed effects of the individual deletions. In an enhanceosome model, they would be highly interdependent, so deletion of the core would be expected to fully ablate expression, but it does not.

Rather, these data support a complex arrangement of sequences, where each contributes to function but when combined, they produce activity that is more (or less) than the sum of the parts, neatly mirroring the activity of the element as a whole and of the genomic region. Fitting this with my hypothesis that DHS 16 is a regulator of ATG5, a gene which is essential and expressed across almost every tissue type, it would make sense for DHS 16 to have redundancy within its TF binding patterns. This would allow it to be activated by a variety of factors expressed in different cell lines, and prevent loss of any one TFBS from ablating its activity and affecting ATG5 expression.

4.6 Methods

4.6.1 Plasmids and Cloning

The Renilla co-transfection control plasmid used is a 3705bp pRL-SV40P containing *Renilla reniformis* luciferase under control of the SV40 enhancer and SV40 promoter (and a same-orientation AmpR and ori). Firefly plasmids all used the same EMMA receptor vector [161] backbone containing ori, AmpR for selection. All SV40 promoter, CMV enhancer, and Gateway sites, and Firefly luciferase genes are the same sequence, rearranged by restriction-digest and ligation, Gateway recombination (attR -> attB), or EMMA assembly. Firefly luciferase and SV40 promoter originated from a pGL3 vector.

DHS regions were isolated by PCR from K562 gDNA, following protocol for Qiagen gDNA extraction kits (#51304). PCR was done using Phusion High-Fidelity polymerase (NEB #M0530S), using 100ng gDNA per 50uL reaction, following kit instructions (1uL 10mM dNTP, 2.5uL each 10uM fw and rv primer, 10uL 5x HF buffer and 0.5uL polymerase to 50uL in nuclease-free H2O), -/+ 3% DMSO, at different temperatures depending on the optimal region conditions. Initial denature 98°C 2min, denature 98°C 10 sec, anneal 30 sec, with extension at 33sec/kb. Final extension 5min.

PCR was cleaned up using QIAquick PCR Purification Kit (Qiagen #28104). DNA measured by qubit on run on a gel for fragment size analysis. Successful PCR products were phosphorylated using PNK and blunt ligated (using 1uL high concentration T4 Ligase (#M0202M NEB) and 1uL T4 buffer, 1:10 backbone:insert, 10uL reaction, 100ng backbone, incubated 16°C overnight) into CIP'd pEntr1a Gateway backbone (Kan selection marker on backbone) between attL sites. Plasmids were transformed into Stbl3 *E. coli* and restriction-digest screened. Successful clones were also screened via Sanger, and fw orientation chosen.

pE-DHS vectors were then combined with reporter assay constructs containing Gateway attR sites with an internal bacterially-expressed RFP for screening (and AmpR selection marker on backbone), using Gateway™ LR Clonase™ II Enzyme mix (ThermoFisher #11791020). Gateway reactions: 50ng attR plasmid, 1:10 molar pE-DHS plasmid, 1uL Gateway LR Clonase mix, to 5uL in H2O. Incubated 1hr-2hr or overnight at RT, stopped with 0.5uL Proteinase K for 10min 37°C, and transformed into 50uL *E. coli*, and selected on Amp. White colonies were selected and screened by digest, and insert site screened by Sanger or using OnRamp.

TF deletion series was created in existing DHS 16-assay plasmids, using fortuitous restriction digest sites within DH16 (BamHI-DraIII for 'up', DraIII-PvuII for 'core', and PvuII-PciI for 'down), treated with klenow to fill in sticky ends, and re-ligated without cleanup at high dilution (1uL T4 Ligase as above, in 100uL reaction) to optimize for backbone self-ligation. Ligations were transformed at 1uL/25uL Stbl3 *E. coli*. Products were screened by digest, or Sanger/OnRamp nanopore sequencing as needed. TF insertions were performed using NEB HiFi assembly protocol for single-stranded small products (NEB #E5520 - see manual for details) into the e-DHS 16 plasmid backbone (with the entire DHS 16 removed by AfeI-SacII restriction digest, blunted by Klenow, and gel extracted) and transformed at 2uL per 50uL cells, and screened similarly.

4.6.2 Cell Culture

All transfections listed in this chapter were done in the human myelogenous leukemia cell line K562 (ATCC CCL-243^{TM}) or adherent hepatocellular carcinoma line

HepG2 (ATCC HB-8065^{TM}). Cells were grown at 37^{o}C and 5% CO_2. K562 were cultured in RPMI-1640+L-glutamate media (ThermoFisher, 11875093) containing 10% heat-inactivated FBS (ThermoFisher, 10437028) and 1x antibiotic/antimycotic (ThermoFisher, 15240112) (K562 complete media). HepG2 were cultured in EMEM (Corning #10-009) with 10% Non-heat inactivated FBS (Corning 35-015-CV) and 1x antibiotic/antimycotic (ThermoFisher, 15240112) (HepG2 complete media). Cells were passed every 2-3 days depending on confluence (4-6 days for HepG2, with fresh media added as needed) and were sacked after passage 10. HepG2 required trypsinization for passage. Supernatant was removed and 1.5mL per 25mL culture of 0.25% Trypsin-EDTA (ThermoFisher 25200056) added, then incubated for 10 min at 37°C. At least 10x (vol Trypsin) media was added to inactivate, and cells split to new flask. Cells were incubated for 48 hours 37°C and 5% CO2 and then collected for readout (see 4.6.5).

4.6.3 Transfections

HepG2 transfections were done using Lipofectamine 3000 (ThermoFisher L3000015). 0.4x10^6 cells were plated in 12-well plates in 1mL HepG2 complete media 24 hours prior to transfection. Biological replicate DNA mixes were made as pools prepped at 1.1x excess. Per biological replicate: 1ug DNA (or molar equivalent, using E control plasmid as standard), 2uL P3000, 2uL Lipofectamine 3000, and Opti-MEM Reduced Serum Medium (ThermoFisher, 31985062) to 100uL. DNA-Lipo mixes were incubated 15 min at RT and then 100uL was added dropwise to each well. No DNA controls were prepared with reagent and no DNA for measurement of background signal. Cells

All transfections of K562 were done via electroporation using a NEPA21 Electroporator (Nepagene). Cells were checked for minimum 75% viability and 50% confluence prior to electroporation. Cells were collected by centrifugation at room temperature (RT) at 100xg for 10min. Supernatant was removed and cell pellets resuspended in 15mL per initial 50mL conical of cell culture of RT Opti-MEM. 500uL was removed and set aside for a cell count and viability check. Remaining cells were spun again at 100xg for 10min at RT and the 500uL aliquot counted during this time. After spin, supernatant was again removed and cells were resuspended with fresh Opti-

MEM to a final concentration of 1x10^6 cells/90uL, measured for accuracy by pipette. For each condition biological replicate, 99uL of cells was added to 1.5mL microcentrifuge tubes containing 11uL pre-aliquoted DNA (see 2.6.3 below). Cell-DNA mixtures were mixed by pipette, then 100uL of mixture (90uL cells, 10uL DNA) was pipetted to each of three 2mM cuvettes (Bulldog Bio,12358-346). Cuvettes were electroporated using poring pulse: 275V, 5ms length, 50ms interval, 1 pulse, 10% D rate, + polarity. Transfer pulse: 20V, 50ms pulse length, 50ms pulse interval, 5 pulses, 40% D rate, +/- polarity. Immediately following electroporation of each cuvette, 900uL RPMI 1640 complete media pre-warmed to 37^{o}C. Cells were transferred to 24-well tissue culture plates by pipette and incubated (as listed above in Cell Culture) for 48 hours.

4.6.4 Preparation of DNA for Electroporation

All data in this chapter were generated using Firefly luciferase expression. For K562, DNA amount per cuvette was set at 1.5ug of a 5416bp pGL3 plasmid, per 1x10^6 cells, and all other plasmids were transfected in molar-equivalent amounts. All were co-transfected with a pRL plasmid containing Renilla luciferase at a 1:25 molar ratio. DNA concentrations were measured by Qubit 4 Fluorometer (ThermoFisher) using the dsDNA BR Assay Kit. All DNA to be used in a transfection was measured at the same time, using the same Qubit dye & buffer mastermix to account for measurement fluctuation due to mastermix preparation and room temperature. DNA mixes for each condition and replicate were made in a 1.5mL microcentrifuge tube containing firefly DNA and renilla DNA in molecular-grade H2O, at a 1.1 scale to account for pipetting error. These were made prior to electroporation to minimize cell time at RT.

4.6.5 Readout of Luciferase Signal

Luciferase readouts used Promega's Dual-Glo Luciferase Assay System (Promega, E2920). Readout was done using a GloMax-Multi+ Detection System (Promega, E7081) plate reader. At 48 hours (+/- 4 hours) post-electroporation, all cells from each well were collected by pipette into individual 1.5mL microcentrifuge tubes. For HepG2, supernatant was collected to tubes, and 400uL 0.25% Trypsin-EDTA added

per well, and incubated 37°C for 10min. Supernatant from respective tubes was used to quench Trypsin for each well, and entire sup-Trypsin mix was collected back to the tube to ensure collection of entire well contents.

Cells (HepG2 or K562) were spun in a centrifuge at 500xg for 5 minutes. Supernatant was removed by either pipette or vacuum except for ~50uL to avoid disrupting the cell pellet. 450uL RT 1x PBS (Phosphate-buffered saline, Invitrogen, 10010023) was added and cells resuspended by vortexing. Immediately before loading onto luciferase readout plates, cells were vortexed again to ensure even suspension. For each biological replicate, 3 wells of a white, flat-bottomed 96-well plate were each loaded with 50uL of cell-PBS suspension. After a full plate was loaded with sample, 50uL of Dual-Glo Luciferase reagent was added and mixed by multichannel pipette. The plate was incubated for a minimum of 10 minutes prior to readout (maximum 30min). Plates were read for Firefly signal at 10 reads per well, removed, and 50uL of Stop & Glo reagent added and mixed by multichannel pipette. Plates were again incubated for a minimum of 10 minutes, and Renilla signal read at 10 reads per well.

4.6.6 Analysis of Luciferase Expression Data

The 10 reads taken per well were averaged separately for Firefly and Renilla signal to get well values. Background signal was calculated by averaging reads across 6 wells (2 biological replicates) of untransfected cell controls. Background Firefly signal was subtracted from each Firefly technical replicate to get background-adjusted values. This was repeated for Renilla signal. Adjusted Firefly technical replicates were then divided by the corresponding adjusted Renilla value for that well. A low-expression control condition was chosen for value normalization (typically a promoter-only plasmid), which varied depending on the experiment. Background adjusted F/R tech rep values were individually divided by this control (average of all of its F/R biological replicate values). This gave a final fold-change value.

Normalized F/R = [(F - F cell background)/(R - R cell background)]

Fold change = (normalized F/R for tech rep) / (bio rep average of normalized F/R for control condition)

Technical replicates were then averaged to get biological replicate values. Biological replicates were averaged and final fold change values plotted on graphs. Where multiple reads for a condition were taken across different plates or days, or where values are represented as % instead of fold change, values were normalized to correct for variation and in order to be able to compare replicates. Normalization was done by dividing all values by the 'high' control readout for that plate - the same high-expressing control plasmid used on every plate within an experiment (for example CMVe-SV40p).

4.6.7 Statistical Testing

Error of biological replicates is given as the standard error for all graphs [standard deviation(bio reps)/square root(# bio reps)]. T-testing was done comparing biological replicates using a one-sided, heteroscedastic model and significance indicated for that comparison by * if the test showed $p<0.05$. Unless otherwise indicated, on bar charts showing fold changes in luciferase act: error bars represent standard error of three or more biological replicates, circles show individual biological replicate values, and the number above each bar is the average value of the three biological replicates. Where there were replicates from multiple plates/days of the same condition, and normalization to the high-expressing control was done, t-testing was done pre- and post-normalization and if either one of these t-tests gave ≥0.05, the conditions were listed as not significant. Percent change in expression for si and eb assays was determined by dividing all fold change values by the fold change of the 'high' expression control for that plate (si or eb control), to give values relative to that control as 100%.

4.7 Notes and Acknowledgements

Thanks especially to Sheila Rasouli and Diana Davis who worked on the PCR isolation of DHS regions, and their screening and cloning into pEntra1 vectors. Thanks to John Moran and Shigeki Iwase who generously allowed me to use their equipment for electroporation and readout. Thanks to Greg Farnum for construction of the tricky 4x F2/3 repetitive enhancer blocker element using Emma Assembly [161] and for establishing the EMMA system in our lab. Thanks to Alan Boyle, Jessica Switzenberg, Torrin McDonald and Sierra Nishizaki-Sweiso for feedback on plasmid design and

thoughts on data interpretation. Special thanks to Jessica for assisting with feedback on many complex cloning situations.

CHAPTER V
Conclusions and Future Directions

5.1 Reporter Assay Design: Modeling Enhancer-Promoter Spacing

In Chapter 2 of this thesis, I addressed the ways that the complexities of sequence- and context-dependencies and interactions of CREs can complicate reporter assay design, and addressed ways to account for these interactions, using supporting data from my work. These results also emphasized the ways that plasmid-based assays can be useful models for CRE activity which are consistent with the elements' genomic behavior. The goal of Chapter 2 is to more thoroughly address the different design components of plasmid-based reporter assays. While I touch on many effects in this chapter, the most important for follow-up studies is the effect of enhancer-promoter spacing.

While the effects of enhancer-promoter spacing are evident in a handful of studies going back decades, the focus of these studies was often other aspects, and this distance effect is mentioned only in passing. Additionally, the context of its impact on reporter assay design particularly in studies of enhancer blockers has not been investigated thoroughly. Two recent studies supporting this effect combine with my data to demonstrate that decreased expression with increased enhancer-promoter distance occurs: at the level of a hundred bases with a minimal paired TFBS [137], at a couple hundred base pairs distance in a full enhancer element (my data), and across a megabase genomic scale with a full enhancer element in a genomic context [138].

These results not only have implications for assay design (discussed in Chapter 2) but may represent a crucial development in our understanding of enhancer function. Many models of enhancer mechanism focus on the importance of enhancer looping for promoter contact, or they treat proximal enhancers as already close enough for contact. However, there might be an intermediate space, where proximity of an enhancer to its target promoter affects action, on a scale where there is not enough distance for looping

to be an efficient method of contact (perhaps due to topological constraints to DNA folding). On a more distal scale, looping can occur, but might occur slightly more rapidly if there is less DNA to loop out due to the enhancer being closer, increasing burst frequency due to more frequent enhancer contact.

Future testing of this phenomenon should leverage both episomal and native systems to interrogate the dynamics of enhancer-promoter (e-p) distance-dependence. One primary gap to address regarding this effect is the precise distance-dependencies of expression for full enhancer elements. While the Davis et al. paper thoroughly tested e-p distances, they did not use an enhancer but rather a pair of single TF binding sites, and only tested up to 190bp from the promoter [137]. They did, however, establish a very robust model system, iterating single base pair distances, across different sequence backgrounds, and testing in both an episomal and genomic contexts. By comparison, Zuin et al. tested greater distances, but due to the random integration nature of their assay they could not control precise e-p distances. Additionally they tested distances on the kilobase to megabase scale [138].

Future work on this topic should leverage the Davis et al. system but test full enhancer elements with e-p distances ranging from 0bp-2kb to fill this intermediate range gap, which is also the distance most relevant to reporter assays. This would expand on my initial test, but include more robust controls and iterate more fine distances rather than the 250bp I initially used. One important control is using multiple alternate sequence backgrounds to control for sequence-context-specific effects that could influence expression. These would also be tested episomally and integrated genomically to determine whether this e-p distance effect on expression is impacted by chromatin state. One reason this crucial distance gap may not have been covered is that library-scale oligo synthesis is limited to ~200bp. In order to iterate over 2kb distances, plasmids must be created without using library synthesis, limiting iteration of distance to every 50bp, rather than every 1bp, requiring 40 plasmids per background. This can be achieved most elegantly by inserting the chosen enhancer into the test plasmid containing the target promoter at the initial 50bp distance, with a fixed restriction insertion site between the two elements. The spacing sequence can then be serially copied via PCR to generate larger and larger fragments increasing by 50bp,

which is then cloned into place between the enhancer and promoter. To account for spacer-specific effects, a library with the spacer fragments upstream of the promoter, but no enhancer, can be used to normalize expression for matched e-p versions. This can be done using barcoded plasmids such that while initial cloning might be laborious, readout can be done in bulk via mRNA expression from the gene used for readout.

These results, in combination with the results from Zuin et al. and Davis et al. have the potential to advance our understanding of the mechanisms by which enhancers contact and activate promoters. Comparing genomic vs episomal results will help determine whether the differences between chromatinized vs plasmid DNA effect distance-dependent effects. A lack of difference could indicate that the effect is solely based on the relationship between DNA folding and conformational likelihoods and kinetics of the frequency of e-p contact is driven by the frequency at which certain DNA loops can occur.

Another crucial test that could be done using this system is to distinguish whether enhancer-promoter contact occurs through a mixed tracking looping mechanism, with tracking occurring at sub-kb distances and looping occurring after some specific distance cutoff (discussed in Chapter 2). Using the genomically integrated version of the plasmids generated for this assay, a set of LacO sites can also be integrated between the enhancer and promoter sequences (during plasmid cloning, prior to genomic integration). Use of LacO in mammalian cells precludes the possibility of a CTCF-based enhancer blocker loop forming to isolate the enhancer from the promoter, as LacO is not native to mammalian cells. We would then express LacI, which would bind to the LacO sites and physically obstruct progression of RNA Pol II should a tracking mechanism occur. However, at distances where looping occurs, the enhancer and promoter should be able to loop in such a way as to establish contact at the base of the loop, with the LacO obstructing element looped out. The use of LacO/I system to test enhancer tracking has previously been established to test tracking of HS2e [64].

Depending on results for the cutoff distance in these tests where expression dropoff occurs (observed in the Davis et al. paper and my data), further testing could be done on a limited subset of the conditions representing the distances just before and after the cutoff. This would allow for inclusion of different spacer sequences to control

for spacer-specific effects, and importantly different enhancer-promoter combinations. Thus far, whether this effect or its location or strength is universal or CRE-specific has not been established.

More broadly, these studies represent an important direction for the study of CREs. Through the use of high-throughput scaled episomal and genomic assays, CRE mechanisms can be interrogated on a scale not previously available. These assays have often been applied to discovery of new enhancers in different cell types, which is crucial work, but should not be overlooked for their ability to provide robust results when applied to single mechanisms. Chapter 2 demonstrates that the complexity of sequence-dependent interactions in plasmids makes them an excellent tool for characterizing these effects, in spite of some differences between episomal and chromatin contexts. Their malleability and the ease of producing sequences at scale provide a powerful tool to go back and re-examine more thoroughly some phenomenon hinted at in earlier literature, where investigators were first establishing the way plasmid sequences behaved but did not have the ability to clone at scale available to us today.

5.2 Using OnRamp to Improve Replicability in Plasmid-Based Research

The results from Chapter 2 showed the potential for functional impact that every piece and sequence of a plasmid can have, making the current standard of spot-checking 1kb sequences of plasmids using Sanger sequencing insufficient to ensure full validation. In chapter three I addressed this gap, by presenting a nanopore-based method for plasmid sequencing which generates full plasmid sequences, PCR-free, with timing and costs similar to that of Sanger sequencing. This protocol and the associated analysis pipeline built into a web tool, are built to make nanopore plasmid sequencing rapid and widely accessible.

Plasmid validation using nanopore seems to be increasingly recognized as an important advancement, as it is being addressed by a number of groups. However, while our method, OnRamp, and these other methods will allow for full-plasmid sequencing in labs with high levels of plasmid cloning where investment in a nanopore system is justified, for other labs this may not be feasible. Companies and university cores providing nanopore plasmid sequencing in a parallel manner to the way Sanger

sequencing cores and companies currently function will hopefully fill this gap, allowing for broader adoption of full-plasmid sequencing as standard practice for any lab running regular plasmid cloning. For labs where adoption of the nanopore platform *is* feasible for plasmid validation, but who otherwise are not invested in learning the programming necessary to run the analysis, or for labs who run data through a device shared with another lab, OnRamp provides a solution for rapid, user-friendly analysis centered on reference-based plasmid sequencing.

An increased awareness of the importance of full plasmid validation will also be an important factor in adoption of this method for validation. This has recently been addressed in part by the group which originally published the STARR-seq assay, one of the cornerstone MPRAs in the field [185]. However this group only addressed the potential functional impact of one element in one plasmid system in that paper. Klein et al. more recently published a paper attempting to iterate various commonly used plasmid designs and test their impact on readout of expression by testing the use of different element (insert element, gene, barcode) arrangements [183]. They did find that element arrangement impacted reproducibility, supporting how variation in plasmid design across different labs and different assays can impact interpretation. They did not however address a second cause of variation between results across labs: how undetected variation in these plasmid sequences can impact function.

I believe an important next step is to more broadly characterize the range of *previously undetected* natural variation that occurs across plasmid systems within and across labs. This is important to establish a few key things; at what rate this effect actually occurs, how often it actually impacts function, and what the likely cause of variation is. Answering these questions could be a huge step toward determining how much of a problem inter-plasmid variation due to cloning is, and depending on whether it impacts function and in what specific patterns it occurs, what can be done to address this issue. OnRamp and nanopore-based full-plasmid sequencing provide an opportunity to address these questions and greatly improve the reproducibility of research within the field of regulatory study.

I propose two experiments to attempt to establish rates of plasmid variation. The first would include a sequencing survey of plasmid sequences across labs. I propose to

collect multiple plasmids from across multiple labs within the University of Michigan and across the country, where labs are willing to participate. Labs would provide multiple plasmids, both multiples with similar backbones and where possible with differing backbones. Information on the lab's predicted sequences for the plasmids as well as validation method used to generate the clones would be collected for comparison. Plasmids would not be collected that had were directly obtained, and not manipulated, from distributors that complete sequence validation. Plasmids would then be directly sequenced using OnRamp, and the expected sequences provided by the labs compared to OnRamp's consensus sequences. Variant locations and types could then be mapped. Degree of discrepancy could be compared to validation methods used by the lab to determine which validation methods may contribute to the least and most undetected variation. For the subset of plasmids where variation did occur, and where the plasmid contains an expression system that can be used, a paired version of the plasmid would be generated that removes the novel variant and restores the expected sequence. The variant and non-variant plasmid versions would then be transfected to test whether the variation impacts expression as read by mRNA levels.

For this project, a key limitation would be willingness of labs to participate. To mitigate this, lab or contributor names would not be associated directly with sequences in any publications except in aggregate, such that we would not be publishing which labs had the best versus poorest sequence quality. Additionally, plasmid ownership and distribution limitations would have to be addressed. Finally, it is unlikely labs would want plasmid sequences to reveal novel design innovations, so plasmids sequences would need to be only partially published/made available. While this work does not directly answer a specific question regarding CRE biology, it could have a large impact on improving the quality of CRE research and address once source for well-established issues with reproducibility. It would also help address whether full-plasmid sequencing is crucial to implement for universities and labs, or whether these variations are less frequent or impactful than previously thought, improving confidence in our current methods.

A second, complimentary project would involve characterizing the rate of clonal error within a lab. Within our lab, we would insert a cloned sequence using 5 different

cloning methods to a plasmid backbone. These would include: blunt and sticky restriction digest and ligation, Gateway cloning, Gibson cloning, and PCR-based insertion using tailed primers. This would be repeated in triplicate for each plasmid and method. 10 clones would be picked for each plasmid-method combination, grown, barcoded, analyzed by restriction digest and gel, and sequenced using OnRamp. This would allow us to directly establish the rates of unwanted variation induced by each method. Additionally, by comparing variant sizes and gel vs OnRamp results, we could determine also the rate at which variation would be missed by this method. Finally, by looking across all 150 plasmid clones, we could determine the frequency of bacterial-replication-induced variation, as measured by variants which occurred outside the region -/+50bp from the cloning insertion site. Establishing variant frequency and detection rates by gel and diagnostic digest would also complementarily support the degree to which full-plasmid sequencing is or is not necessary during cloning. It would also for the first time establish the rate at which diagnostic gel digest can miss variation.

Finally, looking at variation rates for bacterial replication errors outside of cloned regions would strongly impact our understanding of the reproducibility of results from high-throughput regulatory assays (MPRAs). This is because typical MPRA readouts make use of Illumina sequencing, which uses short fragment reads. Consequently, backbone variation cannot be uniquely mapped and results for that plasmid discarded, as backbone sequence cannot be uniquely mapped when the entire library contains identical backbone sequence. Determining the rate of bacterial-induced backbone errors can help determine whether steps need to be taken to improve backbone validation of MPRA libraries subsequent to cloning.

5.3 Further Characterizing the *PRDM1-ATG5* Domain

In Chapter 4 of this thesis I characterize the activities, combinatorial interactions, and regulatory state of CREs in the *PRDM1-ATG5* region of the human genome. I identify a tissue-conserved strong enhancer, DHS 16 which is a candidate enhancer for ATG5, characterize its TF binding, chromatin state, and establish relative functional contributions of its component sequences. I also identify patterns of position-dependent enhancer-enhancer synergy with the HS2 enhancer.

My goals in studying the *PRDM1-ATG5* region were multifaceted. I initially hoped to characterize novel examples of silencer elements which were at the time very limited. Additionally, I hoped not just to study these elements in isolation, but to answer two larger questions that could not be answered just using a high-throughput silencer assay. First, regarding how silencers behave in the context of a larger regulatory unit, and second how they fit into the logic structure of gene regulation - are they active alongside enhancers? Do they actively mediate cell-type specific differences in expression, or are they active on a much more transient scale, establishing a repressive chromatin state and then becoming inactive? Characterizing multiple CREs using a low-throughput assay across a single domain seemed an optimal way to interrogate the combinatorial interactions and contributions of CREs. The region was chosen for its interesting expression dynamics, where the two genes were differentially expressed within and across the two cell lines studies, and where differential sub-TAD and chromatin structure was involved. It was also chosen to contribute to our understanding of the regulation of *ATG5* and *PRDM1*, both of which play important roles in immune function, development and disease. While K562 and HepG2 cell lines are not the most pertinent lines for the study of PRDM1 function in a particular disease (ATG5 functions broadly and is relevant to almost any tissue type), they are tractable and highly-characterized lines. I did not expect to see, but did find, that the larger organizing principles of this region actually apply across many different, and more relevant cell lines. TAD structure, CTCF binding sites, and a number of DHS regions had conserved accessibility and TF binding across a number of lines. By establishing this initial model, in future CRE function could be tested across other cell lines, and through comparison, functional inferences drawn.

The two regions identified as having potential silencer for this function acted as silencers in HepG2, as expected given its more repressed state, however they were not genomically active. Further study of these regions is warranted, in a model cell system where the *PRDM1* region is in the process of undergoing heterochromatinization, in order to determine whether these elements are transiently active. Additionally, remaining DNase-seq accessibility regions should be tested, as it is possible other silencers were missed. However, it seems that the lack of active silencer elements in

this region does provide information regarding the silencing mechanisms. This region provides an example where silencing appears to be mediated through heterochromatin spreading, possibly through the loss of CTCF binding at an element which behaves as an enhancer.

Having established and begun to characterize this region, there are many experiments which could be completed to build upon this foundation. 4C contact data for PRDM1 and ATG5 promoters in these cell lines should be generated to determine which enhancers contact which promoters. Additionally, CRISPR-Cas9 mediated genomic deletion of DHS 16 to determine effect on ATG5 and PRDM1 expression measured by RNAseq (which could be complicated by cell death, if ATG5 is affected strongly enough) would help determine which genes it regulates. Similar deletions of single or combined DHS with contact evidence for PRDM1 could help tease out whether there is multi-enhancer synergy and redundancy in the region, and to what degree each DHS is necessary for proper PRDM1 expression. To support the potential roles of DHS 3 and DHS 25.1 in preventing the spread of heterochromatin in K562 and HepG2 respectively, I would measure heterochromatin and expression of PRDM1 and ATG5 after deletion of these elements.

Silencer and enhancer blocker testing was not completed in HepG2 cells, as the HS2 enhancer used for these assays is not active in HepG2 cells. However in the enhancer assay, DHS 11 and 30 both showed putative silencer activity. It would be interesting to confirm this activity in the si and eb assays using a HepG2-active enhancer. Given the positioning of DHS 11 in the middle of the heterochromatin domain in HepG2, perhaps it is active at a transitory point during cell differentiation and is responsible for establishing a silenced state in the region, but then becomes inactive once it is established. DHS 11 is accessible in GM12878, CD4+ and CLL cell lines, supporting a lymphocyte-related function, and is bound by YY1, a transcription factor with both repressing activating functions, in GM12878 and K562 cells.

In an episomal context, I would selectively test the remaining few DHS from K562 and HepG2 that were not tested here, from further up- and down-stream. I would also further interrogate the combinatorial and synergistic activity of the enhancer DHS in K562 in a manner more relevant to the regions biology by testing the activity of the other

DHS alongside DHS 16 or each other, rather than HS2e, in an episomal context, perhaps in conjunction with the PRDM1 promoter in place of the SV40 promoter. (The PRDM1 promoter was tested, but was not used as a constitutive promoter in the plasmid assays as its activity is very low in K562 (data now shown) and so might make detection and differentiation of DHS effects challenging).

Follow-up studies which build on the foundation established in this region have the potential to address many interesting current questions in the field of gene regulation, and to model how the different principles and layers of gene regulation (chromatin states, redundancy, cell-type specificity, synergy, CRE-CRE contacts) combine for a single domain to regulate expression.

5.4 Characterizing *Cis*-Regulatory Elements: Limitations of Contact Data

Outside of the specific mechanisms of *PRDM1* and *ATG5* regulation, the results in Chapter 4 reveal important areas for ongoing research in *cis*-regulatory element biology in general. One of the primary limitations of the episomal approach used in Chapter 4 is not the potential differences between episomal and chromatin context. In fact, as discussed, for 4/5 of the regions tested, the classification assigned by episomal testing is supported by native indicators including TF binding and histone modifications. However, a primary limitation of the episomal approach is the inability to link a particular DHS to regulation of its target gene. This data can only be obtained through testing the native sequences. CRE-gene linking is essential both to improving our understanding of CRE-gene networks and how and when inter-element coordination occurs, and to linking non-coding genetic variation in humans to specific disease mechanisms.

This is also one of the primary limitations of data in the field of CRE biology. Past studies have relied on enhancer or silencer proximity to a promoter as a proxy for regulation, however this is not direct evidence and is not always accurate. This also limits assignment of enhancers located more distally to target promoters. A more promising technique used co-expression of enhancer RNAs with the timing of mRNA expression subsequent to activation of the cells using a cell signal [71]; [77]. Again however, this evidence provides correlation, not direct evidence of regulation. The best methods available to date are chromatin conformation capture methods, including 3C,

4C, and HiC [68]. These methods link enhancers and silencers to promoters by using physical proximity of the DNA sequences of the elements in a cell population. However a limitation to these methods is resolution of these results. More recently, Hsieh et al. improved chromatin capture resolution with MicroC [339], however this new method has not been yet applied to the many commonly studied cell lines and tissues and so data availability is limited, and mapping these contacts requires not insignificant investment of resources to obtain the high-throughput data. For smaller-scale studies (such as that presented here on *PRDM1* and *ATG5*), 3C or 4C are needed to link CREs to their cognate promoters, but these must be repeated for each specific gene or region as they require selection of a target or 'bait' sequence for which contacts are identified. Future work in the field of regulation is significantly limited by the lack of 1kb-resolution contact data across many cell lines. However work is in progress to address this and the number of cell lines and tissue types with available contact data is increasing.

Another key tool, which would be a primary next assay to use for the PRDM1-ATG5 regulatory domain characterization, is CRISPR-mediated deletions and disruptions. This also mirrors the state of larger field in general, as while it has emerged for use in studying CRE activity, there are still limitations of time and scale. Use of CRISPR/Cas9 facilitates necessity testing in native genomic contexts. Using sgRNAs Cas9 protein can be directed to target specific sequences for deletion, if sgRNAs are designed to target sequences flanking the region of interest, as Cas9 makes double-stranded DNA breaks. Setting aside known limitations of off-target cutting and the availability of genomic sgRNA target sites near the region of interest, larger methodological limitations remain. The key readout to link deletion of a CRE to changes in gene expression is either qPCR or RNA-seq. RNA-seq captures aggregate RNA expression data, and samples would be compared to detect gene expression changes pre- vs post- deletion. This allows for unbiased detection of expression changes in any gene, potentially allowing detection of genes which are distally located and regulated by the CRE. However bulk RNA-seq is often overpowered and resource-intensive for the study of a single gene. However with qPCR, target genes must be chosen for readout using specific primers, which pre-limits detection of changes to the chosen genes and precludes detection of other regulated genes. This is potentially additionally confounded

by enhancer redundancy, where deletion of one enhancer may not have a detectable impact if other enhancers up-regulate or activate to compensate.

In summary, while both chromatin conformation capture and CRISPR deletion methods are crucial for directly linking CREs to their target genes, both have limitations linked to either scale or resolution. Additionally, they are both needed together to complement each other alongside episomal results in order to improve confidence of gene-CRE links. However this becomes limiting at a certain scale for medium-throughput studies like the one in Chapter 4. They are more suited to the study of single elements or on a genome-wide scale but can be limiting on the level of a TAD. They are, however, crucial to mapping and characterizing CRE-gene interactions. Improvements to these methods that increase resolution, for HiC, improve scale limitations, for CRISPR, and allow for capture of data on a single-cell rather than population-level scale, will be important advances for the future of the field.

5.5 Concluding Remarks

The work in this thesis provides a methodical overview of many of the core functional aspects of reporter assay design, and addresses the importance of these aspects to correct interpretation of assay results as well as tools to address potential pitfalls. It also presents novel data on the complete set of paired position- and orientation-dependencies of the cHS4 core CTCF-binding enhancer blocker element in a plasmid context, and data supporting the impact of enhancer-promoter distance on gene expression in episomal assays. It presents a novel tool for barcode-free preparation, sequencing, and analysis of full plasmid sequences for validation using nanopore sequencing and demonstrates the tool's function using real and simulated plasmid data. Finally, it contributes to the mapping and characterization of CRE activity in the regulatory space of the *PRDM1-ATG5*-containing domain of the human genome.

Despite the adoption of high-throughput regulatory assays, much of the human genome remains uncharacterized, despite non-coding sequences increasingly being understood to play an important role in human development and disease. This work attempted to address aspects of two primary limitations to this characterization. The reliance on plasmid-based assays for characterization is reasonable given their

flexibility and the ease of cloning, however a full understanding of the regulatory logic of any plasmid constructs as well as the ability to fully validate plasmid sequences is essential to the ability to obtain useful results. Additionally, once these well-designed and validated assays are applied, individual studies are needed to layer the results from across the many high-throughput assays, combine them together, and ground them in genomic context for a particular location. Without this, the assays will continue to be useful for establishing general principles and patterns of activity, a very important goal, but without other groups focused on contextualizing and interpreting the data, loci in the human genome will remain largely uncharacterized.

BIBLIOGRAPHY

1. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature. 1997. pp. 251–260.

2. Li G, Reinberg D. Chromatin higher-order structures and gene regulation. Curr Opin Genet Dev. 2011;21: 175–186.

3. Noonan JP, McCallion AS. Genomics of long-range regulatory elements. Annu Rev Genomics Hum Genet. 2010;11: 1–23.

4. Lelli KM, Slattery M, Mann RS. Disentangling the Many Layers of Eukaryotic Transcriptional Regulation. Annual Review of Genetics. 2012. pp. 43–68.

5. Istrail S, Davidson EH. Logic functions of the genomic cis-regulatory code. Proceedings of the National Academy of Sciences. 2005. pp. 4954–4959.

6. Hotchkiss RD. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. J Biol Chem. 1948;175: 315–332.

7. Holliday R, Pugh JE. DNA Modification Mechanisms and Gene Activity During Development. Science. 1975;187: 226–232.

8. Compere SJ, Palmiter RD. DNA methylation controls the inducibility of the mouse metallothionein-I gene lymphoid cells. Cell. 1981;25: 233–240.

9. Schulz WA, Steinhoff C, Florl AR. Methylation of endogenous human retroelements in health and disease. Curr Top Microbiol Immunol. 2006;310: 211–250.

10. Zwart R, Sleutels F, Wutz A, Schinkel AH, Barlow DP. Bidirectional action of the *Igf2r* imprint control element on upstream and downstream imprinted genes. Genes & Development. 2001. pp. 2361–2366.

11. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, et al. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell. 2008;30: 755–766.

12. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013;38: 23–38.

13. Tilghman SM, Tiemeier DC, Seidman JG, Peterlin BM, Sullivan M, Maizel JV, et al. Intervening sequence of DNA identified in the structural portion of a mouse beta-globin

gene. Proc Natl Acad Sci U S A. 1978;75: 725–729.

14. Gilbert W. Why genes in pieces? Nature. 1978;271: 501.

15. Lutz CS. Alternative polyadenylation: a twist on mRNA 3’ end formation. ACS Chem Biol. 2008;3: 609–617.

16. Rottman FM, Bokar JA, Narayan P, Shambaugh ME, Ludwiczak R. N6-adenosine methylation in mRNA: substrate specificity and enzyme complexity. Biochimie. 1994;76: 1109–1114.

17. Terns M, Terns R. Noncoding RNAs of the H/ACA family. Cold Spring Harb Symp Quant Biol. 2006;71: 395–405.

18. Guschina E, Benecke B-J. Specific and non-specific mammalian RNA terminal uridylyl transferases. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2008. pp. 281–285.

19. Rodriguez AJ, Czaplinski K, Condeelis JS, Singer RH. Mechanisms and cellular roles of local protein synthesis in mammalian cells. Curr Opin Cell Biol. 2008;20: 144–149.

20. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet. 2008;9: 102–114.

21. Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. Nat Rev Genet. 2010;11: 75–87.

22. Lee MJ, Yaffe MB. Protein Regulation in Signal Transduction. Cold Spring Harb Perspect Biol. 2016;8.

23. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 2009. pp. 9362–9367.

24. Poulos RC, Sloane MA, Hesson LB, Wong JWH. The search for cis-regulatory driver mutations in cancer genomes. Oncotarget. 2015;6: 32509–32525.

25. Conaway JW, Florens L, Sato S, Tomomori-Sato C, Parmely TJ, Yao T, et al. The mammalian Mediator complex. FEBS Lett. 2005;579: 904–908.

26. Orphanides G, Lagrange T, Reinberg D. The general transcription factors of RNA polymerase II. Genes Dev. 1996;10: 2657–2683.

27. Vanaja A, Yella VR. Delineation of the DNA Structural Features of Eukaryotic Core Promoter Classes. ACS Omega. 2022;7: 5657–5669.

28. Grünberg S, Hahn S. Structural insights into transcription initiation by RNA polymerase II. Trends Biochem Sci. 2013;38: 603–611.

29. Juven-Gershon T, Kadonaga JT. Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev Biol. 2010;339: 225–229.

30. Ptashne M, Gann A. Transcriptional activation by recruitment. Nature. 1997;386: 569–577.

31. McKnight SL, Kingsbury R. Transcriptional control signals of a eukaryotic protein-coding gene. Science. 1982;217: 316–324.

32. Claessens F, Gewirth DT. DNA recognition by nuclear receptors. Essays Biochem. 2004;40: 59–72.

33. Lonard DM, O'Malley BW. Expanding functional diversity of the coactivators. Trends Biochem Sci. 2005;30: 126–132.

34. Ogbourne S, Antalis TM. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. Biochem J. 1998;331 (Pt 1): 1–14.

35. Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, et al. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Mol Cell. 2015;58: 1101–1112.

36. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, et al. RNA exosome depletion reveals transcription upstream of active human promoters. Science. 2008;322: 1851–1854.

37. Bernstein BE, Humphrey EL, Erlich RL, Schneider R, Bouman P, Liu JS, et al. Methylation of histone H3 Lys 4 in coding regions of active genes. Proc Natl Acad Sci U S A. 2002;99: 8695–8700.

38. Lee C-K, Shibata Y, Rao B, Strahl BD, Lieb JD. Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet. 2004;36: 900–905.

39. Schübeler D, MacAlpine DM, Scalzo D, Wirbelauer C, Kooperberg C, van Leeuwen F, et al. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev. 2004;18: 1263–1271.

40. Kadonaga JT. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell. 2004;116: 247–257.

41. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet. 2006;7: 29–59.

42. Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated

with preaxial polydactyly. Hum Mol Genet. 2003;12: 1725–1735.

43. Moreau P, Hen R, Wasylyk B, Everett R, Gaub MP, Chambon P. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. Nucleic Acids Res. 1981;9: 6047–6068.

44. Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell. 1981;27: 299–308.

45. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010;107: 21931–21936.

46. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, David Hawkins R, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics. 2007. pp. 311–318.

47. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74.

48. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129: 823–837.

49. Arnosti DN, Kulkarni MM. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem. 2005;94: 890–898.

50. Merika M, Thanos D. Enhanceosomes. Curr Opin Genet Dev. 2001;11: 205–208.

51. Jindal GA, Farley EK. Enhancer grammar in development, evolution, and disease: dependencies and interplay. Dev Cell. 2021;56: 575–587.

52. Beagrie RA, Pombo A. Gene activation by metazoan enhancers: Diverse mechanisms stimulate distinct steps of transcription. Bioessays. 2016;38: 881–893.

53. Sauer F, Hansen SK, Tjian R. Multiple TAFIIs directing synergistic activation of transcription. Science. 1995. pp. 1783–1788.

54. Bashor CJ, Patel N, Choubey S, Beyzavi A, Kondev J, Collins JJ, et al. Complex signal processing in synthetic gene circuits using cooperative regulatory assemblies. Science. 2019;364: 593–597.

55. Panigrahi A, O'Malley BW. Mechanisms of enhancer action: the known and the unknown. Genome Biol. 2021;22: 108.

56. Tunnacliffe E, Chubb JR. What is a transcriptional burst? Trends Genet. 2020;36: 288–297.

57. Larsson AJM, Johnsson P, Hagemann-Jensen M, Hartmanis L, Faridani OR,

Reinius B, et al. Genomic encoding of transcriptional burst kinetics. Nature. 2019;565: 251–254.

58. Fukaya T, Lim B, Levine M. Enhancer Control of Transcriptional Bursting. Cell. 2016;166: 358–368.

59. Walters MC, Fiering S, Eidemiller J, Magis W, Groudine M, Martin DI. Enhancers increase the probability but not the level of gene expression. Proc Natl Acad Sci U S A. 1995;92: 7125–7129.

60. Furlong EEM, Levine M. Developmental enhancers and chromosome topology. Science. 2018;361: 1341–1345.

61. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. Science. 1998;281: 60–63.

62. Bulger M, Groudine M. Looping versus linking: toward a model for long-distance gene activation. Genes Dev. 1999;13: 2465–2477.

63. Kong S, Bohl D, Li C, Tuan D. Transcription of the HS2 enhancer toward a cis-linked gene is independent of the orientation, position, and distance of the enhancer relative to the gene. Mol Cell Biol. 1997;17: 3955–3965.

64. Ling J, Ainol L, Zhang L, Yu X, Pi W, Tuan D. HS2 Enhancer Function Is Blocked by a Transcriptional Terminator Inserted between the Enhancer and the Promoter*[boxs]. J Biol Chem. 2004;279: 51704–51713.

65. Schoenfelder S, Fraser P. Long-range enhancer–promoter contacts in gene expression control. Nat Rev Genet. 2019;20: 437–455.

66. Robson MI, Ringel AR, Mundlos S. Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D. Mol Cell. 2019;74: 1110–1122.

67. Panigrahi AK, Foulds CE, Lanz RB, Hamilton RA, Yi P, Lonard DM, et al. SRC-3 Coactivator Governs Dynamic Estrogen-Induced Chromatin Looping Interactions during Transcription. Mol Cell. 2018;70: 679–694.e7.

68. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295: 1306–1311.

69. Ashe HL, Monks J, Wijgerde M, Fraser P, Proudfoot NJ. Intergenic transcription and transinduction of the human β-globin locus. Genes Dev. 1997;11: 2494–2509.

70. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, et al. A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. PLoS Biol. 2010;8: e1000384.

71. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An

atlas of active enhancers across human cell types and tissues. Nature. 2014;507: 455–461.

72. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465: 182–187.

73. Kouno T, Moody J, Kwon AT-J, Shibayama Y, Kato S, Huang Y, et al. C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. Nat Commun. 2019;10: 360.

74. Natoli G, Andrau J-C. Noncoding transcription at enhancers: general principles and functional models. Annu Rev Genet. 2012;46: 1–19.

75. Young RS, Kumar Y, Bickmore WA, Taylor MS. Bidirectional transcription marks accessible chromatin and is not specific to enhancers.

76. Mikhaylichenko O, Bondarenko V, Harnett D, Schor IE, Males M, Viales RR, et al. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev. 2018;32: 42–57.

77. Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science. 2015. pp. 1010–1014.

78. Brand AH, Breeden L, Abraham J, Sternglanz R, Nasmyth K. Characterization of a “silencer” in yeast: A DNA sequence with properties opposite to those of a transcriptional enhancer. Cell. 1985. pp. 41–48.

79. Jiang J, Cai H, Zhou Q, Levine M. Conversion of a dorsal-dependent silencer into an enhancer: evidence for dorsal corepressors. EMBO J. 1993;12: 3201–3209.

80. Siu G, Wurster AL, Duncan DD, Soliman TM, Hedrick SM. A transcriptional silencer controls the developmental expression of the CD4 gene. EMBO J. 1994;13: 3570–3579.

81. Qi H, Liu M, Emery DW, Stamatoyannopoulos G. Functional validation of a constitutive autonomous silencer element. PLoS One. 2015;10: e0124588.

82. Bruce AW, Donaldson IJ, Wood IC, Yerbury SA, Sadowski MI, Chapman M, et al. Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. Proc Natl Acad Sci U S A. 2004;101: 10458–10463.

83. Gisselbrecht SS, Palagi A, Kurland JV, Rogers JM, Ozadam H, Zhan Y, et al. Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. Mol Cell. 2020;77: 324–337.e8.

84. Doni Jayavelu N, Jajodia A, Mishra A, Hawkins RD. Candidate silencer elements for the human and mouse genomes. Nat Commun. 2020;11: 1061.

85. Pang B, Snyder MP. Systematic identification of silencers in human cells. Nat Genet. 2020;52: 254–263.

86. Ngan CY, Wong CH, Tjong H, Wang W, Goldfeder RL, Choi C, et al. Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. Nature Genetics. 2020. pp. 264–272.

87. Huang D, Petrykowska HM, Miller BF, Elnitski L, Ovcharenko I. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. Genome Res. 2019;29: 657–667.

88. Courey AJ, Jia S. Transcriptional repression: the long and the short of it. Genes Dev. 2001;15: 2786–2796.

89. Petrykowska HM, Vockley CM, Elnitski L. Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. Genome Res. 2008;18: 1238–1246.

90. Zhang Y, See YX, Tergaonkar V, Fullwood MJ. Long-Distance Repression by Human Silencers: Chromatin Interactions and Phase Separation in Silencers. Cells. 2022;11.

91. Halfon MS. Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. Trends Genet. 2019;35: 93–103.

92. Pang B, van Weerd JH, Hamoen FL, Snyder MP. Identification of non-coding silencer elements and their regulation of gene expression. Nat Rev Mol Cell Biol. 2022.

93. Gaston K, Jayaraman PS. Transcriptional repression in eukaryotes: repressors and repression mechanisms. Cell Mol Life Sci. 2003;60: 721–741.

94. Segert JA, Gisselbrecht SS, Bulyk ML. Transcriptional Silencers: Driving Gene Expression with the Brakes On. Trends Genet. 2021;37: 514–527.

95. Halfon MS. Silencers, Enhancers, and the Multifunctional Regulatory Genome. Trends in genetics: TIG. 2020. pp. 149–151.

96. Geyer PK, Spana C, Corces VG. On the molecular mechanism of gypsy-induced mutations at the yellow locus of Drosophila melanogaster. EMBO J. 1986;5: 2657–2662.

97. Meers MP, Janssens DH, Henikoff S. Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded. Mol Cell. 2019;75: 562–575.e5.

98. Jenuwein T, Allis CD. Translating the histone code. Science. 2001;293: 1074–1080.

99. Felsenfeld G, Groudine M. Controlling the double helix. Nature. 2003;421: 448–453.

100. Udvardy A, Maine E, Schedl P. The 87A7 chromomere. Journal of Molecular Biology. 1985. pp. 341–358.

101. Kellum R, Schedl P. A position-effect assay for boundaries of higher order chromosomal domains. Cell. 1991;64: 941–950.

102. Cubeñas-Potts C, Corces VG. Architectural proteins, transcription, and the three-dimensional organization of the genome. FEBS Lett. 2015;589: 2923–2930.

103. Kim S, Yu N-K, Kaang B-K. CTCF as a multifunctional protein in genome regulation and gene expression. Exp Mol Med. 2015;47: e166.

104. Gaszner M, Felsenfeld G. Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet. 2006;7: 703–713.

105. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell. 1999;98: 387–396.

106. Recillas-Targa F, Pikaart MJ, Burgess-Beusse B, Bell AC, Litt MD, West AG, et al. Position-effect protection and enhancer blocking by the chicken β-globin insulator are separable activities. Proceedings of the National Academy of Sciences. 2002;99: 6883–6888.

107. Chung JH, Whiteley M, Felsenfeld G. A 5’ element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. Cell. 1993;74: 505–514.

108. Kumar RP, Krishnan J, Pratap Singh N, Singh L, Mishra RK. GATA simple sequence repeats function as enhancer blocker boundaries. Nat Commun. 2013;4: 1844.

109. Hong CKY, Erickson AA, Li J, Federico AJ, Cohen BA. Massively parallel characterization of insulator activity across the genome. bioRxiv. 2022. p. 2022.11.29.518444.

110. Willoughby DA, Vilalta A, Oshima RG. An Alu element from the K18 gene confers position-independent expression in transgenic mice. J Biol Chem. 2000;275: 759–768.

111. Smirnov NA, Didych DA, Akopov SB, Nikolaev LG, Sverdlov ED. Assay of insulator enhancer-blocking activity with the use of transient transfection. Biochemistry . 2013;78: 895–903.

112. Emery DW, Yannaki E, Tubb J, Nishino T, Li Q, Stamatoyannopoulos G. Development of virus vectors for gene therapy of beta chain hemoglobinopathies: flanking with a chromatin insulator reduces gamma-globin gene silencing in vivo. Blood. 2002;100: 2012–2019.

113. Cavalheiro GR, Pollex T, Furlong EEM. To loop or not to loop: what is the role of

TADs in enhancer function and gene regulation? Curr Opin Genet Dev. 2021;67: 119–129.

114. Zhu X, Ling J, Zhang L, Pi W, Wu M, Tuan D. A facilitated tracking and transcription mechanism of long-range enhancer function. Nucleic Acids Res. 2007;35: 5532–5544.

115. Yusufzai TM, Tagami H, Nakatani Y, Felsenfeld G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. Mol Cell. 2004;13: 291–298.

116. Pekowska A, Benoukraf T, Zacarias-Cabeza J, Belhocine M, Koch F, Holota H, et al. H3K4 tri-methylation provides an epigenetic signature of active enhancers. EMBO J. 2011;30: 4198–4210.

117. Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning. Proceedings of the 25th international conference on Machine learning. New York, NY, USA: Association for Computing Machinery; 2008. pp. 160–167.

118. Andersson R, Sandelin A. Determinants of enhancer and promoter activities of regulatory elements. Nat Rev Genet. 2020;21: 71–87.

119. van Arensbergen J, FitzPatrick VD, de Haas M, Pagie L, Sluimer J, Bussemaker HJ, et al. Genome-wide mapping of autonomous promoter activity in human cells. Nat Biotechnol. 2017;35: 145–153.

120. Nguyen TA, Jones RD, Snavely AR, Pfenning AR, Kirchner R, Hemberg M, et al. High-throughput functional comparison of promoter and enhancer activities. Genome Res. 2016;26: 1023–1033.

121. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009;137: 1194–1211.

122. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013;339: 1074–1077.

123. Inoue F, Ahituv N. Decoding enhancers using massively parallel reporter assays. Genomics. 2015;106: 159–164.

124. Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res. 2017;27: 38–52.

125. Kvon EZ, Kazmar T, Stampfel G, Yáñez-Cuna JO, Pagani M, Schernhuber K, et al. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature. 2014;512: 91–95.

126. Navratilova P, Fredman D, Hawkins TA, Turner K, Lenhard B, Becker TS. Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. Dev Biol. 2009;327: 526–540.

127. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337: 816–821.

128. Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, Qi LS. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nat Protoc. 2013;8: 2180–2196.

129. Dong C, Fontana J, Patel A, Carothers JM, Zalatan JG. Synthetic CRISPR-Cas gene activators for transcriptional reprogramming in bacteria. Nat Commun. 2018;9: 2489.

130. Field A, Adelman K. Evaluating Enhancer Function and Transcription. Annu Rev Biochem. 2020;89: 213–234.

131. Atchison ML. Enhancers: mechanisms of action and cell specificity. Annu Rev Cell Biol. 1988;4: 127–153.

132. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503: 290–294.

133. Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. Nature Genetics. 2019. pp. 1442–1449.

134. Barolo S. Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. Bioessays. 2012;34: 135–141.

135. Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, et al. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature. 2018;554: 239–243.

136. Carleton JB, Berrett KC, Gertz J. Multiplex Enhancer Interference Reveals Collaborative Control of Gene Regulation by Estrogen Receptor α-Bound Enhancers. Cell Syst. 2017;5: 333–344.e5.

137. Davis JE, Insigne KD, Jones EM, Hastings QA, Boldridge WC, Kosuri S. Dissection of c-AMP Response Element Architecture by Using Genomic and Episomal Massively Parallel Reporter Assays. Cell Syst. 2020;11: 75–85.e7.

138. Zuin J, Roth G, Zhan Y, Cramard J, Redolfi J, Piskadlo E, et al. Nonlinear control of transcription through enhancer-promoter interactions. Nature. 2022;604: 571–577.

139. Hong J-W, Hendrix DA, Levine MS. Shadow enhancers as a source of evolutionary

novelty. Science. 2008;321: 1314.

140. Butler JEF, Kadonaga JT. Enhancer–promoter specificity mediated by DPE or TATA core promoter motifs. Genes Dev. 2001;15: 2515–2519.

141. van Arensbergen J, van Steensel B, Bussemaker HJ. In search of the determinants of enhancer-promoter interaction specificity. Trends Cell Biol. 2014;24: 695–702.

142. Harth-Hertle ML, Scholz BA, Erhard F, Glaser LV, Dölken L, Zimmer R, et al. Inactivation of intergenic enhancers by EBNA3A initiates and maintains polycomb signatures across a chromatin domain encoding CXCL10 and CXCL9. PLoS Pathog. 2013;9: e1003638.

143. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012. pp. 376–380.

144. Sun F, Chronis C, Kronenberg M, Chen X-F, Su T, Lay FD, et al. Promoter-Enhancer Communication Occurs Primarily within Insulated Neighborhoods. Mol Cell. 2019;73: 250–263.e5.

145. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nat Rev Genet. 2018;19: 789–800.

146. Dily FL, Le Dily F, Baù D, Pohl A, Vicent GP, Serra F, et al. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes & Development. 2014. pp. 2151–2162.

147. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. 2016;15: 2038–2049.

148. Kim Y, Shi Z, Zhang H, Finkelstein IJ, Yu H. Human cohesin compacts DNA by loop extrusion. Science. 2019. pp. 1345–1349.

149. Kojic A, Cuadrado A, De Koninck M, Giménez-Llorente D, Rodríguez-Corsino M, Gómez-López G, et al. Distinct roles of cohesin-SA1 and cohesin-SA2 in 3D chromosome organization. Nat Struct Mol Biol. 2018;25: 496–504.

150. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape at single-cell level. Nature. 2020;581: 303–309.

151. Struhl K. Fundamentally Different Logic of Gene Regulation in Eukaryotes and Prokaryotes. Cell. 1999. pp. 1–4.

152. Prud'homme B, Gompel N, Carroll SB. Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A. 2007;104 Suppl 1: 8605–8612.

153. Johnson WC, Ordway AJ, Watada M, Pruitt JN, Williams TM, Rebeiz M. Genetic Changes to a Transcriptional Silencer Element Confers Phenotypic Diversity within and between Drosophila Species. PLoS Genet. 2015;11: e1005279.

154. Mandel M, Higa A. Calcium-dependent bacteriophage DNA infection. J Mol Biol. 1970;53: 159–162.

155. Cohen SN, Chang AC, Hsu L. Nonchromosomal antibiotic resistance in bacteria: genetic transformation of Escherichia coli by R-factor DNA. Proc Natl Acad Sci U S A. 1972;69: 2110–2114.

156. O'Donnell M, Langston L, Stillman B. Principles and concepts of DNA replication in bacteria, archaea, and eukarya. Cold Spring Harb Perspect Biol. 2013;5.

157. Tuttle AR, Trahan ND, Son MS. Growth and Maintenance of Escherichia coli Laboratory Strains. Curr Protoc. 2021;1: e20.

158. Arber W, Linn S. DNA modification and restriction. Annu Rev Biochem. 1969;38: 467–500.

159. Hartley JL, Temple GF, Brasch MA. DNA cloning using in vitro site-specific recombination. Genome Res. 2000;10: 1788–1795.

160. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA 3rd, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods. 2009;6: 343–345.

161. Martella A, Matjusaitis M, Auxillos J, Pollard SM, Cai Y. EMMA: An Extensible Mammalian Modular Assembly Toolkit for the Rapid Design and Production of Diverse Expression Vectors. ACS Synth Biol. 2017;6: 1380–1392.

162. Fus-Kujawa A, Prus P, Bajdak-Rusinek K, Teper P, Gawron K, Kowalczuk A, et al. An Overview of Methods and Tools for Transfection of Eukaryotic Cells in vitro. Front Bioeng Biotechnol. 2021;9: 701031.

163. Kim TK, Eberwine JH. Mammalian cell transfection: the present and the future. Anal Bioanal Chem. 2010;397: 3173–3178.

164. Makrides SC. Strategies for achieving high-level expression of genes in Escherichia coli. Microbiol Rev. 1996;60: 512–538.

165. Williams PD, Kingston PA. Plasmid-mediated gene therapy for cardiovascular disease. Cardiovasc Res. 2011;91: 565–576.

166. Lara AR, Ramírez OT, Wunderlich M. Plasmid DNA production for therapeutic applications. Methods Mol Biol. 2012;824: 271–303.

167. Becker S, Boch J. TALE and TALEN genome editing technologies. Gene and

Genome Editing. 2021;2: 100007.

168. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. Nat Protoc. 2013;8: 2281–2308.

169. Kvon EZ. Using transgenic reporter assays to functionally characterize enhancers in animals. Genomics. 2015;106: 185–192.

170. Kornberg RD. Structure of chromatin. Annu Rev Biochem. 1977;46: 931–954.

171. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature. 1997;389: 251–260.

172. Kornberg RD, Thomas JO. Chromatin structure; oligomers of the histones. Science. 1974;184: 865–868.

173. Schultz J. Variegation in Drosophila and the Inert Chromosome Regions. Proc Natl Acad Sci U S A. 1936;22: 27–33.

174. Mladenova V, Mladenov E, Russev G. Organization of Plasmid DNA into Nucleosome-Like Structures after Transfection in Eukaryotic Cells. Biotechnol Biotechnol Equip. 2009;23: 1044–1047.

175. Tong W, Kulaeva OI, Clark DJ, Lutter LC. Topological analysis of plasmid chromatin from yeast and mammalian cells. J Mol Biol. 2006;361: 813–822.

176. Zhao H, Kim A, Song S-H, Dean A. Enhancer blocking by chicken beta-globin 5'-HS4: role of enhancer strength and insulator nucleosome depletion. J Biol Chem. 2006;281: 30573–30580.

177. Nakagawa T, Yoneda M, Higashi M, Ohkuma Y, Ito T. Enhancer function regulated by combinations of transcription factors and cofactors. Genes Cells. 2018;23: 808–821.

178. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol. 2015;16: 144–154.

179. Huang D, Ovcharenko I. Enhancer-silencer transitions in the human genome. Genome Res. 2022;32: 437–448.

180. Lee DSM, Park J, Kromer A, Baras A, Rader DJ, Ritchie MD, et al. Disrupting upstream translation in mRNAs is associated with human disease. Nat Commun. 2021;12: 1515.

181. van der Velden AW, Thomas AA. The role of the 5' untranslated region of an mRNA in translation regulation during development. Int J Biochem Cell Biol. 1999;31: 87–106.

182. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic

dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012;30: 271–277.

183. Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat Methods. 2020;17: 1083–1091.

184. Lee D, Kapoor A, Lee C, Mudgett M, Beer MA, Chakravarti A. Sequence-based correction of barcode bias in massively parallel reporter assays. Genome Res. 2021;31: 1638–1645.

185. Muerdter F, Boryń ŁM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, et al. Resolving systematic errors in widely used enhancer activity assays in human cells. Nat Methods. 2018;15: 141–149.

186. White MA, Myers CA, Corbo JC, Cohen BA. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc Natl Acad Sci U S A. 2013;110: 11952–11957.

187. Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, et al. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat Genet. 2013;45: 1021–1028.

188. Serfling E, Jasin M, Schaffner W. Enhancers and eukaryotic gene transcription. Trends Genet. 1985;1: 224–230.

189. Malone CS, Omori SA, Wall R. Silencer elements controlling the B29 (Igβ) promoter are neither promoter- nor cell-type-specific. Proceedings of the National Academy of Sciences. 1997;94: 12314–12319.

190. Zou Y, Yu Q, Chiu Y-H, Bi X. Position effect on the directionality of silencer function in Saccharomyces cerevisiae. Genetics. 2006;174: 203–213.

191. Yokoshi M, Segawa K, Fukaya T. Visualizing the Role of Boundary Elements in Enhancer-Promoter Communication. Mol Cell. 2020;78: 224–235.e5.

192. Tinti C, Yang C, Seo H, Conti B, Kim C, Joh TH, et al. Structure/function relationship of the cAMP response element in tyrosine hydroxylase gene transcription. J Biol Chem. 1997;272: 19158–19164.

193. Liu Y, Bondarenko V, Ninfa A, Studitsky VM. DNA supercoiling allows enhancer action over a large distance. Proc Natl Acad Sci U S A. 2001;98: 14883–14888.

194. Waters CT, Gisselbrecht SS, Sytnikova YA, Cafarelli TM, Hill DE, Bulyk ML. Quantitative-enhancer-FACS-seq (QeFS) reveals epistatic interactions among motifs within transcriptional enhancers in developing Drosophila tissue. Genome Biol. 2021;22: 348.

195. Asakawa K, Kawakami K. The Tol2-mediated Gal4-UAS method for gene and enhancer trapping in zebrafish. Methods. 2009;49: 275–281.

196. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol. 2012;30: 521–530.

197. Calderon D, Ellis A, Daza RM, Martin B, Tome JM, Chen W, et al. TransMPRA: A framework for assaying the role of many trans-acting factors at many enhancers. bioRxiv. 2020. p. 2020.09.30.321323.

198. Cohen RN, van der Aa MAEM, Macaraeg N, Lee AP, Szoka FC Jr. Quantification of plasmid DNA copies in the nucleus after lipoplex and polyplex transfection. J Control Release. 2009;135: 166–174.

199. Morris JR, Chen JL, Geyer PK, Wu CT. Two modes of transvection: enhancer action in trans and bypass of a chromatin insulator in cis. Proc Natl Acad Sci U S A. 1998;95: 10740–10745.

200. Boshart M, Weber F, Jahn G, Dorsch-Häsler K, Fleckenstein B, Schaffner W. A very strong enhancer is located upstream of an immediate early gene of human cytomegalovirus. Cell. 1985;41: 521–530.

201. Ford JP, Hsu MT. Transcription pattern of in vivo-labeled late simian virus 40 RNA: equimolar transcription beyond the mRNA 3’ terminus. Journal of Virology. 1978. pp. 795–801.

202. Kadesch T, Berg P. Effects of the position of the simian virus 40 enhancer on expression of multiple transcription units in a single plasmid. Mol Cell Biol. 1986;6: 2593–2601.

203. Tokuda N, Sasai M, Chikenji G. Roles of DNA looping in enhancer-blocking activity. Biophys J. 2011;100: 126–134.

204. Musteanu M, Blaas L, Zenz R, Svinka J, Hoffmann T, Grabner B, et al. A mouse model to identify cooperating signaling pathways in cancer. Nat Methods. 2012;9: 897–900.

205. Curtin JA, Dane AP, Swanson A, Alexander IE, Ginn SL. Bidirectional promoter interference between two widely used internal heterologous promoters in a late-generation lentiviral construct. Gene Ther. 2008;15: 384–390.

206. Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. Cell. 2013;154: 914–927.

207. Karpen GH. Position-effect variegation and the new biology of heterochromatin. Curr Opin Genet Dev. 1994;4: 281–291.

208. Murtha M, Tokcaer-Keskin Z, Tang Z, Strino F, Chen X, Wang Y, et al. FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. Nat Methods. 2014;11: 559–565.

209. Browning DL, Trobridge GD. Insulators to Improve the Safety of Retroviral Vectors for HIV Gene Therapy. Biomedicines. 2016;4.

210. Ameres SL, Drueppel L, Pfleiderer K, Schmidt A, Hillen W, Berens C. Inducible DNA-loop formation blocks transcriptional activation by an SV40 enhancer. The EMBO Journal. 2005. pp. 358–367.

211. Parnell TJ, Geyer PK. Differences in insulator properties revealed by enhancer blocking assays on episomes. EMBO J. 2000;19: 5864–5874.

212. Chung JH, Bell AC, Felsenfeld G. Characterization of the chicken β-globin insulator. Proceedings of the National Academy of Sciences. 1997;94: 575–580.

213. Lehner R, Wang X, Hunziker P. Plasmid linearization changes shape and efficiency of transfection complexes. European Journal of Nanomedicine. 2013;5: 205–212.

214. Tuan DY, Solomon WB, London IM, Lee DP. An erythroid-specific, developmental-stage-independent enhancer far upstream of the human “beta-like globin” genes. Proc Natl Acad Sci U S A. 1989;86: 2554–2558.

215. Scott KC, Taubman AD, Geyer PK. Enhancer blocking by the Drosophila gypsy insulator depends upon insulator anatomy and enhancer strength. Genetics. 1999;153: 787–798.

216. Kyrchanova O, Chetverina D, Maksimenko O, Kullyev A, Georgiev P. Orientation-dependent interaction between Drosophila insulators is a property of this class of regulatory elements. Nucleic Acids Res. 2008;36: 7019–7028.

217. Yannaki E, Tubb J, Aker M, Stamatoyannopoulos G, Emery DW. Topological constraints governing the use of the chicken HS4 chromatin insulator in oncoretrovirus vectors. Mol Ther. 2002;5: 589–598.

218. Zhao H, Dean A. An insulator blocks spreading of histone acetylation and interferes with RNA polymerase II transfer between an enhancer and gene. Nucleic Acids Res. 2004;32: 4903–4919.

219. de Wit E, Vos ESM, Holwerda SJB, Valdes-Quezada C, Verstegen MJAM, Teunissen H, et al. CTCF Binding Polarity Determines Chromatin Looping. Mol Cell. 2015;60: 676–684.

220. Huang H, Zhu Q, Jussila A, Han Y, Bintu B, Kern C, et al. CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. Nat Genet. 2021;53: 1064–1074.

221. Ribeiro-Dos-Santos AM, Hogan MS, Luther RD, Brosh R, Maurano MT. Genomic context sensitivity of insulator function. Genome Res. 2022;32: 425–436.

222. Ivashkiv LB, Donlin LT. Regulation of type I interferon responses. Nat Rev Immunol. 2014;14: 36–49.

223. Di Blasi R, Marbiah MM, Siciliano V, Polizzi K, Ceroni F. A call for caution in analysing mammalian co-transfection experiments and implications of resource competition in data misinterpretation. Nat Commun. 2021;12: 2545.

224. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921.

225. Rosano GL, Ceccarelli EA. Recombinant protein expression in Escherichia coli: advances and challenges. Front Microbiol. 2014;5: 172.

226. Mali S. Delivery systems for gene therapy. Indian J Hum Genet. 2013;19: 3–8.

227. Potapov V, Ong JL. Examining Sources of Error in PCR by Single-Molecule Sequencing. PLoS One. 2017;12: e0169774.

228. Conley EC, Saunders VA, Saunders JR. Deletion and rearrangement of plasmid DNA during transformation of Escherichia coli with linear plasmid molecules. Nucleic Acids Res. 1986;14: 8905–8917.

229. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74: 5463–5467.

230. Shinde D, Lai Y, Sun F, Arnheim N. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. Nucleic Acids Res. 2003;31: 974–980.

231. Stranneheim H, Lundeberg J. Stepping stones in DNA sequencing. Biotechnol J. 2012;7: 1063–1073.

232. Kittleson JT, Wu GC, Anderson JC. Successes and failures in modular genetic engineering. Curr Opin Chem Biol. 2012;16: 329–336.

233. Williams JA, Carnes AE, Hodgson CP. Plasmid DNA vaccine vector design: impact on efficacy, safety and upstream production. Biotechnol Adv. 2009;27: 353–370.

234. Gallegos JE, Rogers MF, Cialek CA, Peccoud J. Rapid, robust plasmid verification by de novo assembly of short sequencing reads. Nucleic Acids Res. 2020;48: e106.

235. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012: 251364.

236. McDonald TL, Zhou W, Castro CP, Mumm C, Switzenberg JA, Mills RE, et al.

Cas9 targeted enrichment of mobile elements using nanopore sequencing. Nat Commun. 2021;12: 3586.

237. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol. 2020;38: 433–438.

238. Sanchis-Juan A, Stephens J, French CE, Gleadall N, Mégy K, Penkett C, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Med. 2018;10: 95.

239. Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, et al. Sequencing of human genomes with nanopore technology. Nat Commun. 2019;10: 1869.

240. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39: 1348–1365.

241. Emiliani FE, Hsu I, McKenna A. Circuit-seq: Circular reconstruction of cut in vitro transposed plasmids using Nanopore sequencing. bioRxiv. 2022. p. 2022.01.25.477550.

242. Currin A, Swainston N, Dunstan MS, Jervis AJ, Mulherin P, Robinson CJ, et al. Highly multiplexed, fast and accurate nanopore sequencing for verification of synthetic DNA constructs and sequence libraries. Synth Biol. 2019;4: ysz025.

243. Brown SD, Dreolini L, Wilson JF, Balasundaram M, Holt RA. Complete sequence verification of plasmid DNA using the Oxford Nanopore Technologies' MinION device. bioRxiv. 2022. p. 2022.06.21.497051.

244. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16: 276–277.

245. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14: 178–192.

246. Yang C, Chu J, Warren RL, Birol I. NanoSim: nanopore sequence read simulator based on statistical characterization. Gigascience. 2017;6: 1–6.

247. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19: 90.

248. Chandak S, Neu J, Tatwawadi K, Mardia J, Lau B, Kubit M, et al. Overcoming High Nanopore Basecaller Error Rates for DNA Storage Via Basecaller-Decoder Integration and Convolutional Codes. bioRxiv. 2020. p. 2019.12.20.871939.

249. Kieleczawa J. Fundamentals of sequencing of difficult templates--an overview. J Biomol Tech. 2006;17: 207–217.

250. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022;19: 823–826.

251. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell. 2014;159: 647–661.

252. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34: 2666–2669.

253. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. 2013;23: 800–811.

254. Klein JC, Keith A, Agarwal V, Durham T, Shendure J. Functional characterization of enhancer evolution in the primate lineage. Genome Biol. 2018;19: 99.

255. Mattioli K, Volders P-J, Gerhardinger C, Lee JC, Maass PG, Melé M, et al. High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. Genome Res. 2019;29: 344–355.

256. Kim J, Kang J, Kim YW, Kim A. The human β-globin enhancer LCR HS2 plays a role in forming a TAD by activating chromatin structure at neighboring CTCF sites. FASEB J. 2021;35: e21669.

257. Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature. 2000;405: 486–489.

258. Symmons O, Pan L, Remeseiro S, Aktas T, Klein F, Huber W, et al. The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. Dev Cell. 2016;39: 529–543.

259. Xu Z, Wei G, Chepelev I, Zhao K, Felsenfeld G. Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. Nat Struct Mol Biol. 2011;18: 372–378.

260. Li L, Waymack R, Gad M, Wunderlich Z. Two promoters integrate multiple enhancer inputs to drive wild-type knirps expression in the Drosophila melanogaster embryo. Genetics. 2021;219.

261. Kvon EZ, Waymack R, Gad M, Wunderlich Z. Enhancer redundancy in

development and disease. Nat Rev Genet. 2021;22: 324–336.

262. Choi J, Lysakovskaia K, Stik G, Demel C, Söding J, Tian TV, et al. Evidence for additive and synergistic action of mammalian enhancers during cell fate determination. Elife. 2021;10.

263. Bothma JP, Garcia HG, Ng S, Perry MW, Gregor T, Levine M. Enhancer additivity and non-additivity are determined by enhancer strength in the Drosophila embryo. Elife. 2015;4.

264. Naumova N, Smith EM, Zhan Y, Dekker J. Analysis of long-range chromatin interactions using Chromosome Conformation Capture. Methods. 2012. pp. 192–203.

265. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159: 1665–1680.

266. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485: 381–385.

267. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148: 458–472.

268. Sikorska N, Sexton T. Defining Functionally Relevant Spatial Chromatin Domains: It is a TAD Complicated. J Mol Biol. 2020;432: 653–664.

269. Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science. 2015;347: 1010–1014.

270. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, et al. Functional and topological characteristics of mammalian regulatory domains. Genome Research. 2014. pp. 390–400.

271. Narendra V, Rocha PP, An D, Raviram R, Skok JA, Mazzoni EO, et al. CTCF establishes discrete functional chromatin domains at the *Hox* clusters during differentiation. Science. 2015. pp. 1017–1021.

272. Hanssen LLP, Kassouf MT, Oudelaar AM, Biggs D, Preece C, Downes DJ, et al. Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. Nat Cell Biol. 2017;19: 952–961.

273. Diehl AG, Boyle AP. Conserved and species-specific transcription factor co-binding patterns drive divergent gene regulation in human and mouse. Nucleic Acids Res. 2018;46: 1878–1894.

274. Kulkarni MM, Arnosti DN. Information display by transcriptional enhancers. Development. 2003;130: 6569–6575.

275. Junion G, Spivakov M, Girardot C, Braun M, Hilary Gustafson E, Birney E, et al. A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. Cell. 2012. pp. 473–486.

276. Andersson LC, Nilsson K, Gahmberg CG. K562—A human erythroleukemic cell line. International Journal of Cancer. 1979;23: 143–147.

277. Donato MT, Tolosa L, Gómez-Lechón MJ. Culture and Functional Characterization of Human Hepatoma HepG2 Cells. Methods Mol Biol. 2015;1250: 77–93.

278. Aden DP, Fogel A, Plotkin S, Damjanov I, Knowles BB. Controlled synthesis of HBsAg in a differentiated human liver carcinoma-derived cell line. Nature. 1979;282: 615–616.

279. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017;12: 2478–2492.

280. Huang S. Blimp-1 is the murine homolog of the human transcriptional repressor PRDI-BF1. Cell. 1994;78: 9.

281. Keller AD, Maniatis T. Identification and characterization of a novel repressor of beta-interferon gene expression. Genes Dev. 1991;5: 868–879.

282. Yu J, Angelin-Duclos C, Greenwood J, Liao J, Calame K. Transcriptional repression by blimp-1 (PRDI-BF1) involves recruitment of histone deacetylase. Mol Cell Biol. 2000;20: 2592–2603.

283. Vervoort M, Meulemeester D, Béhague J, Kerner P. Evolution of Prdm Genes in Animals: Insights from Comparative Genomics. Mol Biol Evol. 2016;33: 679–696.

284. Bikoff EK, Morgan MA, Robertson EJ. An expanding job description for Blimp-1/PRDM1. Curr Opin Genet Dev. 2009;19: 379–385.

285. Vincent SD, Mayeuf-Louchart A, Watanabe Y, Brzezinski JA 4th, Miyagawa-Tomita S, Kelly RG, et al. Prdm1 functions in the mesoderm of the second heart field, where it interacts genetically with Tbx1, during outflow tract morphogenesis in the mouse embryo. Hum Mol Genet. 2014;23: 5087–5101.

286. Horsley V, O'Carroll D, Tooze R, Ohinata Y, Saitou M, Obukhanych T, et al. Blimp1 defines a progenitor population that governs cellular input to the sebaceous gland. Cell. 2006;126: 597–609.

287. Ohinata Y, Payer B, O'Carroll D, Ancelin K, Ono Y, Sano M, et al. Blimp1 is a critical determinant of the germ cell lineage in mice. Nature. 2005. pp. 207–213.

288. Magnúsdóttir E, Dietmann S, Murakami K, Günesdogan U, Tang F, Bao S, et al. A tripartite transcription factor network regulates primordial germ cell specification in mice. Nat Cell Biol. 2013;15: 905–915.

289. Roy S, Ng T. Blimp-1 specifies neural crest and sensory neuron progenitors in the zebrafish embryo. Curr Biol. 2004;14: 1772–1777.

290. Smith MA, Maurin M, Cho HI, Becknell B, Freud AG, Yu J, et al. PRDM1/Blimp-1 controls effector cytokine production in human NK cells. J Immunol. 2010;185: 6058–6067.

291. Kallies A, Carotta S, Huntington ND, Bernard NJ, Tarlinton DM, Smyth MJ, et al. A role for Blimp1 in the transcriptional network controlling natural killer cell maturation. Blood. 2011;117: 1869–1879.

292. Chang DH, Angelin-Duclos C, Calame K. BLIMP-1: trigger for differentiation of myeloid lineage. Nat Immunol. 2000;1: 169–176.

293. Johnston RJ, Poholek AC, DiToro D, Yusuf I, Eto D, Barnett B, et al. Bcl6 and Blimp-1 are reciprocal and antagonistic regulators of T follicular helper cell differentiation. Science. 2009;325: 1006–1010.

294. Angelin-Duclos C, Cattoretti G, Lin KI, Calame K. Commitment of B lymphocytes to a plasma cell fate is associated with Blimp-1 expression in vivo. J Immunol. 2000;165: 5462–5471.

295. Kallies A, Hawkins ED, Belz GT, Metcalf D, Hommel M, Corcoran LM, et al. Transcriptional repressor Blimp-1 is essential for T cell homeostasis and self-tolerance. Nature Immunology. 2006. pp. 466–474.

296. Mackay LK, Minnich M, Kragten NAM, Liao Y, Nota B, Seillet C, et al. Hobit and Blimp1 instruct a universal transcriptional program of tissue residency in lymphocytes. Science. 2016;352: 459–463.

297. Yan J, Jiang J, Lim CA, Wu Q, Ng H-H, Chin K-C. BLIMP1 regulates cell growth through repression of p53 transcription. Proc Natl Acad Sci U S A. 2007;104: 1841–1846.

298. Zhu Z, Wang H, Wei Y, Meng F, Liu Z, Zhang Z. Downregulation of PRDM1 promotes cellular invasion and lung cancer metastasis. Tumour Biol. 2017;39: 1010428317695929.

299. Xia Y, Xu-Monette ZY, Tzankov A, Li X, Manyam GC, Murty V, et al. Loss of PRDM1/BLIMP-1 function contributes to poor prognosis of activated B-cell-like diffuse large B-cell lymphoma. Leukemia. 2017;31: 625–636.

300. Boi M, Zucca E, Inghirami G, Bertoni F. PRDM1/BLIMP1: a tumor suppressor gene in B and T cell lymphomas. Leuk Lymphoma. 2015;56: 1223–1228.

301. Chiou S-H, Risca VI, Wang GX, Yang D, Grüner BM, Kathiria AS, et al. BLIMP1 Induces Transient Metastatic Heterogeneity in Pancreatic CancerBLIMP1 Induces Metastasis in Pancreatic Cancer. Cancer Discov. 2017;7: 1184–1199.

302. Li X, He S, Ma B. Autophagy and autophagy-related proteins in cancer. Mol Cancer. 2020;19: 12.

303. Tanida I. Autophagosome formation and molecular mechanism of autophagy. Antioxid Redox Signal. 2011;14: 2201–2214.

304. Grand RJ, Milner AE, Mustoe T, Johnson GD, Owen D, Grant ML, et al. A novel protein expressed in mammalian cells undergoing apoptosis. Exp Cell Res. 1995;218: 439–451.

305. Hammond EM, Brunet CL, Johnson GD, Parkhill J, Milner AE, Brady G, et al. Homology between a human apoptosis specific protein and the product of APG5, a gene involved in autophagy in yeast. FEBS Lett. 1998;425: 391–395.

306. Mizushima N, Yamamoto A, Hatano M, Kobayashi Y, Kabeya Y, Suzuki K, et al. Dissection of autophagosome formation using Apg5-deficient mouse embryonic stem cells. J Cell Biol. 2001;152: 657–668.

307. Georgi B, Voight BF, Bućan M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS Genet. 2013;9: e1003484.

308. Kuma A, Hatano M, Matsui M, Yamamoto A, Nakaya H, Yoshimori T, et al. The role of autophagy during the early neonatal starvation period. Nature. 2004;432: 1032–1036.

309. Tsukamoto S, Kuma A, Murakami M, Kishi C, Yamamoto A, Mizushima N. Autophagy Is Essential for Preimplantation Development of Mouse Embryos. Science. 2008. pp. 117–120.

310. Kim M, Sandford E, Gatica D, Qiu Y, Liu X, Zheng Y, et al. Author response: Mutation in ATG5 reduces autophagy and leads to ataxia with developmental delay. eLife Sciences Publications, Ltd; 2016.

311. Hara T, Nakamura K, Matsui M, Yamamoto A, Nakahara Y, Suzuki-Migishima R, et al. Suppression of basal autophagy in neural cells causes neurodegenerative disease in mice. Nature. 2006;441: 885–889.

312. Codogno P, Meijer AJ. Atg5: more than an autophagy factor. Nature cell biology. 2006. pp. 1045–1047.

313. Singh R, Kaushik S, Wang Y, Xiang Y, Novak I, Komatsu M, et al. Autophagy regulates lipid metabolism. Nature. 2009;458: 1131–1135.

314. Miller BC, Zhao Z, Stephenson LM, Cadwell K, Pua HH, Lee HK, et al. The

autophagy gene ATG5 plays an essential role in B lymphocyte development. Autophagy. 2008;4: 309–314.

315. Pua HH, He Y-W. Maintaining T lymphocyte homeostasis: another duty of autophagy. Autophagy. 2007;3: 266–267.

316. Lee HK, Lund JM, Ramanathan B, Mizushima N, Iwasaki A. Autophagy-dependent viral recognition by plasmacytoid dendritic cells. Science. 2007;315: 1398–1401.

317. Nakagawa I, Amano A, Mizushima N, Yamamoto A, Yamaguchi H, Kamimoto T, et al. Autophagy Defends Cells Against Invading Group A *Streptococcus*. Science. 2004. pp. 1037–1040.

318. Kondo Y, Kanzawa T, Sawaya R, Kondo S. The role of autophagy in cancer development and response to therapy. Nat Rev Cancer. 2005;5: 726–734.

319. Yousefi S, Simon H-U. Apoptosis regulation by autophagy gene 5. Crit Rev Oncol Hematol. 2007;63: 241–244.

320. Cho D-H, Jo YK, Kim SC, Park IJ, Kim JC. Down-regulated expression of ATG5 in colorectal cancer. Anticancer Res. 2012;32: 4091–4096.

321. Drullion C, Trégoat C, Lagarde V, Tan S, Gioia R, Priault M, et al. Apoptosis and autophagy have opposite roles on imatinib-induced K562 leukemia cell senescence. Cell Death Dis. 2012;3: e373.

322. Tong Y, Liu Y-Y, You L-S, Qian W-B. Perifosine induces protective autophagy and upregulation of ATG5 in human chronic myelogenous leukemia cells in vitro. Acta Pharmacol Sin. 2012;33: 542–550.

323. Zhang K, Chen J, Zhou H, Chen Y, Zhi Y, Zhang B, et al. PU.1/microRNA-142-3p targets ATG5/ATG16L1 to inactivate autophagy and sensitize hepatocellular carcinoma cells to sorafenib. Cell Death Dis. 2018;9: 312.

324. Gateva V, Sandling JK, Hom G, Taylor KE, Chung SA, Sun X, et al. A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. Nat Genet. 2009;41: 1228–1233.

325. Han J-W, Zheng H-F, Cui Y, Sun L-D, Ye D-Q, Hu Z, et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat Genet. 2009;41: 1234–1237.

326. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132: 311–322.

327. Byrne BJ, Davis MS, Yamaguchi J, Bergsma DJ, Subramanian KN. Definition of the simian virus 40 early promoter region and demonstration of a host range bias in the

enhancement effect of the simian virus 40 72-base-pair repeat. Proc Natl Acad Sci U S A. 1983;80: 721–725.

328. Raisner R, Kharbanda S, Jin L, Jeng E, Chan E, Merchant M, et al. Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation. Cell Rep. 2018;24: 1722–1729.

329. Kallies A, Hasbold J, Tarlinton DM, Dietrich W, Corcoran LM, Hodgkin PD, et al. Plasma cell ontogeny defined by quantitative changes in blimp-1 expression. J Exp Med. 2004;200: 967–977.

330. Bönelt P, Wöhner M, Minnich M, Tagoh H, Fischer M, Jaritz M, et al. Precocious expression of Blimp1 in B cells causes autoimmune disease with increased self-reactive plasma cells. EMBO J. 2019;38: 1–19.

331. Jackson JD, Miller W, Hardison RC. Sequences within and flanking hypersensitive sites 3 and 2 of the beta-globin locus control region required for synergistic versus additive interaction with the epsilon-globin gene promoter. Nucleic Acids Res. 1996;24: 4327–4335.

332. Ouyang N, Boyle AP. TRACE: transcription factor footprinting using chromatin accessibility data and DNA sequence. Genome Res. 2020;30: 1040–1046.

333. Talbot D, Grosveld F. The 5'HS2 of the globin locus control region enhances transcription through the interaction of a multimeric complex binding at two functionally distinct NF-E2 binding sites. EMBO J. 1991;10: 1391–1398.

334. Kioussis D, Vanin E, deLange T, Flavell RA, Grosveld FG. β-Globin gene inactivation by DNA translocation in γβ-thalassaemi. Nature. 1983;306: 662–666.

335. Taub R, Kirsch I, Morton C, Lenoir G, Swan D, Tronick S, et al. Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells. Proc Natl Acad Sci U S A. 1982;79: 7837–7841.

336. Lettice LA, Horikoshi T, Heaney SJH, van Baren MJ, van der Linde HC, Breedveld GJ, et al. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proceedings of the National Academy of Sciences. 2002. pp. 7548–7553.

337. Ahituv N. Exonic enhancers: proceed with caution in exome and genome sequencing studies. Genome Med. 2016;8: 14.

338. White RJ. RNA Polymerase I and RNA Polymerase III in Eukaryotes. In: Lennarz WJ, Lane MD, editors. Encyclopedia of Biological Chemistry. New York: Elsevier; 2004. pp. 759–762.

339. Hsieh T-HS, Cattoglio C, Slobodyanyuk E, Hansen AS, Rando OJ, Tjian R, et al. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Mol Cell. 2020;78: 539–553.e8.