Winter 12-15-2022

Understanding gene regulation with high-throughput genome-integrated reporter assays

Clarice Hong
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

Part of the Genetics Commons

Recommended Citation
Hong, Clarice, "Understanding gene regulation with high-throughput genome-integrated reporter assays" (2022). Arts & Sciences Electronic Theses and Dissertations. 2798.
https://openscholarship.wustl.edu/art_sci_etds/2798

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
Understanding Gene Regulation with High-throughput Genome-integrated Reporter Assays
by
Clarice Kit Yee Hong

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

December 2022
St. Louis, Missouri
# Table of Contents

List of Figures ........................................................................................................ v  
List of Tables .......................................................................................................... vii  
Acknowledgements ................................................................................................. viii  
Abstract ................................................................................................................. x  
Chapter 1: Introduction .......................................................................................... 1  

1.1 Types of cis-regulatory sequences ................................................................. 1  
  1.1.1 The surprising diversity of core promoters .............................................. 2  
  1.1.2 The importance of enhancers (and silencers) ....................................... 4  
  1.1.3 The perplexing case of insulators ....................................................... 5  
1.2 Chromatin environments and gene expression ............................................... 8  
1.3 The eternal question of specificity ................................................................ 9  
  1.3.1 Diverse core promoter motifs can encode specificity ......................... 9  
  1.3.2 Insulators as mediators of specificity ................................................. 13  
1.4 Methods for understanding interaction between CRSs and specificity .......... 20  
  1.4.1 Leveraging position effects ................................................................ 21  
  1.4.2 Genome-integrated massively parallel reporter assays ....................... 22  
1.5 Noise in gene expression ............................................................................. 23  
  1.5.1 Sources of gene expression noise ....................................................... 24  
  1.5.2 Methods for studying gene expression noise ...................................... 25  
1.6 Conclusion ................................................................................................ 26  

Chapter 2: Genomic Environments Scale the Activities of Diverse Core Promoters .... 28  

2.1 Abstract .................................................................................................... 28  
2.2 Introduction .............................................................................................. 29  
2.3 Results ...................................................................................................... 30  
  2.3.1 Measurement of diverse core promoter activities at different genomic locations ......................................................................................... 30  
  2.3.2 The effects of genomic locations on core promoters ......................... 32
# List of Figures

Table of Figures:

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Measurements of core promoter library by patchMPRA</td>
<td>31</td>
</tr>
<tr>
<td>2.2</td>
<td>Effects of genomic locations on core promoter activity</td>
<td>33</td>
</tr>
<tr>
<td>2.3</td>
<td>Intrinsic promoter strength explains differences between promoter classes</td>
<td>36</td>
</tr>
<tr>
<td>2.4</td>
<td>Core promoter scaling is a genome-wide phenomenon</td>
<td>39</td>
</tr>
<tr>
<td>2.5</td>
<td>Non-linear scaling of core promoters by genomic environments</td>
<td>41</td>
</tr>
<tr>
<td>2.6</td>
<td>Genomic clusters have different chromatin states and sequence features</td>
<td>43</td>
</tr>
<tr>
<td>S2.1</td>
<td>Landing pad locations overview</td>
<td>54</td>
</tr>
<tr>
<td>S2.2</td>
<td>patchMPRA measurements are robust and reproducible</td>
<td>55</td>
</tr>
<tr>
<td>S2.3</td>
<td>Effect of genomic locations on core promoters</td>
<td>56</td>
</tr>
<tr>
<td>S2.4</td>
<td>Core promoter scaling is not dependent on the proximal enhancer</td>
<td>57</td>
</tr>
<tr>
<td>S2.5</td>
<td>Intrinsic promoter strength explains differences between core promoters</td>
<td>58</td>
</tr>
<tr>
<td>S2.6</td>
<td>Specific interactions between landing pads and promoters are rare</td>
<td>59</td>
</tr>
<tr>
<td>S2.7</td>
<td>Core promoter activities in the genome reflect the promoters' intrinsic activity</td>
<td>60</td>
</tr>
<tr>
<td>S2.8</td>
<td>Measurements of six core promoters by TRIP</td>
<td>61</td>
</tr>
<tr>
<td>S2.9</td>
<td>Core promoter scaling is a genome-wide phenomenon</td>
<td>62</td>
</tr>
<tr>
<td>S2.10</td>
<td>Promoter strength determines its interaction with the genomic environment</td>
<td>64</td>
</tr>
<tr>
<td>S2.11</td>
<td>Epigenomic signatures and sequence features of genomic clusters</td>
<td>65</td>
</tr>
<tr>
<td>3.1</td>
<td>MPIRE is an efficient method for measuring CRS activity genome-wide</td>
<td>66</td>
</tr>
<tr>
<td>3.2</td>
<td>Insulators do not have global effects on gene expression</td>
<td>67</td>
</tr>
<tr>
<td>3.3</td>
<td>Insulators function in a small number of distinct locations in the genome</td>
<td>68</td>
</tr>
<tr>
<td>3.4</td>
<td>Insulators block specific enhancers in the genome</td>
<td>69</td>
</tr>
<tr>
<td>3.5</td>
<td>ALOXE3 acts as a heterochromatin barrier, while A2 and cHS4 act as enhancers at “primed” locations.</td>
<td>70</td>
</tr>
<tr>
<td>S3.1</td>
<td>Overview of cell pools for MPIRE</td>
<td>71</td>
</tr>
<tr>
<td>S3.2</td>
<td>MPIRE reliably measures insulator activity</td>
<td>72</td>
</tr>
<tr>
<td>S3.3</td>
<td>Insulators function at distinct locations in the genome</td>
<td>73</td>
</tr>
</tbody>
</table>
Figure S3.4: Insulators are mostly dependent on their respective sequence motifs. ........109
Figure S3.5: Insulators block specific enhancers in the genome. ...........................110
Figure S3.6: ALOXE3 acts as a barrier against heterochromatin. .........................111
Figure S3.7: Enformer predicts ALOXE3 barrier activity and loss of H3K27me3. ........112
Figure S3.8: cHS4 and A2 act as enhancers at “primed” genomic locations.............113
Figure 4.1: Overview of the SARGENT workflow. ..............................................122
Figure 4.2: SARGENT measurements are accurate and reproducible..................124
Figure 4.3: Chromosomal features associated with expression mean and noise.......126
Figure 4.4: SARGENT quantifies the extrinsic portion of expression noise ..........129
Figure 4.5: Cellular information improves classification of MIN IR locations........131
Figure 4.6: SARGENT measures the insertion effect of a transgene..................134
Figure S4.1: smFISH corroborates with SARGENT measurements. .................148
Figure S4.2: Measurements of mean-independent noise across different chromosomal environments. ..........................................................149
Figure S4.3: Model results for high mean vs low mean.................................150
Figure S4.4: Clonal identification allows for separation of intrinsic and extrinsic noise...151
Figure S4.5: Cell cycle and CD24 states partially explain extrinsic noise.............152
Figure S4.6: IR integrations have little impact on endogenous expression...........153
List of Tables

Table S2.1: Composition of promoter classes in the core promoter library. 68
Table S2.2: Locations of four landing pads in patchMPRA. 68
Table S2.3: Promoters selected for TRIP. 69
Table S2.4: Primers used in this study. 70
Table S2.5: Sources of epigenome datasets used in this study. 74
Table S3.1: Primers used in this study. 114
Table S4.1: Primers used in this study. 154
Acknowledgments

This thesis would not have been possible without the help of so many people. I will be eternally grateful to my advisor, Barak Cohen, for letting me be a part of his lab. From him I have received scientific training, support, encouragement and praise even when I didn’t deserve it. I learnt how to be confident, take risks and ask the right questions in my work. I have grown so much as a scientist but also as a person under his mentorship, and I truly enjoyed my entire PhD experience because of Barak.

I am thankful for everyone I met in the Cohen lab as well. Mike White for being my first mentor in the lab and providing invaluable insights on every topic imaginable. Hemangi Chaudhari, Dana King and Max Staller for fostering an inclusive and supportive lab environment for a young graduate student finding her feet. Jeff Hansen for being there through all the failed experiments and for the entertainment throughout. Siqi Zhao and Avinash Ramu for being wonderful collaborators and teaching me statistics and quantitative thinking. Ryan Friedman for being my gossip partner and for always being willing to help. Kai Loell for insightful comments on our work and James Shepherdson for making sure our lab stays organized. Dustin Baldridge for asking the hard questions and making me better scientist. Yawei Wu for being an amazing partner to explore the world of in vivo single-cell experiments with, and Jie Li for experimental wizardry while I was rushing around towards the end. The intellectual curiosity and scientific passion of the lab as a whole has shaped me to be the scientist I am today.

I thank my thesis committee – Robi Mitra, Ting Wang, Cristina Strong and Tychele Turner – for invaluable insights and feedback on my work. I always enjoy my thesis meetings. I am also grateful to Jim Skeath, Tim Schedl, John Edwards and Melanie Relich for administrative help and keeping me on track.
The Center for Genome Sciences and and Systems Biology (CGSSB) fostered an amazing environment to do my science in. In particular, Jess Hoisington-Lopez and MariaLynn Crosby of the DNA Sequencing Innovation Laboratory (DSIL), whose enthusiasm and willingness to help troubleshoot my terrible sequencing libraries made my work easier than it should have been. Shirley McKinney for keeping all of our equipment clean and functioning.

I am indebted to the McDonnell International Scholars Academy for financial support and for exposing me to people outside my program at WashU.

I am not an extroverted person, but my friends in St. Louis outside the lab helped keep me going. The Taiwanese community welcomed me to their events and gave me food. My friends from Singapore kept me entertained even from afar. The BandTogether community gave me a place to relax and make music together.

I am so thankful to have my family, who have always been so supportive and encouraged me to pursue my dreams even while they have no idea what I am up to. My parents’ trust in my choices gave me the confidence to live my life. I am also grateful to my husband’s family for welcoming me into their lives.

Finally, none of this would have been possible without the companionship of my husband, Wang Liang-Bo. He is the kindest, most considerate person I know and has tolerated all of my bitching about work and life without complaint. His ridiculous belief in my abilities has given me more courage than I knew I had. Also, most everything I know about programming I learnt from him. Thank you for being my friend, partner, tech support, moral compass and so much more.

Clarice Kit Yee Hong

Washington University in St. Louis

December 2022
Proper spatiotemporal gene expression depends on specific interactions between \textit{cis}-regulatory sequences (CRSs) and chromatin modifications at a genomic location. However, we do not yet understand the rules that determine how CRSs work together to influence gene expression. The goal of this thesis is to systematically assay and quantify how different genomic environments interact with CRSs to control gene expression output using a variety of genome-integrated massively parallel reporter assays (MPRAs). First, we tested the hypothesis that core promoters confer specificity by only responding to compatible enhancers and chromatin in specific genomic environments. We measured the activities of hundreds of core promoters at four genomic locations and conversely, the activities of six core promoters at thousands of genomic locations. We found that core promoters are modular components of the genome that are independently scaled up or down by diverse genomic environments. From these results, we proposed a model where genomic environments scale core promoter activities in a non-linear manner that depends on the respective strengths of the location and core promoter. Since core promoters do not appear to contribute to specific interactions in the genome, we next asked if insulators might act in a more context-specific manner. Insulators are CRSs that can block enhancers from activating target genes or act as barriers to prevent heterochromatin spreading. To test the modularity of insulators we developed MPIRE (\textit{Massively Parallel Integrated Regulatory}
Elements), a method that allows us to test the activities of multiple reporter genes in the same thousands of locations across the genome in parallel. Using MPIRE, we measured the activities of three insulator sequences and their mutants across thousands of genomic locations. We found that in contrast to core promoters, insulators act in a highly context-dependent manner and only function in a small number of locations across the genome. While all insulators tested demonstrated enhancer-blocking ability, only one insulator was found to have barrier activity, and insulator activity is dependent on their respective sequence motifs. The results of these two sections suggest a model where genomic environments and core promoters are broadly compatible, while insulators organize the genome to modulate specific CRS interactions. In the final section of the thesis, we asked how genomic environments control cell-to-cell variability in gene expression, or gene expression noise. To answer this question we developed SARGENT (Single-cell Analysis of Reporter Gene Expression Noise and Transcriptome), a single-cell method to quantify the variability of hundreds of reporter genes integrated across the genome in parallel. Using SARGENT, we quantified the contributions of genomic environments and cellular components to gene expression noise and identified genomic and cellular features that might regulate expression noise. Collectively, the results and methods in this thesis will contribute to our understanding of how genomic environments regulate various aspects of gene expression.
Chapter 1: Introduction

The same DNA sequence underlies all cell types in a multicellular organism. Distinct cell types are produced from the same sequence by varying the expression levels of different genes through developmental time and space. These differences in gene expression are driven by the collective action of non-coding regulatory elements, or cis-regulatory sequences (CRSs), in the genome. Understanding how these CRSs work together to influence gene expression is important for understanding spatiotemporal gene expression and cell-type specification in development.

1.1 Types of cis-regulatory sequences

CRSs are short stretches of DNA sequence that can influence gene expression in some way. The concept of a CRS is rather arbitrary and usually reflects the nature of the assay rather than true genomic activity. Nonetheless, CRSs can be broadly split into three categories:

1. Promoters: Transcription initiation site
2. Enhancers/Silencers: Sequences that directly increase or decrease expression respectively
3. Insulators: Sequences that indirectly modulate expression levels through various mechanisms

Due to the arbitrary nature of CRS definitions, these categories are not mutually exclusive and that any given CRS may have more than one function. I will first review the features of each CRS in detail, then discuss our current understanding how CRSs interact with each other to coordinate gene expression.
1.1.1 The surprising diversity of core promoters

The promoter is located directly upstream of a gene where transcription initiates, and the main role of the promoter is to bind to the RNA polymerase II to position the transcription initiation complex [1]. Due to its proximity to gene bodies promoters are the easiest CRSs to define. It is generally agreed upon that the core promoter is the 100bp or so around the transcription start site (TSS) containing the minimal sequences required for transcription initiation [2]. From here on any reference to promoters will refer to the core promoter unless otherwise defined.

Because all core promoters appear to have the same immediate function, it was thought that the core promoter is a passive platform the binding of general transcription factors (GTFs). However, there is considerable diversity in core promoter sequences across the genome [3, 4]. The Initiator (Inr) motif encompasses the TSS and is currently the most prevalent known motif in Drosophila and humans [5, 6], but multiple Inr consensus motifs in humans have been proposed and all the proposed motifs appear to be highly degenerate [7–11], suggesting a lack of strong sequence requirements even at the TSS. In ribosomal genes, a slightly different motif, the polypyrimidine initiator (TCT) motif, is found at the TSS [12], perhaps because ribosomal genes need to be regulated differently than other genes [13]. Thus, different motifs at the TSS can be used to regulate distinct subsets of genes.

Upstream of the TSS, the TATA box is the most well-studied and is conserved from bacteria to eukaryotes. It is bound by the TATA-binding protein (TBP), which is thought to be an important component of the transcription initiation complex [14]. However, the TATA box only present in 10-20% of metazoan promoters [15] and TBP has been shown to be dispensable for transcription in mouse embryonic stem cells (mESCs) [16]. A similarly conserved element, the TFIIIB recognition element (BRE), can be found upstream or downstream of the TATA box [17, 18].

Human and Drosophila promoters start to diverge more downstream of the TSS. Both the
downstream promoter element (DPE) and the motif 10 element (MTE) were discovered in *Drosophila* as an alternative motif in TATA-less promoters and were thought to be conserved in human promoters [5, 19–22]. Both motifs bind to distinct components of the RNA polymerase complex. The DPE binds to TAF6 and TAF9 [20], and TAF1 was shown to structurally interact with both the MTE and DPE element *in vitro* [23]. Further analysis of more human promoters showed that the DPE element is hardly present while the MTE element cannot be identified [6, 9]. Instead, a downstream core element (DCE) that is important for human β-globin and adenovirus major late promoter activity [24, 25], and Vo ngoc and colleagues identified downstream core promoter region (DPR) sequences that appear to be important in human promoters [26]. The divergence between core promoters across evolution suggests that they might be important for species-specific gene regulation.

In addition to these more well-defined motifs, computational analyses have identified enriched sequences such as various Ohler motifs and the DNA replication-related element (DRE) in flies [5]. Studies of the Hepatitis B virus (HBV) X gene have also identified XCPE1, which might be present at low frequency in human promoters, and XCPE2 elements [27, 28]. Despite these efforts to identify and characterize core promoter elements, a substantial number of promoters still do not contain any known motifs [29, 30].

Promoters in vertebrates have also been found to be preferentially located in CpG islands, which are regions of the genome with high GC content and a high frequency of CpG dinucleotides [31]. These promoters tend to lack defined motifs [9, 32]. Taken together, while it is possible that there are additional core promoter motifs that have yet to be identified, it is likely that core promoters are very degenerate sequences and that there are no strict sequence requirements for transcription initiation.
1.1.2 The importance of enhancers (and silencers)

Enhancers and silencers are sequences that cause gene expression to increase or decrease respectively. These CRSs are composed of transcription factor (TF) binding sites that recruit TFs to interact with RNA polymerase and influence expression levels [33]. A complete review of enhancers and silencers is beyond the scope of this thesis, but the relevant features will be summarized. Because enhancers are relatively more well-studied than silencers I will focus on enhancers here, but many of the principles will be applicable to silencers as well.

The first enhancer to be discovered is from the SV40 virus. The SV40 enhancer was found to upregulate transcription even when located several kilobases upstream of the $\beta$-globin promoter regardless of its orientation [34]. Enhancers have since been found to be critical regulators of development and evolution by precisely coordinating spatial and temporal patterns of gene expression [35]. For example, the globin LCR can activate different globin genes at different time points during hematopoiesis [36], and the five $eve$ enhancers in various combinations are required for the proper generation for seven stripes during $Drosophila$ development [35].

While enhancers are often defined based on their ability to activate expression of a nearby promoter on episomal plasmid vectors, there are several examples of important developmental promoters that are located far away from their target genes. One such example is the ZRS enhancer, which is located 1Mb upstream of its target gene ($Shh$) [37]. The ENCODE consortium currently estimates that 72% of enhancers are located >2kb away from their target gene [38], while a more targeted study of promoter interactions identified more than 1000 ultra long-range (>500kb) interactions [39]. Thus, enhancers exert their functions over long distances in the human genome.

Enhancers also have to work together to regulate genes. By using chromosome conformation capture techniques to map the interaction landscape of promoters, two studies found that promoters frequently interact with multiple enhancers and/or other promoters [39, 40]. Estimates for the average number of enhancers interacting with a single promoter range from
2-6 enhancers [40–42]. Another study using CRISPR-interference (CRISPRi) of thousands of enhancers found that almost 25% of genes in their assay were affected by >1 enhancer [43]. Additionally, important master regulator genes are often regulated by clusters of enhancers [44–46]. These enhancers can be clustered linearly or in 3D space to form hubs that regulate one or multiple genes [47–49]. It is important to understand how such higher-order interactions between enhancers and other CRSs work together to regulate gene expression.

### 1.1.3 The perplexing case of insulators

Insulators are the least well-understood of the CRSs. Unlike promoters and enhancers, there are multiple assays that can be used to call insulator activity. The earliest definition of an insulator is that it can “insulate” transgenes from chromosomal position effects (PE). In vertebrate cell lines, flanking reporter genes with enhancer-like sequences that drive high levels of expression can protect the gene from PE [50–52]. Subsequently, studies in *Drosophila* identified the *scs* and gypsy insulators which protect the *white* mini- and maxigenes from activating and repressing effects respectively [53, 54]. Mutations in a CRS in the bithorax cluster leads to improper activation of the *Abd-B* gene, and this CRS was later identified to be an enhancer-blocking (EB) insulator, suggesting that EB insulators are important for endogenous gene regulation [55]. Importantly, unlike the earlier mammalian studies, the insulator sequences in flies do not have any regulatory activity of their own.

These studies in flies inspired the search for similarly “neutral” insulators that only function as boundaries in mammalian cells. Chung *et al.* identified a 5′ region of the chicken β-globin domain that acts as an EB insulator in both *Drosophila* and human erythroid cells, which they name cHS4 [56]. In this study, they establish the EB assay in human cells, where insulators are tested for their ability to block an enhancer from activating expression a promoter on episomal constructs. Similarly, other studies showed that the gypsy insulator was found to block enhancer activity when placed between enhancers and promoters [57–62]. This flavor of
assays represents a different way of defining insulators and have since been used to identify a whole suite of EB insulators [63–72].

A third definition of insulators is a sequence that has the ability to block heterochromatin-mediated gene silencing. These insulators act as “barrier” elements to prevent the spreading of heterochromatin into promoters, thereby preventing it from being silenced [73]. They have been shown to work in yeast, Drosophila and mammalian cells alike and are mostly assayed by their ability to prevent transgene silencing, similar to the PE assay above [73, 74]. More recently, genome-wide methods have enabled the identification of putative boundaries by looking for transition points between histone modifications in the genome [64, 75–78]. Using a synthetic experimental setup to recruit different chromatin repressors to plasmids, van der Vlag et al. showed that insulators differ in their abilities to block different repression mechanisms [79]. A more recent study tethering chromatin repressors to reporter genes integrated into the genome showed that cHS4 can only block HDAC- but not KRAB-mediated silencing [80]. Thus, there appears to be heterogeneity in activity even within a given class of insulators.

Finally, there is a lot of recent excitement about insulators due to the renewed interested in 3D chromosome structure. Early on, the matrix attachment region (MAR) in chicken cells was shown to protect against PE [52], suggesting a link between nuclear organization and insulator activity. Since two copies of the insulator flanking the reporter seemed to be important for the scs and cHS4 to protect against enhancers, the authors speculated that the insulators might influence 3D genome organization [53, 56]. Detailed dissections of the requirements for gypsy EB activity also led to idea that the promoter is looped out of the insulators reach [81]. This model gained even more traction when it became clear that the genome is organized into functional domains of gene expression [82].

The subsequent development of chromosome conformation technologies led to a series of papers showing that chromosomes are organized into highly self-interacting genomic units commonly referred to as topologically associating domains (TADs) [83–86]. TADs are separated
by boundaries that are thought to prevent CRS crosstalk between domains, which is conceptually related to the proposed insulator looping model previously proposed. Additionally, the sequence features of TAD boundaries overlap with sequences that are important for insulator activity. The CCCTC-binding factor (CTCF) and RNA polymerase III transcribed sequences such as tRNAs and B1/B2 SINEs (short interspersed nuclear elements) are enriched at boundaries [83]. CTCF had previously been found to be important for EB and/or barrier function in a number of insulators [65–67, 71, 87]. Raab and colleagues screened tRNA genes for EB and barrier activity and found that at least some tRNAs can act as insulators [70], while at least two SINEs have been shown to act as barrier insulators [88, 89]. These apparent similarities led to the current model that boundaries are insulator elements, and that organization of the 3D genome is the predominant mechanism of insulator activity [90]. Despite contradictory evidence about the insulating role of boundaries [91], the terms CTCF binding site, insulator and boundary are often used interchangeably.

It is clear that the term insulator encompasses sequences that perform a wide range of functions, and that it is necessary to clarify the assay used when referring to an insulator. These functions are not mutually exclusive. For example, it is likely that insulators block PE by some combination of EB and silencing protection, while the genome organizing function of insulators may directly or indirectly lead to EB. Some insulators, such as cHS4, have EB activity and can prevent heterochromatin spreading and PE, but distinct sequence features are required for their activities [92]. Other insulators, like gypsy, depend on the same TFBSs for EB and PE protection [82]. Many identified EB insulators do not contain a CTCF binding site, and CTCF is not always necessary for function [68, 69]. Carefully clarifying and separating different insulator modes of action will be required to better understand this class of CRSs.
1.2 Chromatin environments and gene expression

In the nucleus, DNA is packaged as chromatin with nucleotides wrapped around histone proteins to form nucleosomes. The histones can be post-translationally modified and can be important for proper gene regulation [93]. In addition, DNA itself can be methylated to alter gene expression [94]. A classic phenomenon exemplifying chromatin effects is position effect variegation (PEV). When a reporter gene is integrated close to heterochromatin, the histone modifications can spread and silence the gene in a stochastic manner, leading to a variegated phenotype [95]. To understand the potential regulatory roles of chromatin modifications, a large amount of effort has been spent on mapping these modifications across many cell types [38, 96–99]. Correlating these modifications with gene expression patterns of endogenous genes or integrated reporter genes have allowed us to identify and predict CRSs and/or gene expression patterns with some degree of accuracy [100–103].

Different histone modifications can recruit distinct TFs or chromatin remodelers to DNA [104]. Histone modifications can also change the structure of the nucleosome to exclude or promote TF binding in a cell-type specific manner [69, 105–107]. DNA methylation is traditionally thought to repress gene expression, but has recently been associated with active reporter genes [94]. This suggests that chromatin modifications can be an additional regulatory layer on top of CRSs. A key question is how these modifications influence CRS activity and vice versa. Our lab previously showed that histone modifications influence hundreds of enhancers equally regardless of their sequence, suggesting that chromatin modifications and CRSs independently modulate gene expression [108]. A different group, however, found that promoters differ in their sensitivities to repression by lamina-associated domains by measuring multiple promoters across thousands of genomic locations [109]. Additionally, chromatin modifications can influence cell-to-cell variability in gene expression (discussed in Section 1.5). Furthermore, DNA sequence is sufficient to direct DNA methylation and repress transcription, but strong promoters can overcome this repression [110]. A major focus of this thesis is understanding
how chromatin environments and CRSs work together to influence gene expression.

1.3 The eternal question of specificity

The latest ENCODE study estimates that there are approximately 800,000 enhancer-like elements in the human genome [38]. These enhancers regulate the expression of ~20,000 genes with exquisite specificity to generate diverse cell types. Given that the genome is only ~2% coding and that enhancers are often located far away from their target genes, many enhancers skip over the closest promoter to activate a different promoter. One remarkable example is the ZRS enhancer, which is located in an intron of the Lmbr1 gene but regulates the expression of Shh, a gene that is more than 1Mb away from the enhancer [37]. Using genome-wide promoter capture approaches, Li et al. found that ≥40% of enhancers do not interact with the closest promoter [39], while Sanyal et al. showed that ~50% of enhancers do not interact with the closest expressed TSS [40].

There are two prevailing hypotheses for how enhancer-promoter (E-P) specificity is achieved [35, 82]. First, different core promoters may be compatible with distinct subsets of enhancers, allowing them to only respond to the correct enhancers in the genome. Second, insulator sequences along the genome may block enhancers or other CRSs to functionally segregate genes into their appropriate expression domains. The evidence supporting each of these models is reviewed below.

1.3.1 Diverse core promoter motifs can encode specificity

The promoter is responsible for integrating signals from all the other CRSs and the chromatin environment [1]. Given the sequence diversity in core promoter sequences, it is attractive to hypothesize that promoters are highly selective and might only respond to very specific genomic signals, as opposed to the classical model that core promoters are inert platforms for transcription initiation.
There are several examples of promoters that have distinct expression patterns despite their proximity in the genome. In Drosophila, the *gsb* and *gsbn* genes are located close to each other but have distinct tissue-specific expression profiles because their respective enhancers can only activate their target promoter [111]. The *dpp* enhancers also appear to be very specific for some promoters [112]. When a transgene was integrated into an intron of the nearby *oaf* gene, the *dpp* enhancers preferentially activated the transgene over *oaf* even though the *oaf* and *Slh* promoters are closer. Replacing the *oaf* promoter with *dpp* promoters was also sufficient to cause it use *dpp* enhancers. The IAB5 enhancer also preferentially activates the *Abd-B* promoter over the *abd-a* promoter, even though both promoters are approximately the same distance from the enhancer [113]. These results suggested that the core promoters contain the information necessary for responding to specific enhancers.

**Sequence determinants of core promoter specificity**

In some cases, the sequence motifs regulating the specificity have been identified. Promoters containing either a TATA box or DPE integrated into the same location in the fly genome exhibit different expression patterns, suggesting that promoters with different motifs can trap different enhancers from the same genomic location [114]. The AE1 and IAB5 enhancers also preferentially activate promoters with a TATA box over an Inr [115]. Interestingly, the same study also found that the *rhomboid* (*rho*) neuroectoderm enhancer (NEE) does not have a preference for TATA boxes, indicating that perhaps some enhancers act as more general activators. In mouse lymphocyte cells, mutating the Inr motif and adding a TATA box instead abolished the expression of the terminal deoxynucleotidyltransferase gene [116]. Xu *et al.* found that an enhancer of the human β-globin gene selectively activates TATA-containing promoters but not Inr-containing promoters [117]. Additionally, the sequences surrounding the TATA box have also been found to be modulate activation by Gal4 relative to basal transcriptional levels [118].

The specificity of core promoters for some enhancers is likely to be mediated by their
preferences for different TFs. Using a variety of synthetic core promoters containing either a TATA box, Inr or both motifs, Emami et al. showed that promoters containing different motifs responded differently to the Sp1 and VP16 activation domains [119]. The proteins NC2 and Mot1 preferentially activate DPE-containing promoters over TATA-containing promoters by binding to and inhibiting TBP, which binds to the TATA box, and TBP can also repressed DPE-containing promoters [120, 121]. A suite of GTFs that was able to activate transcription from TATA and DCE core promoters was not sufficient to transcribe DPE-containing core promoters in vitro [25]. In Drosophila, both Caudal and Dorsal prefer to activate DPE- over TATA-containing core promoters [122, 123]. More recently, a genome-wide survey of promoter cofactor dependencies showed that promoters containing different promoter motifs were found to depend on distinct sets of cofactors in both fly and human cells [30]. Bell et al. further found that different core promoters have different rate-limiting steps that are released by specific cofactors, which might explain their differential cofactor dependencies [124]. By binding or responding only to specific TFs, core promoters can selectively respond to enhancers that recruit different TFs in the genome.

Even for the same core promoter motif, different versions of the motif are not equivalent. The Drosophila Adh gene has two promoters that are selectively utilized during different developmental stages [125]. The switch in core promoters is partially due to the presence of distinct Inr elements in the two promoters, and swapping the Inr elements in both promoters was sufficient to change promoter utilization [126, 127]. The proximal promoter Inr element was also found to bind to a repressor protein, suggesting that different versions of the Inr may perform slightly different functions [127]. In the human hsp70 promoter, swapping its TATA box motif (TATAA) with the SV40 promoter TATA box (TATTAA) did not impact promoter function but abolished its response to E1A activation [128]. Similarly, replacing the TATA box in the human myoglobin promoter (TATAAAA) with the the SV40 TATA box (TATTTAT) rendered it incompatible with the muscle-specific enhancer (MSE), while the myoglobin TATA
box was sufficient to cause the SV40 promoter to respond to the MSE [129]. Thus, there is functional diversity even within the same class of core promoter motifs.

**High-throughput testing of core promoter specificity**

More high-throughput studies of core promoters in *Drosophila* have now shown that promoters can be classified into at least two classes, the housekeeping (hk) and developmental (dev) core promoters. Hk promoters are broadly expressed across tissues, while dev promoters are cell-type specific genes. The two classes of promoters are enriched in different sets of motifs, suggesting that there might be functional separation between the classes [6, 13]. Using a representative hk and dev core promoter and thousands of random genomic fragments, Zabidi *et al.* showed that the two promoters respond to distinct sets of enhancers in the genome [130]. The enhancers identified here were also shown to bind to distinct architectural proteins with different chromatin interaction features [131]. Later, Arnold *et al.* performed the converse experiment with representative hk and dev enhancers and found that distinct sets of sequences responded to each class of enhancers, with various motifs enriched in one class over the other [132]. These assays were performed on episomal plasmids outside the influence of the endogenous gene environment, suggesting that E-P compatibility is encoded in the DNA sequence itself.

Taken together, the literature suggests that the sequence diversity of core promoters at least partially encodes E-P specificity in the genome. However, many of these studies dissected one promoter at a time, and at least at the time I started my thesis, the genome-wide assays had only been performed in *Drosophila*. It remains to be seen how generalizable these observations are in mammalian cells.
1.3.2 Insulators as mediators of specificity

The diversity in insulator functions suggests that there are different ways the various types of insulators mediate specificity in the genome. The function of EB and loop domains are inextricably linked and will be discussed together, while barrier insulators are likely to function via a different mechanism and will be discussed in a separate section.

Enhancer-blockers regulate cell-type specific genome structure

Multiple models for how EB insulators may regulate E-P specificity have been proposed. However, the model with overwhelming supporting evidence is the structural/looping model. In this model, insulators interact with each other to form domains that include or exclude CRSs, thereby limiting the CRSs that a promoter may encounter in the genome. Other models, including the decoy and inhibition of tracking models, lack direct experimental evidence and will be discussed briefly at the end of this section.

Evidence for the structural model

Insulators have been observed to self-interact to form chromosome loops in the genome. In Drosophila, the scs and scs’ insulators that protect the white transgenes from PE were found to form long-range interactions in the genome, and the proteins that bind to scs and scs’ interact in vitro and in vivo [133]. Pairing of the gypsy insulator can lead to a multitude of outcomes. Flanking either the enhancer or promoter with gypsy insulators is necessary for EB activity, suggesting that gypsy functions by looping the enhancer or promoter out and excluding E-P interactions [134]. Two independent studies showed that when two gypsy insulators are placed between an enhancer and promoter, the EB effect of the insulator was neutralized, a phenomenon also known as insulator bypass [62, 135]. Both studies proposed a model where the insulators pair with each to loop out the intervening DNA and facilitates E-P interaction. This interaction appears to be specific to pairs of gypsy insulators—pairs of scs/scs’ insulators do not
neutralize EB activity, heterologous pairs of gypsy and a different insulator does not neutralize EB, and three gypsy insulators between the E-P pair appears to restore EB activity [136, 137]. On the other hand, gypsy combined with the MCP insulator was able cause insulator bypass, suggesting that gypsy may only pair with specific heterologous insulators [138]. Interestingly, Kuhn et al. also showed that placing the scs insulator between a pair of gypsy insulators also attenuated scs EB activity, supporting the idea that gypsy pairing loops out the intervening DNA and prevents its function [137].

In mammalian cells, the only protein that has been observed to have insulator activity is CTCF. CTCF was originally identified from the cHS4 insulator, where it was shown to be required for EB activity [87]. The observation that two copies of the cHS4 insulator flanking the insulator was required for EB activity provided support for the structural model over the decoy model [56]. There are many examples of the impact of CTCF on looping and gene expression at individual loci, comprehensively reviewed in [91], but two insulators that have bona fide EB activity (as measured by plasmid EB assays) are highlighted here. Along with other hypersensitive sites in the mouse β-globin locus, the orthologous HS5 element mediates looping interactions to bring the locus control region (LCR) into close contact specifically with the active globin gene in a CTCF-dependent manner [139, 140]. The intervening regions, which contains promoters for other non-expressed globin genes, are looped out to allow for specific expression of the correct globin genes during development. The loops are also only present in erythroid cells but not brain cells where globin genes are not expressed, suggesting that at least some CTCF loops can be cell-type dependent [139]. Surprisingly, the loss of these loops does not impact globin gene expression, indicating the chromatin structure may not be causal for gene expression [140]. A different study showed that when a second HS5 element was introduced between the LCR and globin genes, the LCR is looped out and unable to drive globin transcription [141]. In the α-globin locus, CTCF sites flanking the enhancers and α-globin promoters interact in a cell-type specific manner to promote E-P interactions [142]. Deletion of
one of the CTCF sites leads to ectopic interaction and expression of a different promoter. These two examples highlight the divergent roles of CTCF-dependent insulator elements.

To test the sufficiency of CTCF binding sites (CBSs) for insulator activity, Tsujimura et al. concatenated CBSs from the mouse Tgfap2c-Bmp7 locus to generate a putative synthetic insulator called STITCH [143]. Integrating STITCH between the MYC super-enhancer and promoter in human induced pluripotent stem cells (iPSCs) led to decreased interactions between the enhancer and promoter and a dramatic reduction in MYC expression, but integrating STITCH elsewhere in the locus had no effect. Interestingly, deletion of the CBSs individually and in combination revealed that not all the CBSs are equivalent and in some cases decoupled the relationship between E-P interactions and expression, suggesting that other mechanisms besides genome topology may contribute to insulator activity. Similarly, Huang et al. tested the ability of several different CBSs to block the Sox2 super-enhancer in mESCs and found that one CBS alone was not sufficient for insulator activity [144]. Concatenating multiple CBSs can increase EB activity, but unlike the MYC locus, the maximal reduction in gene expression observed was ~35%. Notably, only combinations of CBSs from TAD boundaries could function as insulators—combinations of CBSs from non-TAD boundaries did not display any EB abilities despite high levels of CTCF binding. This suggests that other proteins besides CTCF are required for EB in the genome. A different study at the Sox2 locus showed that the addition of three concatenated CBSs between the Sox2 control region (SCR) and Sox2 in mice caused large changes in TAD structure but had minimal effects on gene expression [145]. In fact, mice containing the additional CBSs developed and implanted normally. Interestingly, cells in the anterior foregut with the same integrated CBSs displayed a reduction in Sox2 expression, which lead to an impairment in lung and stomach development. Based on these observations, the authors suggest that high affinity interactions between the SCR and Sox2 allows it to overcome the inserted boundary element, while lower affinity interactions between the anterior foregut enhancers and the Sox2 promoter renders it susceptible to the novel CBSs, providing a potential
model for how cell-type specificity might be mediated by insulator and enhancer sequences.

CTCF often interacts with Cohesin in the genome to mediate insulator function [146]. For example, Cohesin is required for CTCF-mediated insulation at the H19/IGF2 [146, 147] and HOXA loci [148]. CTCF mediates chromatin looping between pairs of convergent CBSs [149] and is crucial for maintaining 3D genome organization [46, 150–153]. Yet, the loss of CTCF or Cohesin has minimal impact on global gene expression levels [151–154]. Super-resolution chromatin tracing experiment has shown the loss of Cohesin reduces interactions between adjacent domains, while the loss of CTCF increases mixing across boundaries and enables cross-TAD interaction [155, 156]. The change in topological structure can impact the expression of genes near boundaries [155]. However, these single-cell experiments also show that chromatin and TAD structures are exhibit high cell-to-cell variability, arguing against a strong requirement of chromosome structure for gene regulation [157].

TFIIIC-transcribed regions such as tRNAs and SINE retrotransposons have also been implicated in insulator activity and genome organization [158]. In Drosophila, tRNA genes across the genome cluster in the nucleolus and insulator bodies [159, 160]. In human cells, tRNA genes have been shown to interact more frequently with other tRNA genes [70]. SINE B2 elements are also enriched for ChAHP complex binding, which competes with CTCF for binding to modulate genome organization [161]. TFIIIC also appears to interact with Cohesin [162], suggesting that different classes of insulators may converge on similar mechanisms to control the spatial organization of the genome.

Together, these results support a model where looping between insulator sites organizes chromatin structure to modulate E-P interactions, but the impact of these loops on gene expression remains controversial.

Other models for enhancer-blocking mediated specificity

In contrast to the structural model, there are numerous examples where one copy of the insulator between the enhancer and promoter is sufficient to block E-P interactions. Cai &
Levine showed that one copy of gypsy is sufficient to block the eve stripe enhancers, and that gypsy might selectively block enhancers between the stripe 2 enhancer was more efficiently blocked than the stripe 3 enhancer [60]. A similar result was shown with the Fab-7 insulator, which was able to block the ftz NE enhancer more effectively when it was paired with hsp70 compared to the ftz promoter [55]. A fascinating study comparing the ability of a SuHw insulator to block the AE1 enhancer in the presence of other, competing promoters showed that the ability of the insulator to block AE1 depended on the “competitiveness” of the other promoter, where a more “competitive” promoter enhances EB by the insulator [163]. These observations imply that single insulators may have some specificity for certain enhancers in the genome.

This could be interpreted to mean that the insulator is blocking an enhancer “tracking” signal from reaching the promoter, but this model is unlikely due to the lack of support for the enhancer tracking model in the genome [164, 165]. Another popular model is the decoy model, where insulators bind directly to enhancers and/or promoters to prevent E-P interactions [166]. The observation that GTFs are sometimes required for for EB activity has been thought to provide support for this model [167]. The Fab-7 insulator was found to directly interact with the AbdB promoter, but the regulatory role of this interaction remains unknown [168]. The strongest direct evidence for this role comes from a study of the white enhancer on transgenes. Two copies of the gypsy insulator and loop formation is required to block E-P interactions, but direct interaction between the insulator and enhancer is also important for full EB [134]. However, a key criterion of EB insulators is that it only affects E-P interactions when placed between an enhancer and promoter, but not in any other configuration. If the decoy model were true, then the insulator should be able to interfere with enhancers or promoters regardless of its position. Indeed, while it was first suggested that the retrotransposon Idefix can act as an insulator [169], further studies showed that it acts as a decoy to prevent the white enhancer from contacting its promoter even when it was located upstream of the enhancer [170], suggesting
that promoter competition is a separate mechanism from insulator activity. Taken together, the
evidence supporting the idea that insulators act via the decoy model is generally sparse and
indirect, suggesting that the decoy model is not the primary mode of EB activity.

How, then, can we explain the ability of one insulator copy to block E-P interactions? The parsimonious explanation for a genome-integrated assay would be that any integrated
insulator will find a interacting partner in the genome to form loops, fulfilling the structural
model of EB activity. On the other hand, EB insulators are traditionally defined by episomal
assays, where there is no other insulator to drive loop formation. In this case, it is possible
that the insulator might be blocking tracking signals from the enhancer, since the enhancer,
insulator and promoter are usually in close proximity on plasmids. This could explain why EB
insulators that work on plasmids sometimes fail to block enhancers in the genome. EB could
be fundamentally different between short-range enhancers (on plasmids or in the genome) and
long-range enhancers (in the genome). Further research into the requirements for insulator
activity in different E-P contexts will be required to understand the mechanisms of cell-type
specific enhancer blocking.

**Barrier insulators may demarcate chromatin boundaries**

Barrier insulators are most well-characterized in the yeast *S. cerevisiae*. In the yeast mating
type locus, the barriers HMR-L and HMR-R flank the mating-type genes *MATα* and *MATα*
limit the repressive domain to only these two genes [171]. Moving or deleting the elements
leads to spreading of the silenced domain, and the HMR-R can also block telomeric silencing.
Other barrier elements have also been observed to limit telomeric silencing [172, 173].

In *Drosophila*, approximately half of the sharp borders at the boundaries of H3K27me3
contain insulator elements, but these boundaries also tend to coincide with transcription
of active genes [77]. Similarly, most insulator binding sites are not in close proximity of
a H3K27me3 domain [78]. In both studies, knockdown of insulator proteins led to minor
spreading of H3K27me3 and almost no changes in gene expression [77, 78]. Additionally,
degrading Cohesin in mESCs led to limited spread of H3K27me3 domains [151]. These studies argue against the model of barrier insulators acting as chromatin boundaries to limit gene expression domains.

Other lines of evidence seem to suggest otherwise. A more recent study in Drosophila showed that TAD boundaries tend to bind to canonical insulator proteins and terminate H3K27me3 domains, suggesting that insulators may limit H3K27me3 spread [174]. In the bithorax locus, the Fab-7 insulator separates active and inactive subdomains, and loss of the insulator causes the two domains to fuse and behave like a single cis-regulatory domain [175]. In human cells, only a small fraction of the H3K27me3 domain boundaries contain CTCF binding sites, but these CTCF-marked boundaries are cell-type specific, suggesting that insulators may demarcate cell-type specific boundaries [176]. The cHS4 insulator separates a H3K9me3-silenced domain and the hyperacetylated chicken β-globin locus, suggesting that it can prevent spreading of silenced chromatin domains into the globin genes [177]. This has been attributed to high levels of acetylation in the cHS4 element, which initiates the spreading of acetylated histones and counteracts silencing histone modifications [177, 178]. The pattern of modifications is found only in erythroid cells but not brain cells, suggesting that cHS4 acts as a cell-type specific barrier [177]. In transgene assays, cHS4 can sometimes block silencing of integrated reporter genes. Silencing of the reporter gene is correlated with loss of nuclease accessibility and acetylation, while the insulator construct retains accessibility and is hyperacetylated [179].

A more direct test of insulator barrier activity was performed at the HOXA locus in mouse and human cells. In mESCs, the HOXA locus is marked with bivalent chromatin (H3K4me3 and H3K27me3). Upon differentiation into neural progenitors, the locus is split into an active (H3K4me3) and inactive (H3K27me3) domain demarcated by the CBS5 insulator [148]. Reducing CTCF binding at the insulator by knocking down CTCF leads to heterochromatin (H3K27me3) spreading into the active domain and a concomitant decrease in gene expression.
A different study at the Wnt4 locus found that the protein Wt1 recruits different chromatin modifiers in kidney mesenchyme and epicardium cells to generate active or inactive chromatin in different cell types, which is correlated with expression and repression of Wnt4 respectively. This chromatin state domain is demarcated by flanking CTCF/Cohesin boundaries [180]. Knocking down CTCF leads to spreading of the chromatin modifications into flanking chromatin and corresponding upregulation (in kidney mesenchyme cells) or downregulation (in epicardium cells) of genes outside the domain.

While there is conflicting evidence on the global importance of barrier insulators in demarcating chromatin domains, studies at individual loci clearly demonstrate the importance of barriers in setting cell-type specific boundaries. It will be important to understand the genomic context requirements for barrier function in the genome.

1.4 Methods for understanding interaction between CRSs and specificity

Plasmid-based reporter assays have been very useful for studying the influence of CRSs on gene expression. A putative CRS is cloned upstream of a minimal promoter driving the expression of a reporter gene, and the resulting plasmid is transfected into cells or tissues. By measuring the expression of the reporter gene, we can determine how much the CRS changes expression relative to the minimal promoter. The need to clone and transfect individual CRSs made this a relatively low-throughput assay, and the data generated was not sufficient to learn generalizable rules about gene expression. To address this gap our lab and others have previously developed high-throughput massively parallel reporter assays (MPRAs) that allow us to measure the activity of thousands of CRSs at the same time [181–183]. MPRAs have since been used to systematically understand the effects of DNA sequence in individual CRSs [184].

To understand how multiple CRSs function together to generate specificity, we need to be
able to systematically combine and measure the effects of different enhancers/insulators with
different promoters. This has been done for a variety of loci at low-throughput on plasmids
[55, 60, 185, 186]. More recent work has extended plasmid-based MPRAs to test combinations
of enhancers and promoters, allowing for high-throughput testing of CRS compatibility [130,
132, 187, 188]. However, plasmid-based assays are unable to capture the effects of CRSs in
a chromatinized, genomic environment where they usually function and by design requires
researchers to select CRSs to test. Thus, there is a need to understand how CRSs function
together in the genome.

1.4.1 Leveraging position effects

We can leverage PE to understand how CRSs function in the context of chromatin with other
endogenous CRSs. Chromosomal PE occurs because integrated transgenes trap enhancers
and take on the surrounding chromatin environment, both of which can influence expression
of the transgene [95, 189, 190]. Each location in the genome therefore represents a suite of
histone modifications and CRSs that work in concert to influence gene expression. By testing
defined CRSs in multiple locations in the genome, we can learn how the same CRSs interact
with diverse CRSs and chromatin environments across the genome.

To understand how different genomic environments influence gene expression, Akhtar et
al. developed the Thousands of Reporters In Parallel (TRIP) method [103]. Reporter genes
are barcoded with a diverse library of random barcodes and integrated into the genome with
piggyBac transposase, and expression is calculated as RNA/DNA barcode counts. Using the
TRIP method to test multiple promoters, Leemans et al. later found that different promoters
have different intrinsic sensitivities to repression by lamina-associated domains, suggesting that
CRSs can interact with the LAD chromatin environment to effect different outcomes [109].
However, to understand how the same genomic environment interacts with different CRSs, we
need to be able to test multiple CRSs in the same genomic locations.
1.4.2 Genome-integrated massively parallel reporter assays

Over the years there have been several variants of directed genome integration methods to test multiple CRSs at the same genomic locations. The methods usually consist of integrating recombinase sites such as LoxP or FRT into the genome (which we will refer to as genomic “landing pads”, or LPs), then using the CRE or FLP recombinase respectively to recombine transgenes of interest these fixed locations. As a whole, this class of methods was generally referred to as recombinase-mediate cassette exchange (RMCE) [191]. In these early studies, after each LP cell line has been generated, a separate cell line will have to be generated for each transgene of interest. Thus, RMCE was a laborious and time consuming process and was extremely limited in throughput.

To increase the throughput of RMCE, Dickel et al. developed SIF-seq [192]. A library of putative CRSs were integrated into a LP expressing the green fluorescent protein (GFP) in mESCs, ensuring that there is only one CRS per cell. The GFP expression level of the cell therefore reflects the CRS activity. By sorting for GFP-positive cells and sequencing the CRSs in only those cells, the authors could identify active mouse enhancers. The drawback of this method, however, is that it relies on protein rather than RNA measurements, and so is not a direct readout of transcription. Furthermore, this still only allows us to test one genomic environment at a time.

To address this gap, our lab developed patchMPRA, a method to measure the effects of many CRSs at multiple locations at the same [108]. The method is conceptually similar to RMCE in that LP cell lines with one LP per cell are generated to allow targeted CRS integration by CRE recombinase. However, each LP is now barcoded with a unique DNA barcode that is mapped to a genomic location (genomic barcode, gBC). We then generate a library of CRSs where each library is linked to a unique CRS barcode (cBC), which are integrated into a pool of mixed LP cell lines. Because both barcodes are transcribed, we can sequence and tabulate the barcode pairs to get RNA and DNA counts of each cBC-gBC pair. The innovative barcoding method
therefore allows us to test a large library of CRSs at multiple locations in the genome. While this method had already been developed at the beginning of my thesis, there were technical aspects of the method that could be improved. This method is also limited in the numbers of locations that can be tested at the same time. Improving and iterating on this method was therefore one of the main goals of my thesis work.

1.5 Noise in gene expression

Besides regulating gene expression levels, CRSs and chromatin environments can also influence cell-to-cell variability. Populations of cells that contain the same DNA sequence can display remarkable levels of variability in expression of the same genes, with estimates suggesting that RNA counts of the same gene in neighboring identical cells can range from 0-700 copies [193]. This “noise” in gene expression is important for bet hedging in single-cell organisms [194–196] and for cell-type specification in multicellular organisms [197–200], amongst others. While noise can manifest at both the mRNA and protein levels, here I am primarily interested in the transcriptional mechanisms that generate noise and will thus focus on mRNA.

Cell-to-cell variability in mRNA levels is a consequence of thermodynamic fluctuations in the TFs that regulate gene expression. The stochastic changes in TF levels lead to a phenomenon called transcriptional bursting, where mRNA molecules are produced in bursts of several mRNAs at a time rather than as a constitutive Poissonian process [201, 202]. Models are still being developed to describe this process, but the simplest model assumes that the promoter stochastically switches between the active and the inactive states, and RNA can only be produced in the active state [203]. Correspondingly, the rate of switching between the active and inactive states is referred to as the burst frequency, the amount of RNA produced is the burst size and the duration a gene spends in the active state is the burst duration. Factors that modulate a gene’s burst parameters will necessarily impact gene expression noise.
1.5.1 Sources of gene expression noise

Conceptually, there are two major sources of gene expression noise [204, 205]. Intrinsic noise is determined by the genomic features at the gene loci. In the case where intrinsic noise dominates, each gene fluctuates independently and two genes in the same set of cells would not correlate. Extrinsic noise, on the other hand, is caused by fluctuations in the cellular environment. If extrinsic noise is dominant, then two genes in the same cells would exhibit a high degree of covariance. At the mRNA level, intrinsic noise appears to dominate in mammalian cells [193, 206], but the relative concentrations of TFs have also been shown to influence individual genes, suggesting that extrinsic noise may also play a role [207]. The lack of evidence for extrinsic noise could well be due to the lack of methods to study intrinsic vs extrinsic noise at scale in mammalian cells.

The chromosomal position of a gene has been shown to affect gene expression noise, indicating that there are features of genomic environments that can influence cell-to-cell variability [193, 208–212]. At the CRS level, enhancers were observed to increase the probability of gene expression (burst frequency), while promoters control burst sizes [213–218]. Only the gypsy insulator has been implicated in the control of noise, where it was found to also reduce burst frequency but not burst size [215]. Because the 3D genome also exhibits a high degree of cell-to-cell variability [157], it is not unreasonable to speculate that there might be a relationship between genome architecture and gene expression noise [219, 220].

Features of the chromatin environment can also modulate noise. Nucleosome positioning at the promoter is correlated with differences in gene expression noise [221, 222]. Various histone modifications and chromatin states have been correlated with changes in gene expression noise [212, 223–225], and the repressed chromatin state has been shown to increase noise [210, 226]. Histone acetylation has been linked to gene expression noise in a number of studies [227]. Cas9-based acetylation or deacetylation of reporter genes was sufficient to change burst frequencies [228–231], and different histone deacetylases (HDACs) were also found to regulate either burst
noise or frequency, suggesting that chromatin modifiers can influence gene expression noise differently [232]. Similarly, loss of TRIM24, a chromatin reader, changes burst duration but not burst frequency [219].

Contrary to the above studies, Suter et al. showed that while promoter sequence features can influence noise, histone modifications or chromosomal locations did not markedly change gene expression noise [233]. A study of calcium signalling genes also showed that transcriptional bursting at the level of individual alleles does not contribute much to noise [234]. A systematic study of the impact of genomic environments on expression noise will help resolve these conflicting studies.

1.5.2 Methods for studying gene expression noise

The method that enabled the study of mRNA expression distributions is RNA Fluorescence In Situ Hybridization (RNA-FISH). Later developments to RNA-FISH led to single-molecule RNA-FISH (smFISH), which allowed researchers to count individual RNA molecules to generate quantitative expression distributions [235]. With FISH technologies, DNA probes carrying fluorescent tags are hybridized to the mRNAs of interest and imaged [236]. However, smFISH is performed on fixed cells and transcriptional burst parameters are inferred from fitting distributions to the measured data. To directly measure burst parameters, researchers can tag their gene of interest with MS2 or PP7 stem loops [237]. The MS2 or PP7 coat protein is fused to GFP and binds to the stem loops, allowing for live-cell tracking of mRNA molecules by GFP. Both methods have been very useful for measuring the noise of individual RNA molecules, but does not scale very well for larger numbers of perturbations.

For more genome-wide studies, reporter genes are integrated into random locations in the genomes and the single clones carrying a gene in one location are generated. The noisiness of the reporter gene in each clone can then be measured. To ensure that large numbers of locations can be assayed, two studies used flow cytometry to measure reporter gene expression
levels in single-cells, but flow cytometry reads out protein rather than mRNA levels [210, 212]. Dar et al. measured mRNA levels with smFISH but did not map the locations of the integrations and thus could not determine the genomic features that impact gene expression noise [209]. To get high-throughput measurements of more genes, researchers have turned to single-cell RNA sequencing (scRNA-seq) [217, 224, 226, 238–240] or spatial transcriptomics [234, 241, 242], which gives us the number of molecules per gene per cell. In theory, distributions of RNA molecules can be generated for all genes at once. This has been used to correlate various genomic features [217] or chromatin states with gene expression noise [226, 240]. However, scRNA-seq methods are limited in that they measure the noisiness of endogenous genes, which by definition have different promoters. Because promoters also influence gene expression noise, it is difficult to deconvolve the effect of the genomic environment from the promoter.

Taken together, systematic and controlled studies of the impact of genomic environments are limited by our current methodologies. Given that the studies with causative evidence are mostly performed at small numbers of loci while genome-wide studies provide only correlative conclusions, new methods to study the impact of chromosomal environments on gene expression noise are necessary.

1.6 Conclusion

Understanding how different CRSs and the chromatin environment work together to regulate genes has long been a central goal of biology. The idea that E-P specificity could be a result of biochemical compatibility between enhancers and promoters and/or insulators, through whatever mechanism, has been discussed since the early 1990s (and possibly even earlier) [59]. Yet, we remain woefully ignorant of the mechanisms by which different CRSs and the chromatin environment interact. This thesis advances the field by explicitly testing the interactions between different genomic environments with core promoters (Chapter 2) and insulators (Chapter 3) in a high-throughput manner. In Chapter 4, we systematically measured
the impact of genomic environments on gene expression noise to further our understanding of cell-to-cell variability, which is also a key player in developmental processes. We improved and iterated on existing methods to develop new methods that directly answer our questions. These studies shed light on old and new questions in transcriptional regulation and takes us one step closer to finding general rules of \textit{cis}-regulation in the genome.
Chapter 2: Genomic Environments Scale the Activities of Diverse Core Promoters

Barak Cohen and I conceived and designed the project. I designed and conducted all experiments and analyses. Barak Cohen and I wrote the manuscript. This study was published in 2022 in *Genome Research* (doi: 10.1101/gr.276025.121) [243].

2.1 Abstract

A classical model of gene regulation is that enhancers provide specificity while core promoters provide a modular site for the assembly of the basal transcriptional machinery. However, examples of core promoter specificity have led to an alternate hypothesis in which specificity is achieved by core promoters with different sequence motifs that respond differently to genomic environments containing different enhancers and chromatin landscapes. To distinguish between these models, we measured the activities of hundreds of diverse core promoters in four different genomic locations and, in a complementary experiment, six different core promoters at thousands of locations across the genome. While genomic locations had large effects on expression, the intrinsic activities of different classes of promoters were preserved across genomic locations, suggesting that core promoters are modular regulatory elements whose activities are independently scaled up or down by different genomic locations. This scaling of promoter activities is non-linear and depends on the genomic location and the strength of the core promoter. Our results support the classical model of regulation in which diverse core promoter motifs set the intrinsic strengths of core promoters, which are then amplified or dampened by the activities of their genomic environments.
2.2 Introduction

In the classical model of gene regulation, the core promoter serves as a universal platform for the assembly of the basal transcriptional machinery, while the specificity of expression is provided by distal enhancers and the chromatin landscape. However, some examples of core promoter specificity seem to challenge this model. Several studies suggest that different core promoters are specific for distinct sets of enhancers [111, 112, 244, 245], and can even trap different enhancers at the same genomic location [114]. Some transcription factors also preferentially activate core promoters containing specific motifs [12, 30, 119, 246]. More recently, a genome-wide massively parallel reporter assay (MPRA) showed that housekeeping and developmental core promoters respond to distinct classes of enhancers [130, 132], arguing that enhancer-promoter compatibility contributes to specificity in the genome. These data have led to an alternate model in which core promoters with different sequence elements respond specifically to the enhancers and chromatin features in distinct genomic environments, which we refer to as the ‘promoter compatibility’ hypothesis. Determining whether the specificity of gene expression is governed by enhancers and chromatin features or by enhancer-promoter compatibility is crucial to understanding a variety of biological processes including cell-type specific regulatory programs and models of gene evolution.

The core promoter is the ~100bp region around the transcription start site and is responsible for accurately positioning RNA polymerase II and binding general transcription factors [3, 4]. It is now known that core promoters are a diverse set of sequences containing specific DNA sequence motifs, also termed core promoter elements or motifs. The most well-known core promoter motif is the TATA box, yet the TATA box is only present in 10-20% of metazoan core promoters [15], suggesting that other motifs might have evolved for different functions. The different motifs have been associated with different functions, for example, the TATA box is often enriched in developmental promoters and show a ‘sharp’ pattern of transcription initiation, while promoters with high CpG content tend to contain other less well-characterized
motifs and is thought to be associated with a broader pattern of transcription initiation [1].

A strong prediction of the promoter compatibility hypothesis is that the relative strengths of different core promoters will change at different genomic locations because the distal enhancers and chromatin environments at different locations will be compatible with different types of core promoters (Figure 2.1A). Here, we tested the promoter compatibility hypothesis by assaying hundreds of diverse core promoters at four different genomic locations and further extend our results genome-wide by assaying six core promoters across thousands of genomic locations.

2.3 Results

2.3.1 Measurement of diverse core promoter activities at different genomic locations

We first created a library in which diverse core promoters drive the expression of an mScarlet reporter gene. The library contains 676 133bp core promoters spanning a variety of promoter features from Haberle et al. [30]. The core promoters are derived from endogenous promoter sequences and include the most common mammalian core promoter motifs (TATA, DPE and TCT), CpG islands and housekeeping (hk) and developmental (dev) promoters that do not contain any known core promoter motifs (Table S2.1). To provide redundancy in the measurements, we included ten copies of each individual core promoter in the library, each with a unique barcode (promoter BC; pBC) in the 3’ UTR. Because basal expression of the core promoters was expected to be weak, we included a common proximal enhancer directly upstream of the core promoters to boost expression (Methods).

Using patchMPRA (parallel targeting of chromosome positions by MPRA), we measured the expression of the core promoter library in parallel at four genomic locations previously shown to have diverse expression levels and chromatin marks in K562 cell lines (Table S2.2 and Figure S2.1) [108]. Each cell line contains a single ‘landing pad’ at a different genomic location.
Figure 2.1: Measurements of core promoter library at four genomic locations by patchMPRA. (A) Schematic of gene regulation by the core promoter, adjacent cis-regulatory sequences and the genomic environment. (B) Schematic of patchMPRA method (see Methods for details). tagBFP: blue fluorescent protein; HSV-TK: herpes simplex virus thymidine kinase; gBC: genomic barcode; pBC: promoter barcode. (C) Reproducibility of core promoter measurements from independent patchMPRA transfections. (D) The expression of all core promoters in the library at each genomic location.

Each landing pad has a unique genomic barcode (gBC) indicating its location in the genome and a pair of asymmetric Lox sites to facilitate site-specific recombination of the library. We pooled the four landing pad lines and integrated the library into the cells by cotransfection with CRE recombinase [108]. When a library member recombines into a landing pad it produces a
transcript with two unique barcodes in its 3’ UTR; a pBC specifying the core promoter and a
gBC indicating its genomic location. By tabulating the pBC-gBC pairs in the mRNAs from the
pool we obtained expression measurements for every core promoter at each genomic location
in parallel (Figure 2.1B).

We obtained reliable measurements of every core promoter at all four genomic locations.
We recovered 70-80% of all promoter barcodes and 99% of all promoters at all landing pads
(Figure S2.2A-B). The three biological replicates showed high reproducibility (average Pearson’s
\( r = 0.87 \)) (Figure 2.1C; Figure S2.2C-D) and the environments of the landing pads had large
effects on library expression that were consistent with previous studies (compare Figure 2.1D
to Figure S2.3A) [108] indicating that the genomic environment is not drastically altered by a
diverse core promoter library. To ensure that the genomic environment effect is not driven by
the expression of nearby promoters, we examined the expression of the endogenous genes that
are closest to the landing pads and found that the expression of these genes do not correlate
with the average expression in landing pads Figure S2.2E. The data allowed us to compare the
effects of the four genomic environments on the different classes of core promoters.

2.3.2 The effects of genomic locations on core promoters

The promoter compatibility hypothesis predicts that the same genomic environment will
impact different classes of promoters differently. In contrast to this prediction, the genomic
effect was similar on all promoter classes: more permissive genomic locations boosted the
expression of all promoter classes regardless of their motif composition or their hk or dev
designation (Figure 2.2A). However, the magnitude of the genomic effect is not the same for
all promoter classes. To quantify the contribution of the genomic location and core promoters
to gene expression we performed ANOVA on each class of promoters. In general, genomic
locations have a larger effects on dev promoters than hk promoters regardless of their motif
composition (Figure 2.2B). Thus, we did not distinguish between the motif classes and focused
Figure 2.2: Effects of genomic locations on core promoter activity. (A) Expression of each class of core promoter motifs at each genomic location. (B) Amount of variance explained by core promoter and genomic location respectively using linear models fit on each class of core promoters separately. (C) Pairwise correlations (Pearson’s r) of core promoter activity between the different genomic locations. (D) All pairwise correlations (Pearson’s r) between genomic locations for hk and dev core promoters.

We next examined whether hk and dev core promoter activities are scaled by different genomic environments. We define scaling as the degree to which core promoter activities correlate between genomic locations. High correlations between genomic locations indicate that the rank order of core promoter activities is preserved across genomic locations. While promoter activities were highly correlated between genomic locations regardless of the class of promoter (Pearson’s r = 0.74 - 0.9, Spearman’s ρ = 0.72 - 0.88) (Figure 2.2C), dev promoters were consistently less correlated than hk promoters (Figure 2.2D). Dividing the promoters into classes containing different motifs showed that each class also had substantial differences
in correlations between genomic locations (Figure S2.3B). These results do not depend on the proximal enhancer immediately adjacent to the core promoter used to boost expression because replicate experiments at locations 1-3 without the enhancer yielded similar results (Figure S2.4A-C). The expression of libraries with and without the proximal enhancer is also largely correlated at locations 1-3 (Figure S2.4D), which suggests that scaling by different genomic locations does not depend on the proximal enhancer. Taken together these results suggest that genomic environments scale the activities of all core promoters, but the quantitative extent of scaling can differ between promoter classes.

2.3.3 Intrinsic promoter strength explains differences between promoter classes

One difference between hk and dev promoters in our library is that they have different mean levels of expression—hk promoters are consistently stronger than dev promoters at all genomic locations (Figure 2.2A and Figure S2.5A). Thus, any differences between hk and dev promoters might be confounded by their difference in strength. To test if strength explains the differences between hk and dev promoters, we divided all core promoters into strong or weak bins based on their strengths and sampled equal numbers of hk and dev promoters within each bin to avoid confounding the results by hk/dev class. Plotting the effect of genomic position on strong and weak promoters showed that the direction of the effect was the same, but that there were larger differences between genomic locations for weak promoters (Figure 2.3A). We quantified the contributions of genomic locations and promoters within strong and weak bins respectively and found that the genomic environment has a larger impact on weak promoters compared to strong promoters (Figure 2.3B). For strong promoters, genomic environments and core promoters contribute almost equally to gene expression (~33% and ~47% respectively), but for weak promoters, genomic environments contribute ~61%, while core promoters contribute only ~12%. Weak promoters are also consistently less correlated than strong promoters (Figure 2.3C).
Again, assaying the library without an upstream proximal enhancer at locations 1-3 showed similar results (Figure S2.5B-C). Finally, we sampled sets of hk and dev promoters with similar average strengths (Figure S2.5D-F) and compared their correlations across genomic locations. In the strong and intermediate strength subsets, correlations across genomic locations are comparable between hk and dev promoters (Figure S2.5G-H). In the weaker subset of promoters, the correlations of hk promoters are weaker than the dev promoters, a result driven by low correlations in loc4, which might suggest that there are specific interactions between this subset of hk promoters and loc4 (Figure S2.5I). Thus, there may be additional interactions between weak promoters and other genomic locations that fall below our threshold of detection. Our data cannot rule out the possibility of extensive interactions between weak promoters and specific genomic locations. However, our modeling also shows that these specific interactions are small relative to the independent effects of genomic locations and core promoters (see Figure S2.6). The differences in how genomic locations scale the activities of each core promoter subclass is also largely explained by the average strength of each promoter class (Figure S2.5J). These data show that the observed differences between different promoter classes is a consequence of promoter strength rather than a feature of the hk/dev distinction, indicating that the strength of a promoter is a key determinant of its interactions with the genomic environment.

To further probe how much promoter strength contributes to interactions with the genomic environment, we divided all the promoters into the four bins of expression levels based on loc2. We then fit models with and without interactions between genomic environment and expression bin. As expected, a simple linear model without accounting for interactions explains ~80% of the variance, and the addition of the interaction explains an additional 5% (Figure S2.6A). To test whether there are any specific interactions between landing pads and individual promoters, we added an additional interaction term between the landing pads and promoters. However, this interaction only explains 3% of the residual variance (Figure S2.6A), suggesting there are no specific interactions between landing pads and promoters with large effects.
Figure 2.3: Intrinsic promoter strength explains differences between promoter classes. (A) The effect of genomic location on the expression of weak and strong core promoters. (B) Amount of variance explained by core promoters and genomic locations respectively using linear models fit on weak and strong promoters separately. (C) All pairwise correlations (Pearson’s r) between genomic locations for weak and strong core promoters. (D) Correlation (Pearson’s r) between promoter activity measured on plasmids and promoter activity at loc2.

We also tested whether specific genomic environments restricted the expression of specific promoters, causing them to “drop out” of our analyses at certain landing pad locations. Such interactions would be rare since we recovered measurements for 99% of our promoters at each landing pad (Figure S2.2B). For promoters that did drop out of our analyses, we found that most were only lost from a single landing pad, with only fifteen promoters being lost from two or more landing pads (Figure S2.6B). The loss of these promoters can be explained by their low abundance in the cloned plasmid library, where promoters lost from more landing pads are present at lower levels in the library (Figure S2.6C). Thus, there does not appear to be any systematic or specific restriction of expression by different genomic environments.
Given the importance of the interaction between promoter strength and genomic location, we next asked if core promoter strengths, as measured in the genome, reflect the promoters’ intrinsic activities. If this is true, then the measurements in the genome should correlate with measurements on plasmids, assuming that plasmids represent a neutral environment that reflect the intrinsic activities of core promoters. Thus, we performed an episomal MPRA on the core promoter library in K562 cells. The plasmid measurements are well-correlated with expression at each genomic location (Pearson’s $r^2 = 0.59-0.76$; Figure 2.3D and Figure S2.7A), indicating that the relative intrinsic activities of core promoters are preserved when integrated into the genome. We were also able to predict activity in the genome using activity on plasmids (adjusted $r^2 = 0.72$; Figure S2.7B). These results demonstrate that genomic locations scale the intrinsic activities of strong and weak promoters to different extents, suggesting that the main role of diverse core promoter motifs is to set the intrinsic strength of the promoter rather than direct specific interactions with the genomic environment.

### 2.3.4 Core promoter scaling is a genome-wide phenomenon

To extend the results we observed at four genomic locations to diverse locations across the genome, we selected six core promoters (three hk and three dev) spanning a range of expression levels and motifs within each class (Figure 2.4A, Table S2.3). We note that while hk1 drives constitutive expression of the ribosomal gene RPS27 across multiple cell types, this promoter contains a TATA box, a motif that is generally enriched in housekeeping promoters [130, 247]. We then measured their activities genome-wide using the TRIP (Thousands of Reporters Integrated in Parallel) assay [103] in K562 cells (Figure S2.8A). Each core promoter was cloned upstream of a reporter gene with a unique promoter barcode (pBC) in its 3’UTR into a PiggyBac transposon vector for random delivery into the genome. No upstream proximal enhancer was included in these constructs. TRIP libraries were generated by incorporating $>10^5$ random barcodes (tBCs) onto each core promoter reporter plasmid. After transposition,
every genomic integration contains a unique pBC and tBC pair specifying the identity of the core promoter and its location in the genome respectively. This double barcoding strategy allowed us to pool promoter libraries into a single TRIP experiment. The replicates were highly correlated (Pearson's $r^2 = 0.96$, Figure S2.8B). In total, we mapped 41,083 unique integrations in the genome, ranging between 6078-7418 integrations per promoter (Table S2.3).

Genomic positions have large effects on core promoter activities, with expression ranging more than 1000-fold for the same promoter across genomic locations (Figure 2.4B). However, even with these large effects of genomic location, the rank order of promoter strengths is preserved across locations and correlates with mean expression in the landing pads (Figure 2.4C and Figure S2.8C), which suggests that the effect of different genomic locations is to scale intrinsic promoter activities. Because PiggyBac is known to have a preference for H3K27-acetylated regions [248, 249], we grouped the integrations by their locations into three groups: H3K27ac regions (n=14,275), within 50kb of a H3K27ac region (n=21,623) or far away from H3K27ac regions (n=5185) (Figure S2.8D). Integrations that are far from H3K27ac regions are generally weak, consistent with the idea that these locations are less permissive for expression. However, the rank order of promoters in these regions is the same as integrations in the other locations. Furthermore, integrations in or near H3K27ac regions span the entire >1000-fold dynamic range of our library, suggesting that there is still substantial diversity within H3K27ac regions. Taken together, these data indicate that core promoters are scaled by diverse genomic environments.

To compare different promoters in the same genomic environment, we identified 1278 genomic regions in which at least four of the six promoters had integrated <5kb from each other (in separate cells). These genomic regions are located across the entire genome and span diverse ChromHMM annotations [97, 100] (Figure S2.9A-B). Across these locations, expression consistently increases from the weakest (dev2) to strongest (hk1) promoter (Figure 2.4D-E), showing that the relative strengths of core promoters are preserved across >1000 genomic
Figure 2.4: Core promoter scaling is a genome-wide phenomenon. (A) Features of core promoters selected for TRIP experiments. (B) Expression of each core promoter across all mapped genomic locations sorted by increasing means measured by TRIP. Blue-green denotes hk promoters and pink denotes dev promoters. (C) Correlation (Pearson’s $r$) between mean expression of each core promoter genome-wide (measured by TRIP) and loc1. The shaded region around the fitted line represents the 95% confidence interval. (D) Mean expression of each core promoter from four genomic locations as measured by patchMPRA. Error bars represent the SEM. (E) Heatmap of expression of each core promoter (column) at each genomic region (row) that has $\geq4$ different integrated promoters. White boxes represent NA values.
locations with 1000-fold differences in expression. The expression of the promoters in each region also correlates well with expression in the landing pads, with >60% of locations having $r > 0.7$ (Figure S2.9C), and a linear model assuming independent effects of genomic region and promoter explains ~54% of the variance in the data (Figure S2.9D). Thus, measurements of integrated promoters across diverse genomic positions demonstrates that core promoter scaling is a genome-wide phenomenon.

### 2.3.5 Non-linear scaling of core promoters by genomic environments

We next explored the relationship between core promoter strength and genomic environments in the TRIP data. We ranked the TRIP genomic regions based on mean promoter expression and plotted the expression of the promoters (Figure 2.5A). As expected, all six core promoters increase expression as genomic environments become more permissive. However, the rates at which their expression changes are different for strong and weak promoters. In less permissive regions, strong promoters increase rapidly, but then level off in more permissive regions. In contrast, weak promoters increase slowly in less permissive regions and then sharply in more permissive regions. To ensure that hk1 expression in activating regions is not saturated due to the dynamic range limits of TRIP, we tested hk1 with an upstream enhancer and it was expressed at still higher levels (Figure S2.9E). Thus, promoters with different strengths do not respond to differences in genomic environments in the same way.

In agreement with our results from the patchMPRA experiment above, the curves in Fig. 5A separate by the intrinsic strength of the core promoters and not by their hk or dev identity. To illustrate this point we calculated the correlations between the curves of each promoter and show that the promoters cluster based on their intrinsic strengths, with the stronger promoters (dev1 and hk1) in one cluster and the others in another (Figure S2.10A). Integrations within 5kb of endogenous hk or dev promoters in K562 also showed no preference for hk or dev promoters respectively (Figure S2.10B). This result again highlights that a promoter’s strength, not class,
Figure 2.5: Non-linear scaling of core promoters by genomic environments. (A) Genomic regions defined by TRIP were sorted by the mean expression of the promoters in each region. The shaded region around the fitted line represents the 95% confidence interval. (B) Heatmap in Figure 2.3B split into 3 clusters by k-means clustering. Clusters were assigned different activity levels based on the overall expression in the cluster. (C) Expression of core promoters in each genomic cluster. (D) Amount of variance explained by core promoters and genomic locations respectively using linear models fit on each genomic cluster respectively. (E) Summary model of the relationship between core promoter strength and genomic environment activity.
determines its interaction with genomic environments.

The differences in the way core promoters respond to genomic environments in Figure 2.5A also demonstrate that genomic environments do not scale promoter activities linearly. Although the rank order of core promoters is preserved across the genome, the fold change between strong and weak core promoters is different in different parts of the genome. To quantify the effects of different genomic environments, we identified three clusters of TRIP genomic regions that appear to have different levels of activity (Figure 2.5B). While the clusters are defined by their average differences in core promoter expression, the extent of scaling is also different in each cluster (Figure 2.5C). This difference in scaling is due to differences in the contributions of genomic location and promoter effects in the three clusters. In regions of the genome with low activity, genomic location contributes \(~23\%\) to gene expression while core promoters contribute only \(~12\%\). In the cluster with high activity, genomic location also contributes about \(~24\%\), but core promoters contribute \(~31\%\), suggesting that differences in expression at these locations depend more on core promoter strength. In the cluster with medium activity, the core promoter contribution is much larger, explaining \(~64\%\) of the variance compared to \(~16\%\) by genomic location (Figure 2.5D). Thus, the strength of the genomic environment determines how much it will contribute to gene expression, resulting in non-linear scaling of promoter activities across the genome. This is in contrast with the linear scaling we previously observed using a library of proximal enhancers [108], suggesting that core promoters and proximal enhancers may interact with the genomic environment in different ways.

2.3.6 Genomic clusters have different chromatin states and sequence features

Finally, we asked what features of each cluster distinguish them from each other by overlapping our genomic regions with existing epigenomic datasets and sequence features. Previous studies have shown that reporter genes integrated into the genome tend to take on the chromatin
Figure 2.6: Genomic clusters have different chromatin states and sequence features. (A,B,C) Metaplots of H3K27ac (A), H3K4me3 (B) and PolII (C) levels respectively in each genomic cluster. The start and end marks the boundaries of each genomic region, which are determined by the first and last integration in the region. The x-axis extends +/- 5kb around each genomic region. (D,E,F) Performance of gkmSVM used to classify sequences from high vs medium (D), high vs low (E) and medium vs low (F) genomic clusters. Receiver-operating characteristics (ROC) curves were generated using five-fold cross-validation. (G) The GC fraction of each genomic region was calculated and plotted for each cluster. (H) The of TF binding sites in each genomic region was calculated and plotted for each cluster. p values were calculated by Student’s t-tests.

state of the integration site [190, 250]. In general, cluster activity is correlated with chromatin marks associated with active transcription (H3K27ac, H3K4me3) and transcriptional activity (PolII binding, CAGE-seq) (Figure 2.6A-C, Figure S2.11A), while accessible chromatin (ATAC) and CpG methylation do not separate the clusters (Figure S2.11A-C). This suggests that the three clusters are mainly distinguished by their level of transcriptional activity. We also used sequence features to classify the clusters using gapped k-mer SVMs comparing two clusters at a time [251, 252]. The SVMs performed well, with five-fold cross-validated AUCs ranging
from 0.8 to 0.9 (Figure 2.6D-F and Figure S2.11D-F). Scrambling the cluster annotations led to essentially random predictions by the SVM (Figure S2.11G-H). To further validate the model, we used the trained SVM to predict the cluster type of other TRIP integrations that were not in the 5kb region analysis. As expected, clusters that were predicted to be more active also showed higher expression (Figure S2.11I). To identify the motifs that separate the clusters, we performed de novo motif enrichment and identified CG-rich sequences in the more active clusters (Figure S2.11J-K). Similarly, the CG content of each sequence increases from low to high activity clusters on average (Figure 2.6G). Motif enrichment using known TF position weight matrices did not identify any obvious enriched TF motifs, suggesting that the clusters are not defined by any single TF. However, when we scanned each sequence for known TF motifs, we find that sequences in more active clusters have more TF motifs than less active clusters on average (Figure 2.6H). This result suggests that the differences between clusters is partially explained by the number of TFs binding in each cluster.

### 2.4 Discussion

Gene expression results from the integration of multiple inputs including the core promoter, chromatin environment, distal enhancers and the surrounding transcription factor concentrations. Here we present a framework for dissecting the contributions of core promoters and the surrounding genomic environments to gene expression. Using this framework we found that the intrinsic activities of core promoters are preserved across diverse genomic locations, and are consistent with their activities on plasmids. Contrary to the promoter compatibility hypothesis, hk and dev promoters scale similarly across genomic locations when normalized for differences in strength. These results suggest a general lack of specificity between core promoters and the chromatin landscape/enhancers in their genomic environments, which is consistent with the classical idea of core promoters as passive sequences for the assembly of basal transcriptional machinery. While promoter compatibility has been observed for specific promoter-genomic
environment pairs [111, 112, 114, 115, 130], our results suggest that such interactions are relatively rare or have smaller effects than the effects of genomic scaling. Our results are also consistent with recent work showing that enhancers and promoters are broadly compatible and combine multiplicatively to control gene expression [187]. In this model sequence-specific or protein-specific interactions between core promoters and genomic environments contribute less to gene expression than the independent effects of core promoters and genomic environments. This model suggests a modular genome compatible with the evolution of gene expression by genome rearrangements [253, 254]. In a modular genome, core promoters will function in new genomic locations without having to evolve the machinery for a new set of specific interactions at each location.

Unlike our previous results with cis-regulatory sequences upstream of the core promoter, scaling is not a simple linear combination of genomic position effects and promoter effects [108]. In a linear relationship, the genomic environment scales the activity of local promoters such that the rank order and quantitative differences between promoters are always preserved. This occurs when the contribution of genomic and promoter effects remains constant across genomic locations and promoters. Instead, we find that the quantitative differences between strong and weak promoters change in different genomic environments (Figure 2.5E), suggesting that genomic environments scale core promoter activities in a non-linear manner. Such non-linear scaling is characteristic of the thermodynamics of dose response curves, which follow a sigmoidal relationship with three phases. At low and high input levels, increasing input has little to no effect on output because the output levels are below detection or saturated respectively. However, in the linear range, increases in input have large effects on output. We speculate that different core promoter sequence features set the strength of the promoter, causing them to start at different points of the dose/response curve. This in turn determines how the promoter interacts with the genomic environment.

Our data is also consistent with recent simulations showing how promoters starting from
different states representing different promoter strengths can have different responses to increasing enhancer contact frequency, giving the appearance of enhancer-promoter specificity [255]. This result may also explain the apparent differences between our data and previous results showing enhancer-promoter specificity [114, 130, 132]. We hypothesize that enhancer-promoter specificity is not governed by biochemical differences in transcription factor usage between hk and dev promoters, instead, the observed differences may be due to differences in intrinsic strengths in the hk and dev promoters used. Alternatively, the differences might be due to technical differences between the episomal MPRAs used previously and the genome-integrated assays used here, or to biological differences between Drosophila and human cells. Controlling the intrinsic strengths of different classes of promoters will be important for testing enhancer-promoter specificity. In the future, the non-linear relationship between promoter strength and genomic position effects will help us to predict gene expression by measuring core promoter strength and genomic environment activity independently.

2.5 Methods

Library design

We obtained a set of 6916 core promoter sequences from Haberle et al. [30] and selected 672 sequences for our library. Each promoter is 133bp long and centered on the major transcription start site (TSS). We selected the sequences to contain diverse core promoter types and expression patterns (Table S2.1) using the designations obtained from Haberle et al.. We also included the super core promoter (SCP1), as well as versions of SCP1 with TATA and DPE single and double mutants [246]. A library of oligonucleotides (oligos) encoding the selected core promoters was synthesized by Agilent technologies through a limited licensing agreement. Each oligo in the library is 200bp and contains a core promoter, a unique barcode that specifies the identity of the promoter, and flanking sequences for subcloning. An example oligo is shown here:
CCTTACACGGAGTGGATA-SpeI-core promoter-HindIII-NheI-XbaI-12bp barcode-SalI-CATAA CTTTCGTATAATGT

Each promoter is present ten times in the library, each time with a different unique barcode, to provide redundancy in the measurements. In total the oligo pool contains 6760 unique sequences. The barcodes were randomly selected from barcode lists generated by the FREE barcodes software [256].

**patchMPRA library cloning**

The sequences and description of all oligos and primers used in this study can be found in Table S2.4.

We selected a single plasmid from a previous patchMPRA library [108] to serve as the backbone of our promoter library. This plasmid contains a single enhancer and drove robust expression in a previous experiment. The enhancer contains motifs for Fos/Jun and MAF transcription factors and its full sequence is TGCCCCCCTTCTTCTATGTCTGATGGAGTTTCCTCTCTAAGTAGCCATTTTTATTCTGCTGACTCACCCTCTAACTCCCGGTCTTATTCCATCCTGCCTCAGGGTCTGTGGTGTAGTCATAGCAC.

To create our library, we first removed the hsp68-dsRed construct from the selected plasmid with HindIII and XhoI. We then amplified the oligo pool using primers CPL1 and CPL2 and inserted it into the digested backbone using HiFi DNA Assembly (New England Biolabs). Next, we digested the library with HindIII and XbaI and inserted an mScarlet fluorophore between the promoter and barcodes. To test the library without an upstream enhancer we also cloned the library into a vector backbone that does not contain an enhancer. The backbone was digested with SpeI and XhoI, and the oligo pool was amplified with primers CPL2 and CPL3. The fragments were assembled using HiFi DNA Assembly (New England Biolabs), and the mScarlet fluorophore was inserted in the same way as described above.
We replaced the HygTK-GFP cassette in the original landing pad cell lines from Maricque et al. [108] with a reporter expressing both HSV-TK (herpes simplex virus thymidine kinase) and the monomeric blue fluorescent protein tagBFP. The new cassette contains a functional HSV-TK gene, allowing for negative selection of cells that do not have a library member integrated.

K562 cells were maintained in Iscove’s Modified Dulbecco’s Medium (IMDM) + 10% FBS + 1% non-essential amino acids + 1% penicillin/streptomycin. To integrate the library into the genome, we co-transfected the library and CRE recombinase (pBS185 CMV-Cre, Addgene 11916) into 4 K562 ‘landing pad’ cell lines expressing the HSV-TK gene (landing pad details in Table S2.2). For each replicate, we transfected 32μg library with 32μg CRE recombinase into 9.6 million total cells using the Neon Transfection System (Life Technologies). We performed three separate transfections representing three biological replicates. After three days, we treated the cells with 2mM ganciclovir to kill the cells that did not successfully integrate a library element. Cells were treated every day for four days. We then selected for live cells using the MACS Dead Cell Removal Kit (Miltenyi Biotech) and the cells were allowed to grow until there were sufficient cells for DNA/RNA extractions (about 10 million cells).

DNA and RNA was harvested from the cells using the TRIzol reagent (Life Technologies). The RNA was treated with two rounds of DNase using the Rigorous DNase treatment procedure in the Turbo DNase protocol (Ambion), and cDNA was synthesized with Oligo(dT) primers using the SuperScript IV First Strand Synthesis System (Invitrogen). The barcodes were then amplified from the cDNA and genomic DNA (gDNA) using the Q5 High Fidelity 2X Master Mix (New England Biolabs) with primers specific to our reporter gene (CPL4-5). We performed thirty-two PCR reactions per cDNA biological replicate and 48 PCR reactions per gDNA biological replicate, then pooled the PCRs of each replicate for PCR purification. 4ng from each replicate was then further amplified with two rounds of PCR to add Illumina sequencing adapters (CPL6-9). Barcodes were sequenced on the Illumina NextSeq platform.
**Episomal MPRA**

We first digested the patchMPRA library with HindIII and XbaI to replace the mScarlet fluorophore with a tdTomato between the promoter and pBC. We then subcloned the promoter library with the tdTomato fluorophore into the landing pad lentiviral vector from Maricque et al. [108] to ensure that the 3’UTR from the episomal library matches that of the patchMPRA experiment. Briefly, the lentiviral vector and patchMPRA library were digested with XhoI/SpeI and NheI/SalI respectively and the library was ligated into the lentiviral backbone with T4 DNA ligase (New England Biolabs).

For the MPRA, we transfected the library into K562 cells using the Neon Transfection System (Life Technologies). We performed two biological replicates, transfecting 2.4 million cells with 10μg of library per replicate. After 24h, we harvested RNA from the cells using the PureLink RNA Mini Kit (Invitrogen). The RNA was treated with DNase and converted to cDNA in the same way as the patchMPRA library above. We then amplified barcodes from cDNA using primers CPL5 and CPL10 with the Q5 High Fidelity 2X Master Mix (New England Biolabs). We performed four PCR reactions per replicate from cDNA. For DNA normalization, we performed the same PCR (two PCR reactions per replicate; two replicates) on the plasmid library. The PCRs from the same replicates were then pooled and purified. 4ng from each replicate was then further amplified with two rounds of PCR to add Illumina sequencing adapters (CPL6-9). Barcodes were sequenced on the Illumina NextSeq platform.

**TRIP library cloning**

We performed TRIP according to the published protocol [103] with some modifications. We first digested the PiggyBac vector (PBSsplitGFP, gift from Robi Mitra lab) [257] with BamHI and NotI. Each selected promoter was amplified from the promoter library (CPL11-22) and assembled into the vector with a tdTomato fluorophore and the neuropilin 1 poly(A) sequence [103] using HiFi DNA Assembly (New England Biolabs). We then added a unique barcode
that identifies the promoter (pBC) to each promoter construct using the Q5 Site-Directed
Mutagenesis Kit (New England Biolabs). A second random barcode was added to each promoter
construct by digesting with XbaI followed by HiFi DNA Assembly (New England Biolabs) with
a single-stranded oligo containing sixteen random N’s (TRIP barcodes; tBC) and homology arms
to the plasmid (CPL23). The components of the final library are shown in Figure S2.8A. The
PiggyBac ITRs, promoter and tdTomato reporter cassette is located between two parts of a split-
GFP reporter gene which is driven by a separate EF1a promoter. When the barcoded reporter
cassette is integrated into the genome, the split-GFP remaining on the plasmid combines to
produce functional GFP, allowing us to sort for cells that have successfully integrated the
promoters Figure S2.8A. Since each promoter is uniquely barcoded, we combined all the
promoters into a single library for subsequent TRIP experiments.

TRIP

The TRIP library and piggyBac transposase (gift from Robi Mitra lab) were co-transfected
into wild-type K562 cells at a 1:1 ratio using the Neon Transfection System (Life Technologies).
In total, we transfected 4.8 million cells with 16μg each of library and transposase. The cells
were sorted after 24 hours for GFP-positive cells to enrich for cells that have integrated the
reporters. After a week, the cells were sorted into four pools of 7000 cells each to ensure that
each pBC-tBC pair is only integrated once in each pool. The pools were then allowed to grow
until there were sufficient cells for DNA/RNA extractions.

We harvested DNA and RNA from the cells using the TRIzol reagent (Life Technologies).
The RNA was treated with DNase and converted to cDNA in the same way as the patchMPRA
library above. We then amplified barcodes from cDNA and gDNA using primers CPL10 and
CPL24. We performed 4 PCRs per pool from cDNA and gDNA respectively using the Q5 High
Fidelity 2X Master Mix (New England Biolabs), then pooled the PCRs and purified them.
4ng from each replicate was then further amplified with two rounds of PCR to add Illumina
sequencing adapters (CPL25-26, CPL8-9). Barcodes were sequenced on the Illumina NextSeq platform.

To map the locations of TRIP integrations, we digested gDNA with a combination of AvrII, NheI, SpeI and XbaI for 16 hours. The digestions were purified and self-ligated at 4°C for another 16 hours. After purifying the ligations, we performed inverse PCR to amplify the barcodes with the associated genomic DNA region (primers CPL24-25). We did eight PCRs per pool, purified them and used 4ng of each pool for a further two rounds of PCR to add Illumina sequencing adapters (CPL29-31, CPL9). The library was then sequenced on the Illumina NextSeq platform.

**patchMPRA and episomal MPRA data processing**

For patchMPRA, we obtained approximately 11-13 million reads per DNA or RNA replicate from sequencing. For episomal MPRA, we obtained approximately 500,000 reads per DNA or RNA replicate. Reads that contained the barcodes in the proper sequence context were included in subsequent analysis. The pBCs were then decoded using the FREE barcodes software [256] and the expression of each barcode pair was calculated as \( \log_2(\text{RNA/DNA}) \). We averaged the expression of barcodes corresponding to the same promoter within each replicate to get promoter expression per replicate, then averaged across replicates for subsequent downstream analysis.

**TRIP data processing**

We obtained approximately 14-25 million reads per DNA or RNA pool from sequencing. Reads that contained both the tBC and pBC in the proper sequence context were included in subsequent analysis. We further filtered tBCs such that they are at least 3 hamming distance apart from every other barcode to account for mutations that occurred during PCR and sequencing. The expression of each BC pair was calculated as \( \log_2(\text{RNA/DNA}) \). We added a
pseudocount to the RNA counts to include barcode pairs that had DNA but no RNA reads. Data from the four independent pools were combined in all analyses.

For the locations of TRIP integrations, reads containing each barcode pair were matched with the sequence of its integration site. The integration site sequences were then aligned to hg38 using BWA with default parameters. Only barcodes that mapped to a unique location were kept for downstream analyses.

**TRIP data analysis**

We downloaded a list of expressed genes in K562 cells using whole cell long poly(A) RNA-seq data generated by ENCODE [258] from the EMBL-EBI Expression Atlas. We then designated the genes as hk or dev based on the list of hk genes obtained from Eisenberg and Levanon [259]. Using the locations of these promoters (GENCODE Release 36, GRCh38.p13) we identified TRIP integrations located within 5kb of either hk or dev promoters and plotted the expression of these integrations separately.

To increase the resolution of the analysis we identified genomic regions where at least four different promoters integrated within 5kb of each other. For regions in which the same promoter integrated more than once we used the median expression of that promoter. This yielded 1268 genomic regions. All heatmaps were generated using the ComplexHeatmap package in R [260]. To determine the diversity of the identified 5kb regions, we downloaded the 15-state segmentation for K562 (hg19) from the ENCODE portal and converted the genomic coordinates to hg38 using the UCSC liftOver tool [261]. We then overlapped the 5kb regions with ChromHMM regions using a minimum overlap of 200bp using the Genomic Ranges R packages [262].

To rank and cluster the regions we first imputed missing values using the mean of the promoter across all locations. We then used the means of each region to rank the clusters and plotted the smoothed expression of each promoter. To cluster the 5kb genomic regions, we ran
k-means clustering on the imputed data using the ConsensusClusterPlus package in R [263]. The imputed data was only used for ranking and clustering and not downstream analysis.

**Epigenome data analysis**

For the cluster metaplots, we considered the boundaries of each genomic region as the locations of the first and last integrations in each region. We then downloaded various K562 epigenome datasets (Table S2.5). For CpG methylation, we downloaded both replicates and used the averaged signal from both replicates. For H3K27ac, H3K4me3, PolII, CpG methylation and ATAC-seq, we used the EnrichedHeatmap package in R [260] to draw the metaplots for each cluster extending 5kb upstream and downstream of each genomic region. For CAGE-seq, we downloaded the hg19 dataset from the FANTOM5 consortium [264, 265] and converted it to hg38 using the UCSC liftOver tool [261]. Because the signal was relatively sparse across genomic locations, we plotted the total CAGE signal across each genomic region.

**Sequence features analysis**

We obtained the sequences of each region using the BSgenome package in R [266]. For the gapped k-mer predictions, we used the gkmSVM R package [252] with word length = 10 and number of informative columns = 6. We used AME for motif enrichment analysis [267], DREME for de novo motif discovery [268] and FIMO to determine the number of motifs per sequence [269], from MEME suite 5.0.4. For all motif analyses we limited analysis to expressed transcription factors (FPKM ≥ 1) in K562 from whole cell long poly(A) RNA-seq data generated by ENCODE [258] downloaded from the EMBL-EBI Expression Atlas.

To predict the type of genomic region of other integrations not in the defined 5kb regions, we obtained genomic sequences of the 1kb flanking region around the integration (500bp upstream and 500bp downstream). We then used the trained gkmSVM kernels to calculate the weights of each flanking region and assigned the integrations into low, medium or high
activity clusters based on their weights. Only integrations that could be confidently assigned were included.

**Modeling**

We fit log2 expression values with linear models of core promoter and genomic location activities using the `lm` function in R. Variance explained by each term was calculated with one-way ANOVAs of the respective models.

**Data Access**

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO) under accession number GSE173678.

**2.6 Supplementary Figures**

![Supplementary figure](image)

**Figure S2.1:** Landing pad locations have diverse chromatin marks and transcriptional activity. Each section is 10kb surrounding the location of landing pad integration.
Figure S2.2: patchMPRA measurements are robust and reproducible. (A) Number of promoter barcodes recovered from each location per biological replicate. (B) Number of promoters recovered from each location per biological replicate. (C,D) Reproducibility of core promoter measurements from independent patchMPRA transfections for Replicates 1 & 3 (C) and Replicates 2 & 3 (D). (E) Expression of genes that the landing pads are located in (locs 1-3) and closest promoter (loc4) as measured by RNA-seq.
Figure S2.3: Effect of genomic locations on core promoters. (A) Expression of a library of proximal enhancers at each genomic location [108]. (B) All pairwise correlations (Pearson's r) between genomic locations for core promoters with different motifs within each class.
Figure S2.4: Core promoter scaling is not dependent on the proximal enhancer. (A) Pairwise correlations (Pearson’s $r$) of core promoter activity between the different genomic locations after removing the proximal enhancer upstream of the core promoter. (B) Expression of $hk$ and $dev$ promoters at each genomic location without the proximal enhancer. (C) All pairwise correlations (Pearson’s $r$) between genomic locations for $hk$ and $dev$ core promoters without the proximal enhancer. (D) Correlation between expression with and without the proximal enhancer at locations 1-3.
Figure S2.5: Intrinsic promoter strength explains differences between classes of core promoters. (A) Expression of all hk and dev promoters at each genomic location. *p*-values were calculated by Student’s *t*-tests. (B) The effect of genomic location on the expression of weak and strong core promoters after removing the proximal enhancer upstream of the promoter.

(legend continued on next page)
(Figure S2.5 continued) (C) All pairwise correlations between genomic locations for weak and strong core promoters without the proximal enhancer. (D,E,F). Expression of hk and dev promoters at each genomic location after sampling promoters such that the two classes have equivalent strong (D), intermediate (E) and weak (F) strengths respectively. n indicates number of promoters sampled from each class for each subsample. (G,H,I) All pairwise correlations between genomic locations for subsampled strong (G), intermediate (H) and weak (I) hk and dev core promoters respectively. Locations with low correlations in I are highlighted. (J) The pairwise correlations of core promoters with different motifs (from Figure S2.2) are explained by the average expression of each group.
**Figure S2.6:** Specific interactions between landing pads and promoters are rare. (A) Variance explained by each term in models with or without interactions between genomic locations and expression quartile or promoters. (B) Number of promoters that were not detected in one, two, three or four landing pads respectively. (C) Average abundance of promoters in plasmid library that were not detected in zero, one, two, three or four landing pads respectively.
Figure S2.7: Core promoter activities in the genome reflect the promoters’ intrinsic activity. (A) Correlations between expression of core promoter library measured on plasmids and at the indicated genomic location by patchMPRA. (B) Correlation between measured expression by patchMPRA and predicted expression by a linear model using core promoter intrinsic activity measured on plasmids. All correlations were calculated as Pearson’s $r$. 
Figure S2.8: Measurements of six core promoters at thousands of genomic locations by TRIP. (A) Schematic of TRIP experiment. tBC: TRIP barcode; pBC: promoter barcode; ITR: inverted terminal repeats. (legend continued on next page)
(Figure S2.8 continued) (B) Reproducibility between measurements from independent DNA and RNA extractions. (C) Correlations between mean expression of core promoters measured by TRIP and at the indicated genomic location by patchMPRA. (D) Expression of core promoters at locations that overlap with H3K27ac sites, are within 50kb of H3K27ac sites or are outside 50kb of H3K27ac sites. p-values were calculated by Student’s t-tests. n.s.: not significant, * p<0.05, ** p<0.001, *** p<0.005.
Figure S2.9: Core promoter scaling is a genome-wide phenomenon. (A) Regions with ≥4 different promoters integrated within 5kb of each other are located across the genome. Cluster activity was designated by the analysis in Figure 2.5B. (B) Distribution of chromHMM annotations of defined 5kb regions. (C) For each defined 5kb region, correlations (Pearson’s r) between core promoter activity measured by TRIP and by patchMPRA were calculated and all correlations were plotted as a density plot. As a comparison, we randomly grouped promoters without considering their integration locations and calculated the correlations for each group. The p-value was calculated using the Mann–Whitney U test. (D) Correlation between measured expression by TRIP and predicted expression using a model assuming independence between genomic environments and core promoters. (E) Expression of all integrations of hk1 and hk1 with an upstream cis-regulatory enhancer (CRE-hk1).
Figure S2.10: Promoter strength, not class, determines its interaction with the genomic environment. (A) Correlation coefficients between curves fitted on each promoter in Figure 2.5A. (B) Hk and dev integrated core promoters behave similarly near endogenous hk or dev promoters.
Figure S2.11: Epigenomic signatures and sequence features of different genomic activity clusters. (A) CAGE-seq signal was calculated for each genomic region, and the summed signals were plotted for each cluster. *p*-values were calculated by Student’s *t*-tests.

(legend continued on next page)
(Figure S2.11 continued) (B,C) Metaplots of CpG methylation (B) and ATAC-seq (C) signals in each genomic cluster. The start and end mark the boundaries of each genomic region, which are determined by the first and last integration in the region. The x-axis extends +/- 5kb around each genomic region. (D,E,F) Performance of gkmSVM used to classify sequences from the high vs low (D), high vs medium (E) and medium vs low (F) genomic clusters. Precision-recall curves (PRCs) were generated using five-fold cross-validation. (G,H) ROC (G) and PRC (H) of gkmSVM on sequences with scrambled cluster assignments. (I) TRIP integrations that were not included in the 5kb genomic region analysis were assigned to a cluster based on their sequence features from the gkmSVM, and the expression of each promoter was plotted based on their predicted clusters. (J,K) Top 6 motifs identified by de novo motif finding comparing high vs low (J) and medium vs low (K) activity sequences respectively.
2.7 Supplementary Tables

Table S2.1: Composition of promoter classes in the core promoter library.

<table>
<thead>
<tr>
<th>Promoter class</th>
<th>hk/dev</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATA-box dev</td>
<td>dev</td>
<td>102</td>
</tr>
<tr>
<td>CpG island</td>
<td>hk</td>
<td>102</td>
</tr>
<tr>
<td>CpG island dev</td>
<td>dev</td>
<td>102</td>
</tr>
<tr>
<td>TCT</td>
<td>hk</td>
<td>2</td>
</tr>
<tr>
<td>DPE</td>
<td>hk</td>
<td>8</td>
</tr>
<tr>
<td>DPE dev</td>
<td>dev</td>
<td>13</td>
</tr>
<tr>
<td>No known motif</td>
<td>hk</td>
<td>102</td>
</tr>
<tr>
<td>No known motif dev</td>
<td>dev</td>
<td>102</td>
</tr>
<tr>
<td>TATA-box &amp; CpG island dev</td>
<td>dev</td>
<td>62</td>
</tr>
<tr>
<td>TATA-box &amp; DPE</td>
<td>dev</td>
<td>2</td>
</tr>
<tr>
<td>DPE &amp; CpG island</td>
<td>dev</td>
<td>13</td>
</tr>
<tr>
<td>DPE &amp; CpG island</td>
<td>hk</td>
<td>40</td>
</tr>
<tr>
<td>TCT &amp; CpG island</td>
<td>hk</td>
<td>22</td>
</tr>
<tr>
<td>SCP (and mutants)</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>676</strong></td>
</tr>
</tbody>
</table>

Table S2.2: Locations of four landing pads in patchMPRA.

<table>
<thead>
<tr>
<th>Name</th>
<th>Chr</th>
<th>Location (hg19)</th>
<th>Annotation</th>
<th>Name in Maricque et al. [108]</th>
</tr>
</thead>
<tbody>
<tr>
<td>loc1</td>
<td>chr11</td>
<td>16,258,750</td>
<td>Sox6 Intron</td>
<td>LP3</td>
</tr>
<tr>
<td>loc2</td>
<td>chr16</td>
<td>53,275,015</td>
<td>CHD9 Intron</td>
<td>LP4</td>
</tr>
<tr>
<td>loc3</td>
<td>chr17</td>
<td>56,426,171</td>
<td>Intergenic</td>
<td>LP5</td>
</tr>
<tr>
<td>loc4</td>
<td>chr1</td>
<td>156,489,766</td>
<td>Intergenic</td>
<td>LP6</td>
</tr>
<tr>
<td>Name</td>
<td>Oligo_id</td>
<td>Number of integrations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>----------</td>
<td>------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hk1</td>
<td>chr1_153963171_153963304_+</td>
<td>6032</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hk2</td>
<td>chr7_94285368_94285501_-</td>
<td>7157</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hk3</td>
<td>chr1_19638753_19638886_+</td>
<td>6851</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dev1</td>
<td>chr17_5522677_5522810_-</td>
<td>6328</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dev2</td>
<td>chr21_33976756_33976889_-</td>
<td>7079</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dev3</td>
<td>chr1_1009619_1009752_-</td>
<td>7364</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Table S2.4:** Primers used in this study.

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPL1</td>
<td>TCTCCTCCGGCTCGCTGATTATTACTTACACG-GAGTGGGATA</td>
<td>Amplify oligo library for sub-cloning</td>
</tr>
<tr>
<td>CPL2</td>
<td>TACATTATACGAAGTTATGTCGAC</td>
<td>Amplify oligo library for sub-cloning</td>
</tr>
<tr>
<td>CPL3</td>
<td>CCTTACACGGAGTTGGA</td>
<td>Amplify oligo library for sub-cloning</td>
</tr>
<tr>
<td>CPL4</td>
<td>CCCCGTAATGCAGAAGAAGA</td>
<td>Amplify barcodes from integrated promoter library for Illumina sequencing.</td>
</tr>
<tr>
<td>CPL5</td>
<td>GCAGCGTATCCACATAGCGTAAAAG</td>
<td>Amplify barcodes from integrated promoter library for Illumina sequencing.</td>
</tr>
<tr>
<td>CPL6</td>
<td>CTTTCCCTACACGACGCTTTCCGATCT(N[1-4])CATGGACGAGCTGTACAGAATCTAGA</td>
<td>Add first round of adapters to promoter library amplified barcodes. Variable numbers of Ns included to phase the library for Illumina sequencing.</td>
</tr>
<tr>
<td>CPL7</td>
<td>GTGACTGGAGTTGACGCTGTGCTTTCCGATCT(N[0-3])GCAGCGGCTTTAGGATCC</td>
<td>Add first round of adapters to promoter library amplified barcodes. Variable numbers of Ns included to phase the library for Illumina sequencing.</td>
</tr>
<tr>
<td>CPL8</td>
<td>AATGATAACGGCGACCACCCAGATCTACAC-GNNNNNACACTCTTTCCCTACACGAGCT</td>
<td>Add second round of adapters to promoter library amplified barcodes. N’s indicated variable sequences for indexing.</td>
</tr>
</tbody>
</table>
Table S2.4: Primers used in this study. (continued)

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPL9</td>
<td>CAAGCAGAAGACGGCATACGAGATNNNNNNNGTGACTGGAGTTCA-GACGTG</td>
<td>Add second round of adapters to amplified barcodes. N’s indicated variable sequences for indexing.</td>
</tr>
<tr>
<td>CPL10</td>
<td>ACCATCTACATGGCCAAGGAAGC</td>
<td>Amplify barcodes from epiposomal and TRIP library for Illumina sequencing.</td>
</tr>
<tr>
<td>CPL11</td>
<td>ATATCAGGCAGCAGCTGTGATCCGCACAA-GATCCTTGCGTC</td>
<td>Amplify hk1 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
<tr>
<td>CPL12</td>
<td>TCCTCGGCCCTTGCTACCATCCTAGGGCC-AAAACCTGGACAAAA</td>
<td>Amplify hk1 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
<tr>
<td>CPL13</td>
<td>ATATCAGGCAGCAGCTGTGATCCGGGAT-GCTGATGCTGAATCGA</td>
<td>Amplify hk2 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
<tr>
<td>CPL14</td>
<td>TCCTCGGCCCTTGCTACCATCCTAGGAG-CACAGGGGTCTCCCAG</td>
<td>Amplify hk2 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
<tr>
<td>CPL15</td>
<td>ATATCAGGCAGCAGCTGTGATCGCGGCGGGCCTGCGGTTC</td>
<td>Amplify hk3 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
<tr>
<td>CPL16</td>
<td>TCCTCGGCCCTTGCTACCATCCTAGGTT-GAAGTTTGCCACCTGCGGTTC</td>
<td>Amplify hk3 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
<tr>
<td>CPL17</td>
<td>ATATCAGGCAGCAGCTGTGATCCCTGC-GATACTGGAGTGG</td>
<td>Amplify dev1 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
<tr>
<td>CPL18</td>
<td>TCCTCGGCCCTTGCTACCATCCTAGGGAAGCT-GACAATCACGAGC</td>
<td>Amplify dev1 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
</tbody>
</table>
### Table S2.4: Primers used in this study. (continued)

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPL19</td>
<td>ATATCAGGCCTAAGGGTCTCATCTATCTATC-CCGATCCTCACTGCCA</td>
<td>Amplify dev2 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
<tr>
<td>CPL20</td>
<td>TCCTCGCCCTTTGCTCACCATCCTAGGAGCA-CACGGTTGCGCCCTGTA</td>
<td>Amplify dev2 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
<tr>
<td>CPL21</td>
<td>ATATCAGGGCCTAAGGGTCTCATCTATCTATC-CCGATCCTCACTGCCA</td>
<td>Amplify dev3 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
<tr>
<td>CPL22</td>
<td>TCCTCGCCCTTTGCTCACCATCCTAGGAGCA-CACGGTTGCGCCCTGTA</td>
<td>Amplify dev3 from promoter library with homology to backbone for HiFi assembly.</td>
</tr>
<tr>
<td>CPL23</td>
<td>GCTCTATAAGTAAGAGCTCTCCTCGGAGTCATCCTTGGATCCTCACTGCCA</td>
<td>Oligo for adding random barcodes to TRIP library by HiFi assembly.</td>
</tr>
<tr>
<td>CPL24</td>
<td>AACGCCAGGGTTTTCCCA</td>
<td>Amplify barcodes from TRIP library for Illumina sequencing.</td>
</tr>
<tr>
<td>CPL25</td>
<td>CTTTCCCTACAGGACGCTCTCTCTGATCTCTCTCAGCATCTATCCTATCTATC-CCGATCCTCAGTTAGCAGTCTAGA</td>
<td>Add first round of adapters to amplified TRIP barcodes. Variable numbers of Ns included to phase the library for Illumina sequencing.</td>
</tr>
<tr>
<td>CPL26</td>
<td>GTGACTGGAGTTCAGACGCTCTGATCTCAGAGGAGGTTTCCCAAC</td>
<td>Add first round of adapters to amplified TRIP barcodes.</td>
</tr>
<tr>
<td>CPL27</td>
<td>CGCATGATTATCTTTAACGCTACGTCA</td>
<td>Amplify TRIP barcode and associated genomic region by inverse PCR.</td>
</tr>
<tr>
<td>CPL28</td>
<td>GCCAGGGTTTTCCCAAC</td>
<td>Amplify TRIP barcode and associated genomic region by inverse PCR.</td>
</tr>
</tbody>
</table>
Table S2.4: Primers used in this study. (continued)

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPL29</td>
<td>ACGACGCTCTTCCGATCTGCTCGAT(N[0-3])GTACGTCACAATATGATTATCTTTCTAG</td>
<td>Add first round of adapters to TRIP amplified barcodes. Variable numbers of Ns included to phase the library for Illumina sequencing.</td>
</tr>
<tr>
<td>CPL30</td>
<td>GTGACTGGAGTTCAGACGTGTGCTCTTCC-GATCTGCCAGGGTTTTCCCAAC</td>
<td>Add first round of adapters to TRIP amplified barcodes. Variable numbers of Ns included to phase the library for Illumina sequencing.</td>
</tr>
<tr>
<td>CPL31</td>
<td>AATGATACGGCGACCACCGAGATCTA-CACTCTTTCCCTACACGACGCTCTTCCGATCT</td>
<td>Add second round of adapters to TRIP amplified barcodes. N’s indicated variable sequences for indexing.</td>
</tr>
</tbody>
</table>
## Table S2.5: Sources of epigenome datasets used in this study.

<table>
<thead>
<tr>
<th>Data</th>
<th>Source Experiment</th>
<th>Source File</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K27ac ChIP-seq</td>
<td>ENCSR000AKP</td>
<td>ENCF437DPT</td>
</tr>
<tr>
<td>H3K4me3 ChIP-seq</td>
<td>ENCSR000EWA</td>
<td>ENCF916MPM</td>
</tr>
<tr>
<td>PolII ChIP-seq</td>
<td>ENCSR388QZF</td>
<td>ENCF285MBX</td>
</tr>
<tr>
<td>CpG methylation</td>
<td>ENCSR765JPC</td>
<td>ENCF867JRG; ENCF721JMB</td>
</tr>
<tr>
<td>ATAC-seq</td>
<td>ENCSR868FGK</td>
<td>ENCF698MIQ</td>
</tr>
</tbody>
</table>
| CAGE            | FANTOM5           | chronic%20myelogenous%20leukemia%20cell%20line%3aK562.
|                 |                   | CNhs11250.10454-106G4.hg19.ctss.bed.gz                                      |
Chapter 3: Massively parallel characterization of insulator activity across the genome

This work was performed in collaboration with Alyssa Erickson, Jie Li, Arnold Federico and Barak Cohen. I conceived and designed the project. I also did most of the experiments and data analysis, and made all of the figures. Barak Cohen and I wrote the manuscript. This work is currently being prepared for submission to a journal.

3.1 Abstract

Insulators are cis-regulatory sequences (CRSs) that can block enhancers from activating target promoters or act as barriers to the spread of heterochromatin. Their name derives from their ability to “insulate” transgenes from genomic position effects, an important function in gene therapy and biotechnology applications that require high levels of sustained transgene expression. In theory, flanking transgenes with insulators protects them from position effects, but in practice, efforts to insulate transgenes meet with mixed success because the contextual requirements for insulator function in the genome are not well understood. A key question is whether insulators are modular elements that can function anywhere in the genome or whether they are adapted to function only in certain genomic locations. To distinguish between these two possibilities we developed MPIRE (Massively Parallel Integrated Regulatory Elements) and used it to measure the effects of three insulators (A2, cHS4, ALOXE3) and their mutants at thousands of locations across the genome. Our results show that each insulator functions in only a small number of genomic locations, and that these locations depend on the sequence motifs that comprise each insulator. All three insulators can block enhancers in the genome,
but only ALOXE3 can act as a heterochromatin barrier. We conclude that insulator function is highly context dependent and that MIRE is a robust and systematic method for revealing the context dependencies of insulators and other cis-regulatory elements across the genome.

### 3.2 Introduction

A consistent model of insulator function has yet to emerge from one-at-a-time perturbations of specific insulator elements. The effects of perturbations to putative insulators on genome structure and gene expression appear to depend on the specific locus and regulatory sequences involved [91]. If the primary role of insulators is to organize genome structure [90], then these context-specific effects may arise from how insulators facilitate or restrict access to cis-regulatory sequences. Yet, genome-wide depletion of known insulator binding proteins CCCTC-binding factor (CTCF) and Cohesin leads to widespread changes in genome structure but have little impact on gene expression [151, 152], suggesting that insulators may not act solely through genome organization. It is therefore important to directly assess the impact of insulators on gene expression across diverse genomic contexts.

Two classes of insulator activities have been described. Barrier insulators halt the spread of flanking heterochromatin, while enhancer-blocking insulators prevent enhancers from activating their target promoters when placed between the enhancer and promoter [90]. Some insulators can protect transgenes from chromosome position effects and transgene silencing [53, 56, 179], suggesting that insulators can also block trapped enhancers at ectopic locations. However, not all enhancer-blocking insulators protect against position effects [270]. Enhancer-blocking assays are typically performed with one enhancer-promoter pair on plasmids, but some insulators only block specific enhancer-promoter pairs [60], and insulator activity in the genome may be highly dependent on genomic context [271, 272]. It remains unclear whether insulators are modular elements that can block position effects across all locations or whether the activities of insulators are specific for certain locations in the genome.
The sequence features that contribute to insulator activity are not well understood. In mammalian cells, the most well-characterized protein with insulation activity is CTCF. Yet, in the extensively characterized cHS4 insulator element, CTCF is required for enhancer-blocking activity on plasmids, but not for protection against position effects in the genome [87, 92, 273] and CTCF is neither necessary nor sufficient to block the spread of repressive H3K27me3 heterochromatin [79, 151, 176, 271]. When different genomic sequences with strong matches to the CTCF binding site are tested in the plasmid-based assay they display a wide range of activities [71], and some insulator sequences do not require CTCF at all. RNA pol III transcribed elements, including short interspersed nuclear elements (SINEs) and tDNA sequences, can possess both enhancer-blocking and barrier activity [70, 88, 274]. While B box motifs are necessary for both enhancer-blocking and barrier activity in one tRNA insulator, the same motif was insufficient for barrier activity in a different assay [70, 274]. Instead, AT content was important for sustained barrier activity[274]. Given our current understanding, we cannot predict the locations of insulator elements in the genome from their sequence motifs, nor can we reliably predict the effects of genetic variants in insulator sequences on either genome structure or gene expression. Systematic studies of insulator elements and their genetic variants will be required to define how the critical sequences that comprise insulator function across the genome.

A major challenge to studying insulators is that their functions are intimately tied to their locations in the genome. Enhancer-blocking assays can be performed on plasmids, but the results of such experiments do not necessarily correlate with insulator function in the genome. We require a technology to systematically assay the impact of insulators at thousands of locations across the genome. Moreover, to reliably compare different insulators, each insulator sequence must be assayed at the same set of locations in the genome. Such a technology would allow investigators to determine the sequence features of insulator sequences that permit them to function in different epigenetic landscapes.
Massively Parallel Reporter Gene Assays (MPRAs) are a family of approaches that allow investigators to measure the activities of cis-regulatory sequences en masse, but these techniques generally rely on plasmid-based reporter genes [181–183]. Some variants of MPRAs allow investigators to measure the activity of a reporter gene integrated at a large number of genomic locations, but because these techniques rely on the random integration of reporter genes [103, 275, 276], different insulator sequences cannot be directly compared to each other at the same locations [272]. Using recombinase mediated cassette exchange (RCME), some versions of MPRAs let investigators recurrently integrate reporter gene libraries at a small number of defined genomic locations [108, 243]. Walters et al. used a similar approach to assay the ability of the cHS4 insulator to block HS2 enhancer mediated activation of the γ-globin promoter at three genomic locations in K562 cells and found that cHS4 activity was exquisitely context-specific. However, gaining the power to unravel the sequence features of insulators that allow them to function in different genomic environments will require a technology that can recurrently assay different insulator variants in a large number of defined genomic locations.

To address this problem we developed Massively Parallel Integrated Regulatory Elements (MPIRE), a method that measures the effects of different insulator sequences at thousands of defined genomic locations simultaneously. Using MPIRE, we present the largest characterization of three insulator sequences (cHS4, A2 and ALOXE3) at more than 10,000 locations across the genome. Since insulator activity is thought to depend on either CTCF binding sites (cHS4 and A2) or B box motifs (ALOXE3), we also measured the effects of mutations in the respective motifs at the same genomic locations. We show that each insulator functions at specific and distinct locations in the genome, and that this activity is largely dependent on CTCF or the B box motif respectively. While all insulators can block enhancers in the genome, only ALOXE3 can act as a heterochromatin barrier. Our results suggest that protection against position effects results from diverse, context-dependent functions of insulators.
3.3 Results

3.3.1 MPIRE measures CRS activity at defined locations genome-wide

The key reagents in MPIRE are pools of cells that carry large numbers of genome-integrated landing pads [108, 191]. The landing pads are integrated randomly into K562 cells where they can serve as sites for recombination mediated cassette exchange (Figure 3.1A). We use the Sleeping Beauty transposon system to deposit multiple landing pads per cell in an unbiased manner throughout the genome [248]. We selected φC31 and BxbI attP sites for recombination because they were previously shown to be highly efficient [277]. Each individual landing pad integration carries a unique genomic barcode (gBC) and each pool of cells contains between 2000-6000 uniquely mapped gBCs (Figure S3.1A).

The strategy in MPIRE is to recurrently integrate reporter genes carrying different insulators into the landing pad pools in parallel. Reporters genes with different insulator sequences are cloned between recombination sites for integration into landing pad pools (Figure 3.1A) and marked with different cis-regulatory barcodes (cBCs) so that the assay can be multiplexed. The reporters are pooled and recombined into each landing pad pool. After recombination, the expressed mRNAs from each landing pad will contain two barcodes: a cBC indicating the identity of the insulator sequence and a gBC indicating the genomic location of the landing pad. We then sequence and count the number of DNA and RNA BC-pairs to calculate expression of each reporter at each location in the genome.

We first verified the utility of landing pad pools for MPIRE. For these analyses, we focused on pool 1. Using qPCR, we estimated that each cell carries approximately 11 independent landing pads, and that the landing pads are well-distributed across the genome (Figure S3.1B), consistent with previous observations that Sleeping Beauty integrates in an unbiased manner [248]. Expression from these landing pads span a large range of expression (>10⁴-fold) indicating that the integrations are sampling diverse genomic environments (Figure S3.1C). To
Figure 3.1: MPIRE is an efficient method for measuring CRS activity genome-wide. (A) Overview of MPIRE method. Pools of cells are generated by integrating landing pads driven by the Cytomegalovirus (CMV) promoter into the genome using SB transposase. The gBCs are mapped to a unique location in the genome, and can be used to recombine CRSs for parallel measurements of cis-regulatory activity across the genome. gBC: genomic barcode. ITR: inverted terminal repeats. CRS: cis-regulatory sequence. cBC: CRS barcode. (B) Proportion of landing pads mapped to a unique location that were recovered after recombination with different CRS constructs. (C,D) Distribution of the number of recombination events per location for an uninsulated construct (C) and construct with the A2 insulator (D).

Test whether different landing pads in the same cells might recombine with each other, we compared the same cells with and without integrase transfection. After integrase transfection, the landing pads have similar expression levels and mostly map back to the same locations (Figure S3.1D-E). Thus, the landing pads in our pools represent diverse genomic environments.
and landing pads in the same cell do not interfere with each other.

We next evaluated the efficacy of large-scale recombination across locations in pools of landing pad cells. We measured the integration efficiency of two reporter genes, one containing only the minimal hsp68 promoter and one that also contains the A2 insulator [71] to test whether the insulator sequence might impact recombination efficiencies. We barcoded each reporter gene with a highly diverse library of random barcodes (rBCs) and transfected them independently into landing pad pool 1. Since the library contains 500,000 rBCs, each integration will be associated with a unique rBC such that the number of rBCs per gBC represents the number of independent integrations at each location. We obtained expressed barcodes from 80% of all mapped landing pads in pool 1 for both the no insulator and A2 constructs (Figure 3.1B). For each landing pad, we obtained a median of 9 independent integrations for the no insulator reporter and 7 independent integrations for the A2 reporter (Figure 3.1C-D) for a total of 18492 (no insulator) and 19031 (A2) integrations. These results demonstrate that MPIRE is a robust method for measuring cis-regulatory activity at a large number of defined locations across the genome.

3.3.2 Measurement of insulator activity across the genome

We selected three previously characterized insulators to test by MPIRE (Figure 3.2A). cHS4 is the most well-characterized insulator and is derived from 5′ end of the chicken β-globin locus [56]. A2 was identified in a screen of endogenous CTCF-containing sequences for enhancer-blocking activity [71], and ALOXE3 tDNA was discovered in a screen for tDNA sequences with insulator activity [70]. All three insulators function in canonical enhancer-blocking assays in K562 cells. The enhancer-blocking activity of cHS4 depends on its CTCF binding site [87, 278], while the enhancer-blocking activity of ALOXE3 depends on its two B box motifs, which recruit RNA polymerase III [279]. A2 binds to CTCF at 100% occupancy in the K562 genome, suggesting that CTCF is likely to be important for its activity [71]. Thus, we mutated the
Figure 3.2: Insulators do not have global effects on gene expression. (A) Insulator constructs and mutants used in this experiment. (B) Insulator activity on plasmids measured by qPCR calculated as fold-change over the uninsulated construct (dotted line). Error bars represent the SEM from three biological replicates. (C) Reproducibility of measurements from two biological replicates. Correlation indicated is Pearson’s r. Color represents the density of neighboring points. (D) Expression distribution of each insulator across all locations. n indicates the number of locations measured for each construct. (E) Heatmap of expression across the genome, where each column represents a different location. Only locations where the “no insulator” construct was measured is plotted to allow for comparisons against the uninsulated reporter.

CTCF binding sites in cHS4 and A2 such that they can no longer bind CTCF (cHS4-mut and A2-mut respectively) and deleted both instances of the B box motif in ALOXE3 (ALOXE3-mut) (Figure 3.2A). In addition, we scrambled the A2 insulator (A2-scrambled) to test the effects of a random DNA sequence that lacks identifiable cis-regulatory elements.
Before performing MPIRE with these insulators, we confirmed their enhancer-blocking activity in a plasmid-based assay (Figure S3.2A). The A2-mut and ALOXE3-mut constructs had less enhancer-blocking activity than their wild-type counterparts in the plasmid assay, but not the cHS4-mut construct. We also measured whether any of the insulator sequences had cis-regulatory activity that was independent of their enhancer-blocking activity. In the absence of an enhancer, the insulators had minimal effects on reporter gene expression (Figure 3.2B), with the exception of the A2-mut, which had some independent repressive activity. However, this repressive activity was not recapitulated globally in the genome (see Figure 3.2D below). Thus, the insulators we tested do not directly influence promoter activity, but do have enhancer-blocking activity.

We performed MPIRE by pooling the insulator constructs and transfecting them into six pools of landing pad cells. In total, we measured expression from each insulator construct in 11365-18031 landing pad locations. The measurements are reproducible across biological replicates consisting of independent transfections (Figure 3.2C). The expression of the non-insulated construct was well correlated with the expression of the landing pads before recombination (Figure S3.2B). These results suggest we were able to obtain accurate and reproducible measurements of multiple cis-regulatory sequences across thousands of genomic locations in parallel.

We first asked whether any of the insulators had widespread effects across the genome. In general, insulators did not have a large impact on the overall distribution of expression from locations across the genome (Figure 3.2D). Formally, this result could arise if insulators increased and decreased expression at different locations in a way that did not alter the overall distribution of expression. However, we find that expression largely depends on genomic location regardless of the identity of the insulator (Figure 3.2E), and that the expression of landing pads with and without insulators are well-correlated, with mutants generally correlating a little better (Figure S3.2C-D). These results are inconsistent with a model of insulators as
modular elements that can function in a wide variety of genomic locations.

### 3.3.3 Insulators function in a small subset of locations that are distinct from each other

While insulators do not appear to have global effects on expression across the genome, there are some locations where expression of the reporter gene with insulators is higher or lower than expected given its location in the genome (Figure 3.2D). We sought to identify the specific locations where insulators might be influencing expression. We first calculated fold-changes in expression for each insulated and non-insulated construct at each landing pad location in the genome. The number of landing pads with matched insulated and non-insulated measurements ranged from 10944-14390 for each insulator. A positive fold-change indicates a location where the insulator upregulates expression relative to the non-insulated construct, while a negative fold-change indicates that the insulator downregulates expression. We defined locations as “insulated” if the insulator caused a greater than 2-fold change in expression (in either direction) with false discovery rate < 0.05 (Figure 3.3A-C, Figure S3.3A-D). As expected from the lack of global insulator activity, we identified only a small number of genomic locations insulated by each insulator (ranging from 10-15%), with even fewer insulated by the mutant insulators (Figure 3.3D). All insulator and mutant constructs can lead to both upregulation and downregulation of expression, but the proportion of locations that are upregulated or downregulated varied by insulator (Figure 3.3E). This suggests that insulators may perform more than one function in a context- and insulator sequence-dependent manner. For the rest of this manuscript we refer to upregulated and downregulated locations as insulator-up (eg. A2-up) or insulator-down (eg. A2-down) respectively.

The insulators could be insulating against common features in the genome, or each insulator may act at distinct locations. The overlap between insulated locations for the three insulators is small, with only 148 insulated locations shared between insulators (Figure 3.3F), suggesting
that each insulator mostly functions at distinct genomic locations. However, this overlap is still greater than expected by chance (bootstrap p-value = 1x10^{-4}), indicating that there might be some shared characteristics between insulated locations. The overlap is similar for both up- and downregulated locations (Figure S3.3E-F), and most of the shared locations (92%) are insulated in the same direction by all three insulators (Figure S3.3G). Because A2 and
cHS4 both contain CTCF binding sites, we asked if there was more overlap between cHS4/A2 shared regions compared to ALOXE3, but the overlap between any two insulators is similar (Figure S3.3H). These results suggest that each insulator functions in a unique type of genomic environment.

We next asked if insulator activities depend on their respective CTCF or B-box motifs. The distribution of fold-changes of mutant versus wild-type insulators was very similar to the fold-changes of insulated versus uninsulated constructs, indicating that even a few nucleotide changes in the insulator sequence is sufficient to generate large changes in gene expression at the same location (Figure S3.4A-C). Using our definition of “insulated” locations above, mutating an insulator can lead to three possible outcomes – the location could still be insulated in the same direction (motif-independent), insulated but in the opposite direction (motif-gained), or could no longer be insulated (motif-dependent). We find that 70-80% of the insulated locations are motif-dependent (Figure 3.3G), with the other locations being mostly motif-independent. The proportion of motif-dependent locations is also similar for both up- and downregulated locations (Figure S3.4D-E), suggesting that the motifs are necessary for both classes of insulator activity. The A2-scrambled control also has higher numbers of motif-dependent locations compared to the A2-mut, suggesting that there might be other sequences in A2 that contribute to insulator activity in addition to the CTCF sites (Figure S3.4F). Our results suggest that insulators only function in small numbers of locations in the genome and depend on their CTCF or B-box motifs.

3.3.4 Insulators block specific enhancers to downregulate expression

We next focused on locations where the addition of an insulator leads to a decrease in gene expression. At these downregulated locations, insulators may be blocking the effects of nearby enhancers. Thus, we predicted that down-regulated locations are normally highly active genomic environments and should contain higher levels of active chromatin modifications.
Consistent with this prediction, insulator-down locations are indeed highly enriched in active chromatin modifications, and depleted for H3K27me3, a marker of polycomb-repressed heterochromatin [280] (Figure 3.4A). Since the insulator is only present on the 5’ end of the reporter, we asked if there was more enrichment upstream compared to downstream of insulator-down locations, but found similar enrichment of signals both upstream and downstream (Figure S3.5A-B). We did not observe the same enrichment at locations insulator-down by the mutant constructs (Figure 3.4B), consistent with the idea that the mutations abrogate insulator activity. Thus, insulator-down locations are enriched for active histone modifications, supporting our model that insulators are blocking enhancers in a motif-dependent fashion at these locations.

Although the same histone modifications were enriched at the insulator-down locations for all three insulators, the different insulators still functioned at distinct genomic locations (Figure S3.3A above). This observation suggests that the insulators are not simply protecting the reporter gene against active chromatin modifications. Instead, each insulator may block distinct transcription factors (TFs) that are present in different genomic environments. In this model, insulators may block TFs that directly occupy the locations surrounding the landing pad, or the insulators may block TFs that bind distal enhancers which loop to these locations.

We first addressed whether TFs that directly occupy insulator-down locations mediate the specificity of different insulators. This model predicts that different TFs will be enriched at locations that are downregulated by the different insulators. In contrast to this prediction, the TFs that are enriched in insulator-down locations are similar for each insulator (Figure 3.4C), suggesting that the TFs that occupy insulator-down locations do not mediate the specificity of different insulators.

Instead, we hypothesized specificity may arise because each insulator blocks different TFs that bind the distal enhancers that loop to insulator-down locations. To address this hypothesis we identified putative enhancers for each insulator-down location from Hi-C data
Figure 3.4: Insulators block specific enhancers in the genome. (A) Mean normalized histone modifications for insulator-unchanged and insulator-down locations. Histone signals are calculated as the mean of 10kb surrounding each location and standardized across locations for each modification. Insulator-down locations are enriched for “active” histone modifications. (B) Mean normalized histone modifications for mutant-unchanged and mutant-down locations. Mutant-down locations are generally not enriched for any histone modifications. (C) Venn diagram of overlap in enriched TFs between insulator-down and insulator-unchanged locations. The number of TF binding peaks in the 10kb surrounding each location was used to calculate enrichment by chi-squared tests. Only TFs that are significantly differentially bound (p < 0.05 after Benjamini & Hochberg (BH) correction) are included in this plot.

(legend continued on next page)
(Figure 3.4 continued) **(D)** Violin plot of number of looped enhancers per location for insulator-unchanged vs insulator-down locations.  **(E)** Violin plot of number of looped transcribed regions per location for insulator-unchanged vs insulator-down locations. *p*-values (two-sided Wilcoxon test) are shown above each plot. **(F)** For all insulators, enhancers that are looped that the insulator-down locations were compared for differences in TF binding measured by ChIP (Chi-squared test). The residuals of TFs that are significantly different (*p* < 0.05) after BH correction was plotted. **(G)** Model for how insulators downregulate expression at specific locations. All insulators are located in open chromatin regions as depicted by the nucleosomes. The insulator blocks enhancers interacting with the location from activating expression, leading to downregulation of expression.

and overlapped these predictions with genomic annotations from chromHMM [100]. Insulator-down locations tend to interact with more enhancers and transcribed regions than insulator-unchanged regions (Figure 3.4D-E), consistent with the idea that insulator-down locations are in more active genomic regions. This difference was reduced in cHS4-mut-down locations and not present in A2-mut-down and ALOXE3-mut-down locations (Figure S3.5C-D). Similarly, there were more ATAC-seq peaks in the 10kb surrounding insulator-down locations compared to insulator-unchanged locations (Figure S3.5E). We then compared TF binding between enhancers looped to each of the insulator-down locations and identified TFs and motifs that are differentially enriched in each group of enhancers (Figure 3.4F, Figure S3.5F). These results support the hypothesis that each insulator blocks a different set of enhancer bound TFs. The general depletion of TF motifs in cHS4-down enhancers could be due to the fact that cHS4 is a longer sequence that contains more TF binding sites than the other two insulators and may therefore interact with a more diverse set of TFs (Figure S3.5G). The lack of overlap in enriched TFs between A2 and cHS4, both of which bind to CTCF, also suggests that the specificity is not purely driven by CTCF. Taken together, insulators appear to block endogenous enhancers from influencing gene expression, and the specificity of insulator activity seems to be at least partially explained by the suite of TFs that are present in the enhancers interacting with each location (Figure 3.4G).
3.3.5  ALOXE3, but not cHS4 or A2, acts as a heterochromatin barrier

Insulator-up locations are upregulated by insulators and are candidates for locations where insulators block the spread of repressive chromatin. This model predicts that insulator-up locations will be enriched for markers of heterochromatin. In contrast to the insulator-down locations, the chromatin landscape of insulator-up regions is not the same for the three different insulators (Figure 3.5A). ALOXE3-up locations are depleted of active histone modifications and enriched for H3K27me3, suggesting that ALOXE3 can block repressive histone modifications as expected for a heterochromatin barrier. Only ALOXE3-up locations are close to heterochromatin (Figure S3.6A) and enriched for looping with repressed (polycomb) regions (Figure S3.6B). Similarly, ALOXE3 constructs are repressed at the least number of locations across the genome, suggesting that ALOXE3 can efficiently block heterochromatin silencing at many locations across the genome (Figure S3.6C). Notably, only H3K27me3, an indicator of facultative heterochromatin, is enriched at ALOXE3-up locations. H3K9me3, which is associated with constitutive heterochromatin, is not enriched in ALOXE3-up locations, consistent with the idea that only facultative heterochromatin remains accessible to TF binding, allowing for upregulation once ALOXE3 is present to block H3K27me3 spreading [281]. This result is also consistent with endogenous ALOXE3 being at the boundary of a H3K27me3 domain in K562 cells [70]. In contrast to ALOXE3-up locations, cHS4-up locations are not enriched for any histone modifications, and A2-up locations are enriched for similar chromatin markers as A2-down locations, with a stronger enrichment for H3K36me3 (Figure 3.5A). From these results we propose that only ALOXE3 acts as a barrier against H3K27me3 spreading, while cHS4 and A2 upregulate expression by other mechanisms. These results suggest that A2 and cHS4 acting via different mechanisms to block repression in the genome.

Our working model predicts that repressive H3K27me3 marks should decrease at insulated locations when ALOXE3 is present. We attempted to test this prediction computationally using Enformer [282], a machine learning model that integrates long-range sequence information
Figure 3.5: ALOXE3 acts as a heterochromatin barrier, while A2 and cHS4 act as enhancers at “primed” locations. (A) Mean normalized histone modifications for insulator-unchanged and insulator-up locations. Histone signals are calculated as the mean of 10kb surrounding each location and standardized across locations for each modification. (B) Violin plot of number of looped enhancers per location for insulator-unchanged vs insulator-up locations. (C) Violin plot of number of looped transcribed regions per location for insulator-unchanged vs insulator-up locations. *p*-values (two-sided Wilcoxon test) are shown above each plot. (D) Expression distribution of the uninsulated reporter at insulator-unchanged vs insulator-up locations. (E) Mean normalized histone modifications for locations upregulated by all insulators, locations upregulated by only one or two locations and uninsulated locations. (legend continued on next page)
**(Figure 3.5 continued)** **(F)** Model for barrier activity at ALOXE3-up locations. ALOXE3 blocks heterochromatin from silencing the gene. **(G)** Model for cHS4/A2-up locations. Instead of acting like an insulator, we propose that cHS4 and A2 act like enhancers in “primed” genomic environments to increase expression.

Enformer can be used to predict gene expression and other histone modifications. Crucially, Enformer only uses sequence information to make predictions, allowing us to compare predictions at genomic locations with and without integrated insulator constructs. Enformer predicted new Cap Analysis of Gene Expression (CAGE) peaks at locations with integrated reporter genes (Figure S3.7A) and higher CAGE fold-changes (insulator/uninsulated) in ALOXE3-up and A2-up locations (Figure S3.7B), so we focused on these two insulators. In the regions flanking ALOXE3-up locations, Enformer predicted higher levels of H3K27me3 (Figure S3.7C), which agrees with our observation that ALOXE3-up are enriched in H3K27me3 chromatin.

At the uninsulated locations, we expect that H3K27me3 from flanking chromatin spreads into the reporter gene to silence expression, resulting in a correlation between H3K27me3 levels in flanking and reporter chromatin. If the insulators block H3K27me3 spreading, then we would expect a decrease in correlation between H3K27me3 levels in flanking and reporter chromatin. We find that the correlation between flanking and reporter chromatin was high in the uninsulated and A2 locations, but decreased in ALOXE3 locations, suggesting that only ALOXE3 can block the spread of H3K27me3 (Figure S3.7D). To test whether ALOXE3-up locations are specifically contributing to the decrease in correlation, we compared the correlations between flanking and reporter gene chromatin for ALOXE3-unchanged vs ALOXE3-up locations. ALOXE3-up locations tend to have lower H3K27me3 than expected compared to ALOXE3-unchanged locations (Figure S3.7E). In contrast, A2-up locations and A2-unchanged locations have the same slopes (Figure S3.7F). These results support the idea that ALOXE3 blocks H3K27me3 spreading from flanking chromatin into the reporter genes, leading to upregulation in ALOXE3-up locations.
3.3.6 cHS4 and A2 enhance expression in “primed” genomic environments

We observed that unlike ALOXE3-up locations, both A2-up and cHS4-up locations seem to be interacting with more enhancers and transcribed regions (Figure 3.5B-C). This suggested that the non-insulated reporters may be more highly expressed in A2-up and cHS4-up locations. Instead, the expression levels of the non-insulated reporters are generally lower in A2-up and cHS4-up locations compared to the unchanged locations, albeit marginally higher ALOXE3-up locations (Figure 3.5D). Thus, we speculate that A2-up and cHS4-up locations are primed for high expression and that A2 and cHS4 act as “enhancers” at these locations by recruiting additional TFs that upregulate expression. This model can also explain the discrepancies between the chromatin environments of A2-up and cHS4-up locations (Figure 3.5A). While A2 requires a more highly active chromatin environment to act as an enhancer, cHS4 can upregulate expression at more neutral environments because it can recruit many more TFs. Consistently, mutant-up locations are all enriched for active histone modifications and looped to more enhancers/transcribed regions (Figure S3.8A-C). Despite that, the uninsulated reporters are also not highly expressed at mutant-up locations (Figure S3.8D). This suggests that deleting the B-box motifs in ALOXE3 abolishes its ability to act as a heterochromatin barrier.

Finally, there are also a small number of locations (86) that are upregulated by all three insulators. These locations are enriched for both active (particularly H3K4me2 and me3) and inactive histone modifications (Figure 3.5E), which could represent a poised chromatin environment. In these environments, ALOXE3 can block heterochromatin-mediated silencing, while cHS4 and A2 can also gain enhancer activity to upregulate gene expression. Taken together, our results suggest that only ALOXE3 acts as a classical heterochromatin blocker, while the other constructs appear to act as enhancers at primed locations across the genome (Figure 3.5F).
3.4 Discussion

Insulators are a class of cis-regulatory sequences that have been associated with several functions. Due to the array of functional assays used to measure insulator activity, it has been difficult to understand or predict insulator function in the genome. Here we developed a high-throughput genome-wide assay, MPIRE, to systematically measure the impact of insulators on gene expression at thousands of locations across the genome and show that insulators have distinct, pleiotropic activities depending on their location in the genome. Our results show that insulator activity is very context-specific and reconcile the differences between genome-wide analyses and analyses of individual insulators. Insulators only function in small numbers of genomic locations, which explains why depletion of insulator-binding proteins such as CTCF or Cohesin have minimal changes on gene expression. Where the insulators do function, different insulators preferentially block specific enhancers that recruit different complements of TFs, and only ALOXE3 protects against heterochromatin silencing. These results explain why one-at-a-time experiments can sometimes show that insulators have large and specific effects on gene expression.

A long-standing question in gene regulation is how enhancers contact and activate target genes in the genome. Recent evidence has shown that enhancers and promoters are broadly compatible with each other, raising the question of how specific interactions are achieved [187, 188, 243]. Insulators are thought to facilitate this process by blocking enhancers from interacting with their target promoters, but it remains unclear whether they act globally across locations in the context of the genome. Our results show that insulators block enhancers at specific locations in the genome, thereby leading to a reduction in gene expression at those locations (Figure 3.4G). For example, ADNP is enriched in enhancers blocked by ALOXE3 (Figure 3.4F), and has previously been implicated in modulating 3D genome interactions by competing with CTCF for binding at SINE B2 transposable elements [161]. Both SINE B2 and ALOXE3 elements are transcribed by RNA Pol III and have been shown to act as insulators [70,
which may explain ALOXE3’s ability to block ADNP-bound enhancers. RNA Pol III has also been suggested to aid ADNP in recruiting distal CTCF sites, which may help ALOXE3 act as an insulator in ALOXE3-down locations [283]. The two insulators that bind CTCF (A2, cHS4) are not any more similar compared to ALOXE3, implying that CTCF itself is not sufficient to explain insulator activity. The specificity of different insulators for different enhancers suggests a model where insulators interact with select enhancers to promote specific enhancer-promoter interactions.

While we only tested three insulators in this study, our results demonstrate that not all insulators possess barrier activity against heterochromatin. Only ALOXE3 appears to be able to block reporter gene silencing. A parsimonious explanation for this difference might be that ALOXE3 binds to RNA PolIII and gets transcribed [70], allowing it to generate an open chromatin environment that can overcome heterochromatin spreading. Indeed, the recruitment of transcriptional activators can insulate against repression [173, 284], and transcription of various TEs have been shown to be sufficient to generate a boundary in mammalian cells [285, 286]. Other groups have also observed that cHS4 is not sufficient to protect against silencing at all genomic locations [79, 271], and can block HDAC-mediated but not KRAB-mediated silencing [80]. We find instead that cHS4 and A2 can act as enhancers in primed environments to upregulate gene expression (Figure 3.5F). Neither cHS4 nor A2 show enhancer activity in plasmid reporter assays [56, 71] (Figure 3.2B), likely because plasmids do not provide the necessary primed environment. Our results are consistent with the previously proposed model that insulation is a dynamic contest between activation and repression [287]. ALOXE3 is transcribed and has the highest activation rate, allowing it to overcome repressed heterochromatin, while A2 has the lowest activation rate and can only overcome very weakly repressive environments. Moving forward, testing more insulators across more locations in the genome will allow us to build a comprehensive and predictive model of insulator activity. These results will also be useful to improve synthetic biology and gene therapy applications where high levels of
sustained gene expression are required.

We developed MPIRE to study insulator activity, but we envision that MPIRE can be used for many other questions involving genomic or chromatin context, for example, to study the interactions between TF binding and chromatin or splicing and chromatin or to assay allelic variants in the the same genomic contexts. MPIRE can also be coupled with other DNA-based readouts such as Cut&Tag or 4C to measure other aspects of genome regulation in addition to expression. Understanding how the different processes in the genome interact with each other will be important for determining and quantifying their contributions to gene expression regulation.

3.5 Methods

All oligos and primers in this study can be found in Table S3.1.

Landing pad library cloning

We first cloned each landing pad construct into a Sleeping Beauty (SB) transposon plasmid containing SB ITRs (gift from Robi Mitra lab). Each landing pad consists of a hsp68 promoter driving the expression the mEmerald reporter gene and is flanked with \( \Phi C31 \) and BxbI attP sites for recombination. The \( \Phi C31 \) attP site, hsp68 promoter, mEmerald fluorophore, poly(A) signal constructs were amplified by PCR from different plasmids and assembled using the NEBuilder HiFi DNA Assembly Master Mix (HiFi Assembly, NEB #E2621). The BxbI attP site was then added using the Q5 Site-Directed Mutagenesis Kit (Q5 SDM, NEB #E0554). To add a library of diverse random barcodes, we first digested the landing pad plasmid with XbaI at 37°C for 16 hours. We then ordered an oligo containing 16 Ns with flanking homology arms to the landing pad plasmid (GWLP P1) and used HiFi Assembly (NEB #E2621) to assemble the oligo to the plasmid (50°C, 15 min).
**Generation of cell lines**

K562 cells were maintained in Iscove’s Modified Dulbecco’s Medium (IMDM) + 10% FBS + 1% non-essential amino acids + 1% penicillin/streptomycin. To generate pools of cells containing landing pads, K562 cells were electroporated with the SB100x transposase (gift from Robi Mitra lab) and landing pad library at a 1:1 ratio. Specifically, we electroporated 4.8 million cells with 20μg SB100x and 20μg landing pad library with the Neon Transfection System 100μL Kit (Life Technologies, #MPK10025). The cells were cultured for a week then sorted into pools of 2000 cells to bottleneck the numbers of gBCs per pool, ensuring that each gBC is associated with a unique location in the cell [103]. In total we generated 6 pools of cells, with pools 1-2 from the same initial transfection pool and pools 3-6 from a second initial transfected pool of cells.

To measure expression of each landing pad we harvested RNA and genomic DNA for each pool of cells using TRIzol reagent (Invitrogen #15596026). 4μg RNA was treated with DNAse using the Rigorous DNase treatment procedure from the Turbo DNase protocol (Invitrogen #AM2238) and reverse transcribed with the SuperScript IV First Strand Synthesis System (Invitrogen #18091050). Barcodes were amplified from the library using Q5 High-Fidelity 2X Master Mix (Q5, NEB #M0492) using primers GWLP P2-3 with 20 cycles. We performed 32 PCR reactions per cDNA sample. For genomic DNA, we performed 4 PCR reactions with 1μg DNA per PCR. PCRs from the same samples were then pooled and purified. Sequencing adapters were then added with two rounds of PCR with Q5 (NEB #M0492) (GWLP P4-7). The resulting libraries were sequenced on the Illumina NextSeq platform.

To map the locations of each landing pad we followed the protocol for mapping TRIP integrations as described in our previous study [243]. The resulting libraries were sequenced on either the Illumina NextSeq or NovaSeq platform.
MPIRE insulator library construction

We first generated a transfer vector containing a reporter gene (hsp68 promoter + mScarlet) flanked by φC31 and BxbI attB sites using HiFi Assembly (NEB #E2621). The insulator sequences were obtained from their respective papers (cHS4 from Chung et al. [56], A2 from Liu et al. [71] and ALOXE from Raab et al. [70]). For ALOXE3, we selected to use the sequence with 2 tDNAs rather than 4 tDNAs because it is a tractable length for cloning (~1kb), does not contain a CTCF binding site and is sufficient for enhancer-blocking activity [70]. cHS4 and A2 were synthesized as gBlocks (Integrated DNA Technologies, IDT) and homology arms for the backbone were added by PCR with GWLP P8-9 (cHS4) or GWLP P10-11 (A2). ALOXE3 was amplified from K562 genomic DNA using primers GWLP P12-13 and homology arms were added by PCR with GWLP P14-15. The constructs were then inserted between the φC31 attB site and hsp68 promoter in the transfer vector by HiFi Assembly (NEB #E2621). The cHS4 x3 mutant design is from Farrell et al. [278] and was introduced to the cHS4 transfer vector with Q5 SDM (NEB #E0554) using primers GWLP P16-17. While the CTCF binding site in A2 has been identified, it has not been tested for its dependence on CTCF. Thus, we used the CTCF PWM (JASPAR [288] MA0139.1) to identify and mutate the three most informative sites in the motif (CACCAGGTGGCGCT → TACCACGTTGCGCT). The mutation was introduced to the A2 transfer vector with Q5 SDM (NEB #E0554) using primers GWLP P18-19. For ALOXE3, the B-box motifs were identified using the B-box frequency matrix from Pavesi et al. [289] with FIMO [269], and both motifs were deleted with Q5 SDM (NEB #E0554) using primers GWLP P20-23. Finally, the A2 sequence was randomly scrambled to generate A2 scrambled and synthesized with homology arms as a gBlock (IDT), then inserted between the φC31 attB site and hsp68 promoter in the transfer vector by HiFi Assembly (NEB #E2621). Insulator barcodes were then added to each construct by site-directed mutagenesis.

To test the efficiency of recombination, we generated constructs barcoded with a diverse library of random barcodes. We digested the no insulator and A2 constructs with NheI at 3°C.
for 16 hours. We then ordered an oligo containing 16 Ns with flanking homology arms to the plasmid (GWLP P24) and used NEBuilder HiFi DNA Assembly to assemble the oligo to the plasmid (50°C, 15 min).

**MPIRE cell pool validation**

For the pilot experiments we focused on cell pool 1. To determine whether the landing pads would recombine with each other, we cotransfected 2.4 million pool 1 cells with 10μg φC31 integrase and 10μg BxbI integrase as described above. The cells were then allowed to grow for a week, and RNA and DNA was harvested using TRIzol reagent (Invitrogen #15596026). We then amplified barcodes from both RNA and DNA to measure expression and mapped the barcodes as described above in the Generation of cell lines section.

To determine the efficiency of recombination we focused on pools 1 and 2. For each pool, we cotransfected 2.4 million cells with a mix of 2.5μg φC31 integrase, 2.5μg BxbI integrase and 5μg of either the no insulator or A2 construct as described above. The cells were then allowed to grow for a week, and RNA and DNA was harvested using TRIzol reagent (Invitrogen #15596026). The RNA was treated with DNase and reverse transcribed to generate cDNA as described above. We then amplified barcodes from both cDNA and DNA in 4 PCR reactions per sample using Q5 (NEB #M0492) and primers specific to the reporter genes (GWLP P3, P25). We pooled PCRs from the same samples for PCR purification, then used 4ng of the product for further amplification with two rounds of PCR to add Illumina sequencing adapters (GWLP P3, P33, P6-7). The resulting libraries were sequenced on the Illumina NextSeq platform.

**Insulator plasmid activity**

To measure expression from plasmids alone, we transfected K562 wild-type cells with each construct separately. For each construct, we performed three biological replicates. We transfected 1.2 million cells with 5μg plasmid DNA per replicate using the Neon Transfection System
100μL Kit (Life Technologies #MPK10025). RNA was harvested 72 hours after transfection using the Monarch Total RNA Miniprep Kit (NEB #T2010). mScarlet and HPRT (housekeeping gene for normalization) levels were then measured by qPCR using the Luna Universal One-Step RT-qPCR Kit (NEB #E3005) with 100ng total RNA per reaction, three reactions per replicate (mScarlet primers: GWLP P25-26, HPRT primers: GWLP P27-28). Expression was calculated using the $\delta\delta$Ct method, normalized to the no insulator construct.

**Enhancer-blocking assay**

For the enhancer-blocking assay, we designed plasmids that contain either enhancer-promoter or enhancer-insulator-promoter sequences to test whether the insulator can block enhancer-mediated activation of promoter activity. We first selected an enhancer that was previously shown to be active with the hsp68 promoter in K562 cells [290]. The full sequence of the enhancer is:

```
GCCCCCCTTCTTCTATGTCTGATGGAGTTTCCTCTCTAAGTAGCCATTTTATTCTGCTGACT-
CACCCTCTAAACTCCCGGTCTTATTCCATCCTGCCTCAGGGTCTGTGGTGTAGTCATAGCACAT-
GCATCTCCTCCGGCTCGCTGATT
```

The no insulator construct was digested with EcoRI to insert the enhancer upstream of the hsp68 promoter. The enhancer was amplified by PCR to add overhangs to the backbone vector (GWLP P29-30) and assembled into the backbone by HiFi Assembly (NEB #E2621).

To add the different insulator constructs, the +enhancer plasmid was digested with AgeI and EcoRI. For A2, A2-mut, A2-scrambled, cHS4 and cHS4-mut, we digested the respective plasmids with AgeI and EcoRI and ligated them to the backbone with T4 DNA Ligase (NEB #M0202). For ALOXE3 and ALOXE3-mut, we amplified the insulators by PCR to add overhangs to the backbone vector and assembled the insulators into the backbone by HiFi Assembly (GWLP P31-32).

We measured expression from these constructs in the same way as described in the Insulator
MPIRE

We pooled insulator constructs for transfection into landing pad cell pools. In the first experiment, we combined the no insulator, A2, A2 mut, A2 scrambled, cHS4, and ALOXE3 constructs and tested them in pools 1 and 2. In the second experiment, all the constructs were combined in equal amounts and tested in pools 3-6. We used the Neon Transfection System 100μL Kit (Invitrogen #MPK10025) for transfection. Each pool was transfected separately and we performed two replicates per pool. For each replicate, we performed four transfections, each with 1.2 million cells and a mix of 2μg φC31 integrase, 2μg BxbI integrase and 4μg of the insulator constructs pooled in equal ratios.

We harvested DNA and RNA from cells one week after transfection using the TRIzol Reagent (Invitrogen #15596018). The RNA was treated with DNase and reverse transcribed to generate cDNA as described above. Barcodes were amplified from cDNA using Q5 (NEB #M0492) with primers specific to the reporter gene (GWLP P3, P25), with 32 PCR reactions per biological replicate. Similarly, barcodes were amplified from genomic DNA with 8 PCR reactions per biological replicate. We pooled PCRs from the same biological replicates for PCR purification, then used 4ng of the product for further amplification with two rounds of PCR to add Illumina sequencing adapters (GWLP P4, P33, P6-7). The resulting libraries were sequenced on the Illumina NextSeq platform.

Data analysis

The preprocessing of sequencing reads was performed with Python 3.9. All statistical analyses and figures were done in R 4.2.0. We used bedtools v2.30.0 [291] for genomics analyses.
MPIRE cell pools data processing

For expression from the initial cell pools and cells transfected with integrase only, we first filtered for reads that contained the gBC in the correct sequence context. We then filtered for gBCs with >20 reads and normalized the reads by sequencing depth. We then calculated expression as log2(RNA counts/DNA counts) for all barcodes.

To identify the locations of the landing pads, we obtained paired-end reads containing the barcode on one read and the sequence of the integration site on the other. We matched the barcodes with the integration site sequence, then aligned them to hg38 with BWA using default parameters. We only kept barcodes where the reads for one location represented at least 80% of all the locations for that barcode. For pools from the same initial transfection, a small number of barcodes are shared between them because the same cells were sorted into multiple pools. These barcodes are indicated with Px instead of a pool number.

Recombinase efficiency data processing

To evaluate the efficiency of recombination, we first filtered for reads that contained both barcodes (rBC and gBC) in the correct sequence context. We then filtered for barcode pairs that had >20 reads associated with them and tabulated the number of rBCs/gBC for either the no insulator or A2 construct (Supplementary Table X). For the proportion of recovered LPs, we compared the DNA gBCs recovered from cells transfected with either the no insulator or A2 construct to the DNA gBCs recovered from the original cell pools. To compare locations before and after integrase-only transfection, we compared the mapped locations in pool 1, only considering gBCs with exact matches.

MPIRE data processing

For the recombined MPIRE libraries, we first filtered for reads that contained both barcodes (iBC and gBC) in the correct sequence context. The gBCs were then compared to the list of
mapped gBCs (gBCs that have a unique location associated with them). A gBC was assigned to a mapped gBC if both barcodes have a hamming distance <5 and the gBC is ≥5 hamming distance from all other gBCs (the average hamming distance between mapped gBCs is 9). We then tabulated the number of barcodes or barcode pairs after assigning gBCs to a mapped gBC for both RNA and DNA.

To calculate reproducibility and to compare expression before and after recombination, we calculated expression as \( \log_2(\text{RNA counts/DNA counts}) \) for all barcodes that are present in both the RNA and DNA pools.

To calculate expression after recombination of the insulator library, including those that might not have an RNA count because they are lowly expressed, we turned to the MPRAnalyze tool [292]. We added a pseudocount of 1 to all counts and used the analyzeQuantification function to calculate the transcription rate (\( \alpha \)) for each construct at each location using counts from both biological replicates. If the same barcode is present in multiple pools, then the counts from all replicates across the pools were used as input, so that each location only gets one \( \alpha \) value. To quantify the effects of the insulators, we used the analyzeComparative function in MPRAnalyze to calculate the fold-change of the insulated vs uninsulated locations.

**Histone modifications**

We downloaded ChIP data for various histone modifications in K562 from the ENCODE database [38]. We used bedtools to map the mean histone signals in the 20kb surrounding each landing pad (10kb upstream and downstream). For each histone modification, we standardized the signal values, then computed the average signal for ins-unchanged, ins-up and ins-down locations respectively. All heatmaps were generated using the ComplexHeatmap package in R [260].
**Other genomics analysis**

The core 15-state chromHMM annotation for K562 cells was downloaded from the Roadmap Epigenomics Project [99] and similar annotations were grouped. We overlapped the landing pads with the corresponding annotation using the GenomicRanges R package [262].

ATAC-seq data for K562 cells was downloaded from the ENCODE database [38]. We used bedtools to count the number of ATAC-seq peaks around each landing pad.

To identify locations that are looped to each location, we used K562 Hi-C data that was previously generated and processed [293]. We then used the peakHiC tool [294] to call loops for each landing pad with the following parameters: window size = 80, alphaFDR = 0.5, minimum distance = 10kb, qWr = 1. The regions identified by peakHiC were overlapped with chromHMM annotations using the GenomicRanges R package [262] to identify putatively interacting enhancers with each landing pad.

**Transcription factors enrichment analysis**

We downloaded all available TF ChIP data in K562 from the ENCODE database [38]. We used bedtools to count the number of TF ChIP peaks in the surrounding 20kb of each landing pad (10kb upstream and downstream) or to count the number of peaks in each putative enhancer identified above. We compared the number of peaks between different groups of landing pads or enhancers using chi-squared tests followed by p-value correction with the Benjamini & Hochberg (BH) method.

**Enformer predictions**

We decided to use Enformer because it can predict histone modifications and expression directly from sequence [282]. We first downloaded the published, trained model from TensorFlow Hub (https://tfhub.dev/deepmind/enformer/1). For each landing pad, we considered the landing pad to be the center of the sequence to use as input and extended it out on both
flanks using the hg38 genome for a total length of 393,216bp, which was the length used in the training model. For each location, we used the model to predict CAGE and H3K27me3 in K562 cells. To calculate CAGE signals, we summed CAGE scores across all windows that contain the landing pad. We then calculated fold-changes for each insulator by taking log₂(predicted_insulator_CAGE/predicted_uninsulated_CAGE). To calculate H3K27me3 signals, we either summed across all windows that contain the landing to get the landing pad H3K27me3 signals, or we summed across all windows outside the landing pad (that were used in the input data) to get flanking H3K27me3 signals. The values were normalized by z-scoring the predicted values for each insulator.

3.6 Supplementary Figures
Figure S3.1: Overview of cell pools for MPIRE. (A) Barplot of the number of mapped landing pads per pool. The exact number is indicated above each bar. (B) Barplot of the number of landing pads in each chromHMM type per pool. (C) Expression across all landing pads in each pool. (D) Expression of landing pads before and after transfection with ϕC31 and BxbI integrases only. Correlation indicated is Pearson's r. (E) Percentage of locations that mapped to the same or different locations after transfection with the integrases only.
Figure S3.2: MPIRE reliably measures insulator activity. (A) Enhancer-blocking assay for testing insulator activity. Expression was measured by qPCR and normalized to the promoter-only (hsp68) construct. Each dot represents a measurement from one replicate, two transfection replicates were performed for each construct. (B) Correlation between reporter gene measurements from the initial landing pad and from the recombined uninsulated construct for each pool. (C) Correlation between uninsulated construct and respective insulator at each location. (D) Correlation between uninsulated construct and respective mutant construct at each location. Correlations shown are Pearson’s $r$. 

107
Figure S3.3: Insulators function at distinct locations in the genome. (A-D) Volcano plot of insulator effect across locations for A2 mutant (A), A2 scrambled (B), cHS4 mutant (C) and ALOXE3 mutant (D). Colored points (insulated locations) represent locations where $\log_2(\text{insulator/no insulator}) > 2$ and fdr < 0.05. (E) Overlap between locations insulator-down by the respective insulators. (F) Overlap between locations upregulated by the respective insulators. (G) Percentage of shared insulated locations (insulated by all insulators) that are insulated in the same or different direction. (H) Upset plot of the number of shared insulated locations by the indicated pairs.
Figure S3.4: Insulators are mostly dependent on their respective sequence motifs. (A, B, C) Fold-changes between either the insulator and uninsulated constructs or insulator and its mutant construct for A2 (A), cHS4 (B) or ALOXE3 (C). (D, E) Percentage of insulated locations that are motif-gained, motif-independent or motif-dependent for insulator-down (D) or insulator-up (E) locations. (F) Percentage of insulated locations in each category for A2-scrambled-up and A2-scrambled-down locations.
Figure S3.5: Insulators block specific enhancers in the genome. (A, B) Mean normalized histone modifications either 10kb upstream (A) or downstream (B) of insulator-unchanged and insulator-down location. Histone signals are standardized across locations for each modification. (C) Violin plot of number of looped enhancers per location for mutant-unchanged vs mutant-down locations. (D) Violin plot of number of looped transcribed regions per location for mutant-unchanged vs mutant-down locations. (E) Violin plot of the number of ATAC peaks for insulator-unchanged vs insulator-down locations. p-values (two-sided Wilcoxon test) are shown above each plot. (F) For all insulators, enhancers that are looped that the insulator-down locations were compared for differences in TF motif content (Chi-squared test). The residuals of motifs that are significantly different ($p < 0.05$) after Benjamini & Hochberg (BH) correction was plotted. (G) Number of TF motifs in each insulator.
Figure S3.6: ALOXE3 acts as a barrier against heterochromatin. (A) Distance to closest heterochromatic region for each insulator-unchanged and insulator-up location. (B) Number of regions looped to each location designated as repressed by chromHMM. p-values (two-sided Wilcoxon test) are shown above each plot. (C) Proportion of locations that are expressed (contains detectable RNA counts) vs repressed (no detectable RNA counts) for each insulator.
Figure S3.7: Enformer predicts ALOXE3 barrier activity and loss of H3K27me3. (A) Representative example of predicted CAGE signals by Enformer for all insulator constructs. Each track is 114,688bp centered on the site of the landing pad insertion (chr5: 151225955). (B) Violin plot of fold-changes in CAGE signal (calculated as log$_2$(insulator/uninsulated)) in insulator-unchanged vs insulator-up locations. (C) Violin plot of summed H3K27me3 levels in regions flanking the landing pad for insulator-unchanged vs insulator-up locations. p-values (two-sided Wilcoxon test) are shown above each plot. (D) Correlation between summed H3K27me3 levels in the LP and flanking regions. Correlations shown are Pearson’s r.

(legend continued on next page)
**Figure S3.7 continued** (E, F) Same correlation plot in D for the ALOXE3 (E) and A2 (F) constructs but colored by whether the points are insulator-unchanged or insulator-up locations. The line presents the best fit linear model with 95% confidence interval surrounding the line in light gray. All H3K27me3 values are z-scored for normalization.

---

**Figure S3.8: cHS4 and A2 act as enhancers at “primed” genomic locations.** (A) Mean normalized histone modifications for mutant-unchanged and mutant-down locations. Histone signals are calculated as the mean of 10kb surrounding each location and standardized across locations for each modification. All mutants are enriched for “active” histone modifications. (B) Violin plot of number of looped enhancers per location for mutant-unchanged vs mutant-up locations. (C) Violin plot of number of looped transcribed regions per location for mutant-unchanged vs mutant-up locations. *p*-values (two-sided Wilcoxon test) are shown above each plot. (D) Expression distribution of the uninsulated reporter at mutant-unchanged vs mutant-up locations.
## 3.7 Supplementary Tables

**Table S3.1:** Primers used in this study

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWLP P1</td>
<td>TGTACGGTACAAACCCGCTTCGAGTCTCAGANNNNNNNNNNNNNNGATCACTCGAGTTGTGGCCGCGCTTT</td>
<td>Oligonucleotide to add random barcodes to GWLP plasmid</td>
</tr>
<tr>
<td>GWLP P2</td>
<td>AGCAAAAGACCCCAACGAGAA</td>
<td>F primer for barcode amplification from GWLPs</td>
</tr>
<tr>
<td>GWLP P3</td>
<td>AACGCCAGGGTTTTCCCAA</td>
<td>R primer for barcode amplification from GWLPs</td>
</tr>
<tr>
<td>GWLP P4</td>
<td>CTTTCCCTACACGCGCTCTTCCGATCT[N]1-4CAAACCCGCTTCGAGTCTAGA</td>
<td>F primer to add partial sequencing adapters to GWLP barcodes, N represents random bases of length 1-4 to add diversity to the sequencing pool</td>
</tr>
<tr>
<td>GWLP P5</td>
<td>GTGACTGGAGTTCCAGACGCTGTGCTCTTCCGATCT[N]1-4GCCAGGGTTTTCCCAAC</td>
<td>R primer to add partial sequencing adapters to GWLP barcodes, N represents random bases of length 1-4 to add diversity to the sequencing pool</td>
</tr>
<tr>
<td>GWLP P6</td>
<td>AATGATACGCGCGACCACCGAGATCTACACNNNNNNNACTCTTTCCCTACACGACGCT</td>
<td>Final Illumina sequencing primer containing partial sequencing adapters to amplify barcode libraries, N represents index (I2) sequences for multiplexing</td>
</tr>
</tbody>
</table>
Table S3.1: Primers used in this study. (continued)

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWLP P7</td>
<td>CAAGCAGAAGACGGCATACGA-GATNNNNNNNNNTGACTGGAGTTCA-GACGTG</td>
<td>Final Illumina sequencing primer containing partial sequencing adapters to amplify barcode libraries, N represents index (I1) sequences for multiplexing</td>
</tr>
<tr>
<td>GWLP P8</td>
<td>TCCACCTCACCCCATCGAATTGAGCT-CACGGGGACAGCC</td>
<td>F primer to add homology arms to cHS4</td>
</tr>
<tr>
<td>GWLP P9</td>
<td>ACTCTCCGCTGGGCCACCGTGTTACATTTCTCAGCT-GACTCCGTCTCTGG</td>
<td>R primer to add homology arms to cHS4</td>
</tr>
<tr>
<td>GWLP P10</td>
<td>CGTACTCCACCTCACCAATCGAATTCAATCGG-GCTGGCCCATAGTA</td>
<td>F primer to add homology arms to A2</td>
</tr>
<tr>
<td>GWLP P11</td>
<td>CGCCCACTCTCCGCTGGGCCACCGGTAGCGA-GATTTCCGTCTTAA</td>
<td>R primer to add homology arms to A2</td>
</tr>
<tr>
<td>GWLP P12</td>
<td>CAGATAAAATCCTCTGCGGTTCCTACAAGC</td>
<td>F primer to amplify ALOXE3 from genome</td>
</tr>
<tr>
<td>GWLP P13</td>
<td>AGCTAAGACCTGTGGTAACTGCTG</td>
<td>R primer to amplify ALOXE3 from genome</td>
</tr>
<tr>
<td>GWLP P14</td>
<td>TCCACCTCACTCATCGAATTCCAGATCCTCTGTCGG</td>
<td>F primer to add homology arms to ALOXE3</td>
</tr>
<tr>
<td>GWLP P15</td>
<td>ACTCTCCGCTGGGCCACCGGTAGCTAAGCTGCTG</td>
<td>R primer to add homology arms to ALOXE3</td>
</tr>
<tr>
<td>GWLP P16</td>
<td>CTTTTTCAGGCAGCGACGCCGCGCATCCGCGG</td>
<td>Site-directed mutagenesis primer to generate cHS4 x3 mutant</td>
</tr>
<tr>
<td>GWLP P17</td>
<td>CTATTTGGAGGGACGTAAATTACATCC-CTGGGGG</td>
<td>Site-directed mutagenesis primer to generate cHS4 x3 mutant</td>
</tr>
<tr>
<td>Name</td>
<td>Sequence</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>GWLP P18</td>
<td>TTGCGCTCAGAGTCTGCGGAGGCAT</td>
<td>Site-directed mutagenesis primer to generate A2 mutant</td>
</tr>
<tr>
<td>GWLP P19</td>
<td>CGTGGTAGTAATTGTGCTGTGCTCTCTTTCTG</td>
<td>Site-directed mutagenesis primer to generate A2 mutant</td>
</tr>
<tr>
<td>GWLP P20</td>
<td>CTGTTCGGGCAGGATGTC</td>
<td>Site-directed mutagenesis primer to generate delete ALOXE3 B-Box 1</td>
</tr>
<tr>
<td>GWLP P21</td>
<td>CTGGACCCCTCAGATTTAAAGTCTG</td>
<td>Site-directed mutagenesis primer to generate delete ALOXE3 B-Box 1</td>
</tr>
<tr>
<td>GWLP P22</td>
<td>CGGTGGAACCTTAAGCTTTT</td>
<td>Site-directed mutagenesis primer to generate delete ALOXE3 B-Box 2</td>
</tr>
<tr>
<td>GWLP P23</td>
<td>CGGATCGCTGGATTCAGA</td>
<td>Site-directed mutagenesis primer to generate delete ALOXE3 B-Box 2</td>
</tr>
<tr>
<td>GWLP P24</td>
<td>GGACGAGCTGTCAAGTAAGCTAGC-NNNNNNNNNNGAGCGGATAAGTAGCTGA</td>
<td>Oligonucleotide to add random barcodes to insulator constructs</td>
</tr>
<tr>
<td>GWLP P25</td>
<td>GACTACACCGTGGTGGGAACA</td>
<td>mScarlet F qPCR primer</td>
</tr>
<tr>
<td>GWLP P26</td>
<td>TTGGTCACCTTCAGCTGG</td>
<td>mScarlet R qPCR primer</td>
</tr>
<tr>
<td>GWLP P27</td>
<td>TGACACTGGCAAACAATAATGCA</td>
<td>HPRT F qPCR primer</td>
</tr>
<tr>
<td>GWLP P28</td>
<td>GGTCCTTTTCACCCAGCAAGCT</td>
<td>HPRT R qPCR primer</td>
</tr>
<tr>
<td>GWLP P29</td>
<td>TCCACCTACCCCATCTCTAGGCCCCCTTCTTCTATG</td>
<td>F primer to add overhangs to enhancer for HiFi assembly</td>
</tr>
<tr>
<td>GWLP P30</td>
<td>ACTCTCCGGCTGGGCCACCGGTGATC-GAATTCAATCCAGCGGAGCAGGAGAGA</td>
<td>R primer to add overhangs to enhancer for HiFi assembly</td>
</tr>
<tr>
<td>Name</td>
<td>Sequence</td>
<td>Description</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>GWLP P31</td>
<td>TCCGGCTCGCTGATTGCATGCCAGATAAAATC-CTCTGTGTCGG</td>
<td>F primer to add overhangs to ALOXE3 or ALOXE3-mut for HiFi assembly</td>
</tr>
<tr>
<td>GWLP P32</td>
<td>ACTCTCCGCTGGGCCACCGGTAGCTAAGACCTGATGGTA</td>
<td>R primer to add overhangs to ALOXE3 or ALOXE3-mut for HiFi assembly</td>
</tr>
<tr>
<td>GWLP P33</td>
<td>CTTTCCCTACACGACGCTTCCGTCTCCGATCT[N]1-4AGCTGTACAAGTAAGCTAGC</td>
<td>F primer to add partial sequencing adapters to GWLP barcodes, N represents random bases of length 1-4 to add diversity to the sequencing pool</td>
</tr>
</tbody>
</table>
Chapter 4: Effect of genomic and cellular environments on gene expression noise

This work was performed in collaboration in with Avinash Ramu, Siqi Zhao and Barak Cohen. All authors conceived and designed the project. I generated the cell lines, performed bulk measurements, mapped the locations of the integrations and helped with 10x library preparation. I also performed the chromatin and transcription factor analyses. Avinash Ramu, Siqi Zhao and I are co-first authors on this manuscript that all authors wrote together. The manuscript is currently under review and has been uploaded to Biorxiv (doi: 10.1101/2022.08.31.506082) [293].

4.1 Abstract

Individual cells from isogenic populations often display large cell-to-cell differences in gene expression. This “noise” in expression derives from several sources, including the genomic and cellular environment in which a gene resides. Large-scale maps of genomic environments have revealed the effects of epigenetic modifications and transcription factor occupancy on mean expression levels, but leveraging such maps to explain expression noise will require new methods to assay how expression noise changes at locations across the genome. To address this gap, we present Single-cell Analysis of Reporter Gene Expression Noise and Transcriptome (SARGENT), a method that simultaneously measures the noisiness of reporter genes integrated throughout the genome and the global mRNA profiles of individual reporter-gene-containing cells. Using SARGENT, we performed the first comprehensive genome-wide survey of how genomic locations impact gene expression noise. We found that the mean and
noise of expression correlate with different histone modifications. We quantified the intrinsic and extrinsic components of reporter gene noise and, using the associated mRNA profiles, assigned the extrinsic component to differences between the CD24+ “stem-like” sub-state and the more “differentiated” sub-state. SARGENT also reveals the effects of transgene integrations on endogenous gene expression, which will help guide the search for “safe-harbor” loci. Taken together, we show that SARGENT is a powerful tool to measure both the mean and noise of gene expression at locations across the genome, and that the data generated by SARGENT reveals important insights into the regulation of gene expression noise genome-wide.

4.2 Introduction

Gene expression is noisy, even among individual cells from an isogenic population [295]. Noisy gene expression leads to variable cellular outcomes in differentiation [198–200, 296], the response to environmental stimuli [297, 298], viral latency [299], and chemotherapeutic drug resistance [300–302]. Explaining the causes of noisy expression remains an important challenge.

A gene’s genomic environment, defined here as the composition of nearby cis-regulatory elements and local epigenetic marks, can influence its expression noise. Some features of genomic environments that can affect noise include enhancers, histone modifications, and transcription factor (TF) occupancy [207, 209, 213, 214, 223, 224, 232]. These observations raise the possibility that genome-wide patterns of expression noise could be explained using the large-scale epigenetic maps that have proved useful in explaining mean expression levels [99, 103, 303]. Leveraging these resources to explain expression noise will require maps of the genome that show the influence of diverse genomic environments on this noise. Producing these maps will require new experimental approaches, because the existing studies demonstrating the effects of epigenetic marks on expression noise have either been performed on endogenous genes, where the effects of different chromosomal locations are confounded with the effects
of the different endogenous promoters or rely on low-throughput imaging methods. Dar et al. assayed the noisiness of large numbers of genomic integrations but was unable to assign genomic locations to the measured reporter genes [209]. Two other studies have assayed integrations in a high-throughput manner but measured protein levels by flow cytometry rather than mRNA levels [210, 212]. Thus, we still lack a high-throughput, systematic way of quantifying the impact of genomic environments on expression noise.

In addition to intrinsic features such as the local genomic environment, extrinsic features, such as the global cellular state of a cell, can also influence gene expression noise [204, 304–307]. For example, variation in the cell cycle, cell size, or signaling pathways can all impact gene expression noise [295, 308, 309]. However, the relative contributions of intrinsic vs extrinsic features on gene expression noise in mammalian cells remains unclear.

Here we report Single-cell Analysis of Reporter Gene Expression Noise and Transcriptome (SARGENT), a highly parallel method to measure the mean and noise of a common reporter gene that has been integrated at locations across the genome. Analysis of SARGENT data showed that different histone modifications explain the mean and noise produced across the genome. In SARGENT, multiple reporters are integrated in each cell, allowing us to separate the intrinsic and extrinsic contributions to noise. Sequencing the associated single-cell mRNA transcriptomes further enabled us to attribute the extrinsic noise to differences in the cellular substates between isogenic cells. To our knowledge, this is the largest genome-wide survey of the impact of intrinsic and extrinsic noise in gene expression. Taken together, our results show that SARGENT is a powerful tool to study how genomic environments and cellular context control expression noise.
4.3 Results

4.3.1 A high-throughput method to measure mean and noise across the genome

We developed a high-throughput method to test the effects of genomic environments on the mean and noise of gene expression. Our goal was to integrate a common transgene across the genome and then, for individual cells, measure both the transcripts produced from the transgene and the global mRNA profile. This allows us to compute the mean and noise of reporter gene expression at each location, and correlate reporter gene expression with the cellular mRNA state of each cell. Because every unique integration contains the same transgene, the measured differences in the mean and noise of reporter gene expression are directly attributable to the influence of genomic environments or cellular states.

We first generated a reporter gene with a library of 16bp random barcodes (Location barcode, locBC) in its 3’UTR (Figure 4.1). Due to the diversity of the locBCs, each locBC is only associated with a single location in the genome [103]. The reporter gene consists of a Cytomegalovirus (CMV) promoter driving the expression of a fluorescent protein and contains a capture sequence from the 10x Genomics Single Cell Gene Expression 3’ v3.1 with Feature Barcoding Kit. The 10x gel beads contain both the complementary capture sequence and polyT sequences, allowing us to isolate the transcripts produced from the reporter gene and the cellular transcriptome.

To generate chromosomal integrations across the genome we cloned the reporter gene library onto a piggyBac transposon vector. The library was transfected into cells along with piggyBac transposase to allow random integrations of the reporter into the genome. After selecting for integrations in K562 cells, we mapped the locations of each integrated reporter (IR) and assigned each locBC to a specific genomic location. We then captured the reporter gene transcripts from single cells and amplified the barcodes (10x cell barcode, UMI, and locBC) using primers specific to our reporter gene (Methods). After sequencing and tabulating
Figure 4.1: Overview of the SARGENT workflow. In step 1, a reporter gene driven by the CMV promoter is randomly barcoded with a diverse library of location barcodes (locBC) upstream of the 10x capture sequence (CS). The reporter genes are randomly integrated into K562 cells and sorted for cells with successful integrations (step 2), then sorted again after a week into pools to ensure that each barcode is only represented once per pool (step 3). We then performed scRNA-seq to capture the transcriptome and amplify the expressed barcodes from integrated reporter genes (step 4). The number of expressed barcodes per cell were then tabulated (step 5). To identify the genomic locations of the integrations we also mapped the location of each locBC in bulk (step 5). (ITR = Inverted terminal repeat; prom = promoter)

the mRNA counts for each IR we computed the expression level of the reporter gene at each genomic location in each single cell. For a subset of cells we also sequenced the mRNA profiles to simultaneously reveal the cell state of each individual cell.

122
4.3.2 SARGENT measurements are accurate and reproducible

We performed SARGENT in K562 cells because of the abundance of public epigenetic data available for this cell line. We first assessed the reproducibility of the SARGENT method. Because replicate infections result in pools of cells with insertions at different genomic locations, we could not assess the reproducibility of independently transfected pools of cells. Instead, we assessed the reproducibility of SARGENT by growing the same pool of insertions twice and performing the SARGENT workflow independently on each sample. We detected 589 identical IR locations in both replicates, which represented 96% of the total IRs observed in both replicates. After quality control, we obtained data from 7680 single cells across replicates, and a total of 2,940,912 unique molecular identifiers (UMIs) representing expressed barcodes from the IRs in these cells. The replicates were well correlated for measurements of both mean and noise measured at each IR location (Figure 4.2A-B, mean Pearson’s $r = 0.76$, noise Pearson’s $r = 0.72$) indicating that measurements obtained by SARGENT are reproducible.

To validate the single-cell measurements made by SARGENT, we also performed single-molecule Fluorescence In Situ Hybridization (smFISH) on two known locations. The measurements of mean and noise made by smFISH agree with the SARGENT measurements for those locations (Figure S4.1) demonstrating that our method is accurate and reproducible for measuring the mean and noise of expression.

4.3.3 Measurements of mean-independent noise across different chromosomal environments

In total, we performed four experiments and generated mean and noise measurements for 939 integrations. The integrations were spread across the genome and found in regions with different chromHMM annotations [100] (Figure S4.2A-B), allowing us to study the effects of diverse chromosomal environments on expression noise.

The mean and noise of expression are often highly correlated [310, 311]. Similarly, we found
a strong correlation between the mean and noise in SARGENT data, indicating that a large proportion of an IR’s noise is explained by its mean level of expression (Figure 4.2C). To identify chromosomal features that control expression noise independent of mean levels we regressed out the effect of mean levels on noise, leaving us with a metric we refer to as mean-independent noise (MIN) [310]. By design, MIN levels of IRs are uncorrelated with their mean expression levels (Figure 4.2D) whereas other measures of noise, such as the coefficient of variance or the Fano factor, retain residual correlation with mean levels in our data (Figure S4.2C-D). Thus,
we used MIN as a measure of expression noise for all following analyses.

4.3.4 Expression mean and noise are associated with different chromosomal features

We sought to identify chromatin features that would explain differences in MIN levels between genomic locations. Studies of genome-wide chromatin features in many cell lines and tissues have shown that the mean expression of a gene is correlated with its surrounding chromatin marks [98, 103]. Thus, we asked whether chromatin features might also explain patterns of MIN across the genome. We split the IRs into bins of high or low mean levels, or high or low MIN levels, and identified chromatin features that were enriched in specific bins. As expected, IRs with high mean expression had higher levels of active chromatin marks such as H3K27ac, H3K4 methylation, H3K79me2 and H3K9ac (Figure 4.3A). Conversely, IRs with high MIN did not exhibit significant differences between H3K27ac or H3K4me1 levels, and low MIN locations showed slightly elevated levels of H3K4me2/3, H3K79me2 and H3K9ac (Figure 4.3B). These results suggest that different chromatin modifications influence the mean and noisiness of expression, and that more active genomic locations might also reduce MIN. This observation is consistent with previous studies showing that repressed chromatin is associated with high MIN [210, 224].

The binding of TFs also impacts noise in gene expression. To identify TFs that might affect noise, we identified TFs whose occupancy is enriched near either high or low MIN IRs. Sequences at low MIN IRs are enriched for transcriptional activators such as SP1 and E2F4, while sequences at high MIN IRs are enriched for other TFs including TFs containing basic helix-loop-helix (bHLH) domains (Figure 4.3C), suggesting that the cofactors recruited by different TFs have separable effects on expression mean and noise.

To assess the power of genomic features to predict the MIN of IR locations we trained a logistic regression model using chromatin modifications and DNA sequence features to classify
Figure 4.3: Expression mean and noise are associated with different chromosomal features. (A) Active histone modifications associated with high or low mean IRs. Start indicates the location of the IR, and each location was extended 5kb on either side. (B) Active histone modifications associated with high or low MIN IRs are different from those associated with mean. (C) Motifs enriched in high or low MIN IRs respectively, and potential TFs that match these discovered motifs. (D) Logistic regression weights of various intrinsic features associated with high or low MIN IRs. Red bars: $p$-value < 0.05; Pink bars: 0.05 < $p$-value < 0.1 from the logistic regression model.
high and low MIN locations and achieved an accuracy of 76%. When applied to data from a different pool, the trained model achieved 67% accuracy. The features with significant weights are the H3K4me3 mark, TF motifs (RARG, FOXO4, HIF1A, TFAP4, CREM, ATF1, NFIC, and NFIA) and whether the IR location was inside a gene (Figure 4.3D). Being inside a gene reduced the probability of being a high noise IR location, which could be due to local regulatory elements that dampen the noisiness of a gene's expression. Similar to our results above, lower H3K4me3 increased the probability of being a high noise IR location. H3K4me3 is associated with active chromatin and supports the hypothesis that higher activity reduces IR MIN. Our observation is consistent with a previous study showing that H3K4me3 correlates with reduced noise at endogenous genes [224]. With respect to the effects of TFs on noise, the presence of some TF motifs increase the probability of being a high noise IR location (NFIC, CREM, TFAP4, CLOCK), whereas other TFs reduce the probability of being a high noise location (RARG, NFIA, ATF1, FOXO4, HIF1A).

We used a similar logistic regression framework to identify features that separate IR locations with high or low mean levels of expression and achieved an accuracy of 82%. When applied to holdout data, the trained model achieved 62% accuracy. The chromatin features that increase the probability of being a high mean IR location are lower levels of H3K27me3, lower levels of H3K4me2 and a higher number of ATAC-seq peaks, which agrees with the known effects of these features in bulk mean expression. The motifs that increased the probability of being a high mean IR location are higher numbers of motifs of the ZNF76, BACH1 and E2F3 TFs and fewer instances of the E2F7, SMAD3 and SOX5 motifs (Figure S4.3). Comparisons of the models explaining either mean or noise again show that different genomic features are correlated with gene expression mean and noise.
4.3.5 **Intrinsic and extrinsic factors have similar effects on gene expression noise**

Expression noise caused by fluctuations in global factors affects all genes and is referred to as extrinsic noise, whereas intrinsic sources of noise are specific to individual genes [210, 306–310]. The correlation between identical reporter genes in the same cell measures the balance between extrinsic and intrinsic noise, with extrinsic factors increasing the correlation [204]. In SARGENT, the correlation between IRs in the same cells is a measure of extrinsic factors that affect noise across IR locations.

For our analysis of extrinsic noise we first identified IRs in the same clonal cells using the co-occurrence of locBCs between single cells (Figure 4.4A). We identified 192 clones, with a mean of three integrations per clone (Figure S4.4A). Of these 192 clones, 45 contain more than one integration (Figure 4.4B), making them suitable for an analysis of extrinsic noise. To validate the identified clones, we individually mapped IR barcodes in sixteen clones and found that 94% of the individually mapped IR locations could be uniquely assigned to an identified clone (Figure 4.4B).

We next asked if extrinsic factors also contribute to the observed gene expression noise. For each cell in a clone, we calculated the standard deviation relative to the mean of all IRs in that cell, which we define as the fluctuation index. Lower fluctuation indices indicate that the IRs in a clone fluctuate in sync (high extrinsic noise), while higher fluctuation indexes indicate that each IR varies independently (high intrinsic noise). To simulate intrinsic noise, we first shuffled the cell labels of all the IRs within a clone and computed a distribution of fluctuation indexes for the shuffled population. If all the measured noise was intrinsic, then the measured distribution would perfectly overlap the shuffled distribution. If all the measured noise was extrinsic, then all the cells would have fluctuation indexes of 0 (Figure S4.4B). We found that all clones show a distribution of fluctuation indexes that is lower than that of the shuffled distribution and above zero (Figure S4.4C). This suggests that some portion of the expression
Figure 4.4: SARGENT quantifies the extrinsic portion of expression noise. (A) Schematic for identifying different initial TRIP clones. (B) A network representation of the different clones identified, red nodes indicate IR locations that were independently validated by sequencing individual clones. (C) Pairwise expression for any two IR locations observed in the same cell. The trend along the diagonal suggests the existence of extrinsic noise, and the anti-correlation indicates the amount of intrinsic noise. (D) Quantification of intrinsic and extrinsic proportion of noise. Error bars from two technical replicates.

noise can be explained by extrinsic factors that impact all IRs within a cell in different genomic environments.

To quantify the contribution of intrinsic and extrinsic noise in each clone we employed an established statistical framework [312]. Using the pairwise IR single cell expressions for all clones that contain more than one IR as input, we found that intrinsic noise comprises approximately 54% of the total noise (Figure 4.4C-D). This analysis suggests that both the intrinsic chromatin and extrinsic cellular context explains about half of the total noise in each clone. These results show that SARGENT can quantify both intrinsic and extrinsic contributions to expression noise.
4.3.6 Cell substates are a source of expression noise

What cellular mechanisms control expression noise? We hypothesized that differences between cellular substates within isogenic populations are an important source of noise. Iso-
genic K562 cells transition between “stem-like” and “more differentiated” substates [249, 313]. The stem-like substate is marked by high CD24 expression and proliferates at a higher rate, which we hypothesized would contribute to extrinsic noise. This hypothesis predicts that the same IRs will have higher MIN in stem-like cells compared to more differentiated cells. To test this prediction we sequenced the single-cell transcriptomes associated with 356 of the 939 genomic locations in parallel with the IRs. Using the transcriptomes we identified clusters of cells with high CD24 expression and confirmed that these clusters had the signatures of high-proliferating cells (Figure S4.5A-B). We then calculated the expression mean and MIN for each IR location separately in the two substates. IR locations in the stem-like substate have higher mean and lower MIN (Figure 4.5A-B) suggesting that the global differences between the two substates are a source of MIN.

Given the differences in mean and MIN between the substates, the MIN of the IR locations in a given clone should be partly explained by the proportion of its cells in each substate. Consistent with this prediction, we found that clones with a higher proportion of cells in the stem-like substate have slightly higher average mean expression (Spearman’s $\rho = 0.22$, $p$-value = 0.008), and lower average MIN (Spearman’s $\rho = -0.27$, $p$-value = 0.0015) across all IRs in the clone (Figure 4.5C-D). We hypothesized that this was due to the slightly higher proliferation rates of cells in the stem-like phase. As expected, there are more cells in the S phase in the stem-like substate compared to the more differentiated state (Figure 4.5E). We then examined the differences of mean and MIN in different cell cycle phases and found that expression mean is higher and MIN is lower in the S phase compared to other phases (Figure S4.5C-D). These results suggest that differences in proliferation rates is an important source of extrinsic noise, and that SARGENT is a powerful tool to dissect the extrinsic sources of expression noise.

130
**Figure 4.5: Cellular information improves classification of low vs high MIN IR locations.** (A,B) Violin plots of expression mean (A) and MIN (B) at two substates (student t-test, \(* * * *: p < 0.0001\)), each dot is an IR location. (C,D) Scatterplots of proportion of cells in the “stem-like” substate against mean (C) and MIN (D). Each dot is the average mean expression or MIN from a clone. Line: linear fit with 95% confidence interval. (legend continued on next page)
Barplot of the fraction of cells in different cell cycle phases for cells in the “stem-like” substate and the “differentiated” substate (Binomial test: S phase $p < 2.2e-16$, G1 phase $p < 5.9e-5$, G2M phase $p < 2.2e-16$). The error bars are derived from the two replicates. (F) Weights of logistic regression model using extrinsic (cellular) features alone. (G) Addition of extrinsic features helps to improve the accuracy of the model. (H) Weights of logistic regression model using both intrinsic and extrinsic features. The most significant features are still the proportion of cells in the G2 phase and CD24+ phase. Red bars: $p$-value < 0.05; Pink bars: 0.05 < $p$-value < 0.1 from the logistic regression model.

4.3.7 Cellular information improves classification of low vs high MIN IR locations

Since extrinsic factors play an important role in determining expression noise, we trained a logistic regression model to predict MIN using three extrinsic features (proportion of cells in S, proportion of cells in G2, and proportion of CD24+ cells). Using only the global features, the model achieved 75% accuracy (albeit without a holdout set to test on due to the small numbers of locations with associated extrinsic features). This result implies that these cellular features explain a significant portion of the variance in MIN between high and low IR locations. The proportion of cells in G2 and the proportion of cells in the CD24+ state were significant predictors in this model. Being in G2 increases the probability of a high MIN IR location whereas having a higher proportion of CD24 cells reduced the probability of being a high MIN IR location (Figure 4.5F). When we combined the significant intrinsic features from the previous model with these extrinsic features, the model accuracy increased to 84% (Figure 4.5G). In the combined model, the extrinsic features have higher weights than the intrinsic genomic environment features (Figure 4.5H), suggesting that the cell-state information may play a larger role in regulating MIN compared to genomic environments.

We observed a similar role for extrinsic features in classifying IR locations with high mean levels from IR locations with low mean levels. The model accuracy for just the extrinsic feature model is 80% and increases to 89% for the combined model with both intrinsic and extrinsic
features (Figure S4.5E). In the combined model, the proportion of cells in the CD24 cell-state is the most highly weighted feature (Figure S4.5F). In contrast to the MIN model, the proportion of cells in the CD24 state increases the probability of being a high-mean IR location (Figure 4.5H, Figure S4.5F), which is consistent with our observations in Figure 4.5B and D. Thus, while cellular information plays an important role in gene expression regulation, these features have orthogonal impacts on expression mean and single-cell variability.

4.3.8 Effects of transgenes integration on endogenous genes

Finally, SARGENT can be used for purposes beyond studying gene expression noise. One such application is screening for “safe harbor” loci in the genome. To achieve safe and effective gene therapy, we need to identify genomic locations that have stable expression of the transgene of interest (high mean expression and low noise) and have minimal effects on endogenous gene expression. Historically, transgenes are often integrated into several known “safe harbor” loci [314]. Those loci are mainly located in the introns of stably expressed genes to prevent silencing. Because SARGENT can be used to measure gene expression mean, noise and endogenous gene expression simultaneously, we can leverage SARGENT to screen for potential safe harbors in a high-throughput manner.

We examined how our reporter gene integrations altered the expression of the gene into which it integrated. We focused on the 65 IR locations that are integrated into gene bodies. These integrations were distributed across different clones (Figure S4.6A) and should not be confounded by clonal effects. We calculated pseudo-bulk expression for each gene from clones that contain the integration and compared that to the expression from other clones that do not have the IR integration (Figure 4.6A). We found that in most cases (61/65), transgene integration does not alter the endogenous gene expression (Figure 4.6B). We also randomly shuffled the gene labels to compute the background differential expression, and found that there were no significantly differentially expressed genes once the labels were shuffled (Figure S4.6B). Among
Figure 4.6: SARGENT measures the insertion effect of a transgene. (A) Schematic for expression change detection in the transcriptome data. (B) Volcano plot of $\log_2$ fold change and $-\log_{10}(p$-value) from a Fisher’s Exact Test. Red dotted lines: cut off for fold change (0.5), cut off for $p$-value (0.05). (C) Barplots of difference of expression between genes without IRs (control) and genes with IRs (insert). The clone where the IR is integrated is indicated. Error bars are derived from two technical replicates.

The locations with significantly differentially expressed genes, 3 out of 4 IR integrations increases gene expression (Figure 4.6C), consistent with previous studies showing that the integration of a transgene often increases endogenous gene expression [315]. Taken together, our results suggest that most endogenous genes are not impacted by the integration of exogenous genes. This result illustrates that SARGENT could be a powerful tool to screen for “safe harbor” loci for transgene integration.
4.4 Discussion

Since the early single-cell studies showing the variability of gene expression in isogenic populations [204], many individual chromatin and sequence features have been suggested to modulate expression noise [193, 295, 296, 316]. However, there has yet to be a systematic study of the impact of different genomic features on large numbers of identical genes.

We developed SARGENT, a high throughput method to measure the expression mean and noise at different genomic locations in parallel. One key advantage of SARGENT is that the reporter gene used in all locations is identical, which allows us to isolate the effects of the genomic environments without being confounded by the effects of different promoters. We identified different chromatin marks that are associated with high or low MIN, and used a logistic regression model to identify features of the genomic environments that might control MIN. Our observations indicate that the features that control expression noise are independent of the features controlling expression mean. Several recent studies have developed tools for the orthogonal control of mean and gene expression noise [316–318]. To this end, our results suggest potential mechanisms that can be targeted for independent modulation of expression mean and single-cell variability.

We also quantified the extrinsic portion of expression noise and identified that the oscillation between a “stem-like” substate and a “differentiated” substate in K562 cells is an important source of extrinsic noise. Our data suggests that extrinsic noise might be more important in regulating MIN than genomic environments. This indicates that the regulation of noise of individual genes might be at the level of the promoter, rather than through its chromatin or genomic environment.

We envision that SARGENT will be a useful tool for other synthetic biology applications. While advances in genome engineering technologies now allow researchers to integrate transgenes at most desired genomic locations, the selection of appropriate sites for transgene overexpression remains non-trivial, with no location in human cells validated as a safe harbor locus.
This is mainly due to the lack of methods to systematically screen for loci that have high expression, low variability and do not impact cellular function. Here we showed that SARGENT can be used to read out a transgene’s impact on global expression as well as the endogenous gene that it is integrated into. Surprisingly, most endogenous genes are not impacted by the insertion of reporter genes. With SARGENT, we can quickly screen genomic locations to find the best locations for human transgene integration which will prove useful gene therapy applications.

More broadly, we envision that SARGENT will be a useful technology for many different applications including mechanistic studies of gene expression noise and synthetic biology applications. The 10x Genomics platform used in this study is limited by throughput, but improvements to scRNA-seq technologies will increase the scope of SARGENT. For example, coupling sci-RNA-seq [320] or SPLiT-seq [321] to SARGENT would allow for many more locations to be assayed in parallel. A larger goal will be to construct a detailed map of the MIN landscape across the genome, much like the maps of mean expression levels generated by ENCODE.

4.5 Methods

SARGENT library cloning

All primers and oligonucleotides used in this study are listed in Table S4.1. To clone the reporter gene for SARGENT, we first cloned a CMV-BFP reporter gene containing the 10x capture sequence 1 (CS1) into a piggyBac vector containing two parts of a split-GFP reporter gene [257]. When the reporter gene construct is integrated into the genome, the split-GFP combines to produce functional GFP, allowing us to sort for cells that have successful reporter gene integrations. We next added a library of random barcodes to the plasmid by digesting the plasmid with XbaI followed by NEBuilder HiFi DNA Assembly (NEB #E2621) with a
single-stranded oligo containing 16 random N’s (Location barcodes; locBC) and homology arms to the plasmid (CAS P57).

**Generation of cell lines for SARGENT**

K562 cells were maintained in Iscove’s Modified Dulbecco’s Medium (IMDM) + 10% FBS + 1% non-essential amino acids + 1% penicillin/streptomycin. We selected two K562 cell lines previously used in our lab that each contain a “landing pad” at a unique location with a pair of asymmetric Lox sites for recombination (loc1 - chr8:144,796,786, loc2 - chr11: 16,237,204; hg38 coordinates). Using these “landing pad” cell lines allows us to perform smFISH on the landing pad to directly compare SARGENT and smFISH results. For each cell line, we replaced the original landing pad cassette with the same reporter gene in the SARGENT library so that we can capture the reporters from the landing pad and reporters from other genomic locations in SARGENT using the same primers. Pool 1 was derived from the loc2 cell line, while Pools 2, 3 and 4 were derived from the loc1 cell line.

The SARGENT library and piggyBac transposase were co-transfected into K562 (LP cell lines) cells at a 3:1 ratio using the Neon Transfection System (Life Technologies). For each experiment, we transfected 2.4 million cells with 9μg of SARGENT library and 3μg of transposase. The cells were sorted after 24 hours for GFP-positive cells to enrich for cells that have integrated SARGENT reporters. We reasoned that ~100 single cells for each Integrated Reporter (IR) location would be required to obtain a good estimate of mean and variance. Each SARGENT experiment contains many single cell clone expansions: all the cells from the same clone share the same genomic integrations. Since we targeted approximately 20,000 cells per 10x run, the upper limit of the numbers of clones we can test in one experiment is 200. Because 10x also has a high dropout rate, we targeted 100 clones per experiment in order to ensure that we obtained high quality data. Each clone has an average of 5 integrations, which theoretically allows us to assay 500 IR locations in one experiment. Since the clones did not all grow at the same rate,
practically we obtained fewer than 500 IRs per experiment. 

For Pools 1 and 2, cells were sorted into pools of 100 cells each and allowed to grow until there were sufficient cells for RNA/DNA extraction and SARGENT experiments. Pool 3 contained the same cells as Pool 2, except that single cells were allowed to grow individually and pooled by hand just before the SARGENT experiments. This allowed for a more even representation of each individual clone (which contains unique integrations) in the final pool. For Pool 4, transfected cells were first sorted into 96-well plates with 2 cells/well and allowed to grow individually and 100 wells were manually pooled for SARGENT experiments. We used cells from Pool 4 to compute technical reproducibility.

**SARGENT integration mapping**

We harvested DNA from SARGENT pools using the TRIzol reagent (Life Technologies). To map the locations of SARGENT integrations, we digested gDNA for each pool with a combination of AvrII, NheI, SpeI and XbaI for 16 hours. The digestions were purified and self-ligated at 16°C for another 16 hours. After purifying the ligations, we performed inverse PCR to amplify the barcodes with the associated genomic DNA region (CAS P59 and P64). For each pool, we performed 2 technical replicates with 8 PCRs per replicate and pooled the PCRs of each replicate for purification. We then used 8ng of each replicate for further amplification with 2 rounds of PCR to add Illumina sequencing adapters (CAS P55 and P65). The sequencing library was sequenced on the Illumina NextSeq platform.

The barcodes of each read were matched with the sequence of its integration site. The integration site sequences were then aligned to hg38 using BWA with default parameters [322]. Only barcodes that mapped to a unique location were kept for downstream analyses.
ClampFISH

Single-molecule FISH was performed on the two “landing pad” locations that were in the original cell lines used for SARGENT (see Generation of cell lines for SARGENT above). ClampFISH probes for the reporter genes were designed using the Raj Lab Probe Design Tool. Each probe was broken into three arms to be synthesized by IDT. The 5’ of the left arm is labeled by a hexynyl group, and the 3’ of the right arm is labeled by NHS-azide. The right arm fragment was purified by HPLC. All three components were resuspended in nuclease-free H$_2$O to a concentration of 400μM. The three arms were ligated by T7 ligase (NEB # M0318), at 25°C overnight. then purified using the Monarch PCR & DNA cleanup Kit (NEB #T1030) and eluted with 40μl of nuclease-free water. After the ligation, each probe is stored at -20°C. ClampFISH was performed according to the suspension cell line protocol of clampFISH [323]. 0.7 million cells were collected and fixed in 2mL of fixing buffer containing 4% formaldehyde for 10 min, then permeabilized in 70% EtOH at 4°C for 24 hours. The primary ClampFISH probes were then hybridized for 4 hours at 37°C in the hybridization buffer (10% Dextran Sulfate, 10% Formamide, 2X SSC, 0.25% Triton X). After hybridization, cells were spun down gently at 1000 rcf for 2 min. Cells were washed twice with the washing buffer (20% formamide, 2X SSC, 0.25% Triton X) for 30 min at 37°C. The secondary probes were then hybridized to cells at 37°C for 2 hours and the cells were then washed twice with washing buffer for 30 min at 37°C. The primary and secondary probes are “clamped” in place through a click reaction (CuSO4 75μM, BTTAA 150μM, Sodium Ascorbate 2.5 mM in 2X SSC) for 20 min at 37°C. The cells were then washed twice in the washing buffer at 37°C for 30 min each wash. Then, the cells were hybridized with the hybridization buffer with tertiary probes for 2 hours at 37°C. We complete 6 cycles of hybridization for all our experiments. After the final washes, cells were incubated at 37°C with 100mM DAPI for 20 min, washed twice with PBS, resuspended in the anti-fade buffer, and spun onto a #1.5 coverslip using a cytospin cytocentrifuge (Thermo Scientific), mounted onto a glass slide, sealed with a sealant, and stored at 4°C.

139
SARGENT library using the 10x Genomics platform

Cell preparation

We used the Chromium Single Cell 3’ Kit (v3.1) from 10x Genomics for SARGENT. We followed the manufacturer’s instructions for preparing single-cell suspensions. We used a cell counter to measure the number of cells and viability, and used cell preparations with greater than 95% cell viability.

Cell barcoding and reverse transcription

We followed the manufacturer’s instructions with the following modifications in Pools 1-3: no 10x template switching oligo (PN3000228) was added to the Master Mix (Step 1.1). To correct for the missing volume, 2.4μl of H2O was added to the master mix per reaction. For pool 4, the template switching oligo was included as written. For the cDNA amplification (Step 2.2), no 10x provided reagents were used. Instead, a custom primer (CAS P20) was used with 14 cycles of amplification with the provided 10x protocol (Step 2.2 d). For the pool where we also sequenced transcriptomes (Pool 4), we followed the 10x protocol as written for cDNA amplification.

Barcode PCR and library preparation

We performed nested PCRs to amplify barcodes from 10x cDNA. For pools 1-2, PCR library construction was split into two pools for amplification of transcripts captured by capture sequence 1 and poly(A) respectively. Both PCR reactions were done with 2μl purified cDNA, 2.5μl 10μM reporter-specific forward primer (CAS P45), 2.5μl 10 uM poly(A) (CAS P20) or capture sequence adapter-specific primers (CAS P32), and 25μl Q5 High Fidelity 2X Master Mix (NEB #M0492) in 50μl total volume with 10 cycles amplification. The PCRs were then purified with Monarch PCR & DNA Cleanup Kit (NEB #T1030) and Illumina adapters were
added in another 2 rounds of PCR, with a PCR purification step with the Monarch kit between
PCRs. For poly(A) amplicons, we used CAS P42 and CAS PP2, followed by CAS P48 and CAS
PP4. For capture sequence amplicons, we used CAS P41 and CAS CS2, followed by CAS P48
and CAS CS4. The reactions were then pooled and purified with SPIselect Beads (Beckman
Coulter) at 0.65x volume. For pool 4, we performed the PCRs for the poly(A) fraction using 2μl
purified cDNA as described above, but not the capture sequence transcripts.

SARGENT data processing

Read parsing

We first identified the reads that match the constant sequence in our reporter gene. We used
two versions of constant sequence to match against, depending on if the read was captured
using the poly(A) sequence on the mRNA or the capture sequence specific to the 10x beads. We
used a fuzzy match algorithm to capture reads that have a mismatch at these positions due to
sequencing error. From each read, we parsed out the cell barcode, 10x UMI and locBC. We then
collapsed reads with identical cell barcodes, UMI and locBCs into one “trio” and kept track
of the number of reads supporting each trio. For downstream analysis, we filtered out trios
with low numbers of supporting reads since these are likely to be enriched for PCR artifacts.
We next processed the trios to error correct the cell barcodes and locBCs before estimating the
mean and variance.

Barcode error correction

To correct for PCR artifact and sequencing errors, a custom script was used to error-correct
for 10x cell barcodes. Briefly, we first acquired the empirical distribution of the Hamming
distances among observed 10x cell barcodes. We found that more than 99% of 10x cell barcodes
are more than 6 hamming distances away from each other, making error correction a feasible
approach to denoise the data. We first identify cell barcodes that match perfectly to the 10x cell barcode whitelist, then we order them based on their abundance. The cell barcodes that are not in the whitelist are then compared to the ordered whitelisted cell barcodes, if the Hamming distance between the non-whitelisted cell barcodes is within 2 Hamming distances of a whitelisted cell barcodes, we correct the unwhitelisted cell barcode. With cell barcode correction, we recovered ~12% of reads that would have been discarded.

Due to the random synthesis of the locBC, a slightly different approach was taken for error correction for the locBCs. Briefly, all the locBCs are ranked based on abundance. Starting from the most abundant barcode, we look for locBCs that are within 4 Hamming distance to that barcode and correct them. We then remove that barcode and any corrected barcodes, and repeat this process until we have iterated through all locBCs.

Calculating mean and variance of each IR

We filtered out cells that had less than 5 IR integrations (locBCs) and less than 10 UMIs. We also filtered out locBCs that were seen in less than 5 cells and UMIs that had less than 2 supporting reads. We then computed the number of UMIs per locBC in each cell to calculate the expression level of each locBC. For each locBC, mean expression was calculated as the average UMI count across all cells that expressed that locBC. Expression variance was calculated as the variance in UMI counts across all cells that expressed that locBC.

Mean-independent noise (MIN) metric

In order to remove the effect of the mean on the variance we first fit a linear model: $\log_2(\text{variance of IR location}) \sim \log_2(\text{mean of IR location})$ for each experimental pool and used the residuals of the model as the mean-independent noise metric. For each IR location, the MIN is the residual variance after removing the effect of the mean. Analyses of genomic environment effects on mean-independent noise
Chromatin environment association with mean/MIN

We downloaded the Core 15-state chromHMM annotations for K562 cells from the Roadmap Epigenomics Project [99]. We then collapsed similar annotations and overlapped the IR locations with the corresponding annotation using the GenomicRanges R package [262].

We split the IRs into locations with high vs low mean or high vs low MIN respectively. We then downloaded histone ChIP-seq datasets from ENCODE [38] and plotted the signals 10kb surrounding each class of IRs using the ComplexHeatmap package in R [260].

To look for enriched TF motifs we first downloaded all human motifs from the HOCOMOCO v11 database. We then filtered the motifs for TFs that are expressed (FPKM ≥ 1) in the K562 cell line using whole-cell long poly(A) RNA-seq data generated by ENCODE (downloaded from the EMBL-EBI Expression Atlas). We then used the STREME package [324] (MEME suite 5.4.1) with sequences of 1kb surrounding each IR to identify enriched de novo motifs in high or low MIN regions, using the other class as the control set of sequences. We then took the top 2 motifs for each bin and matched it against a list of TFs expressed in K562s using TOMTOM [325] (MEME suite 5.4.1). We reported the top 6 TOMTOM matches.

K562 Hi-C

We performed Hi-C on wild-type K562 cells with the Arima Hi-C kit (A510008) according to the manufacturer’s protocols (3 replicates, 870 million reads total). The reads were then processed with the Juicer pipeline [326] to generate HiC contact files for each replicate. We then used the peakHiC tool [294] to call loops from each IR with the following parameters: window size = 80, alphaFDR = 0.5, minimum distance = 10kb, qWr = 1. Using these parameters each IR was looped to a median of 3 regions (range 0-7).
Logistic regression model for intrinsic and extrinsic features associated with MIN

We used histone ChIP-seq and ATAC-seq datasets from ENCODE [98] and overlapped their signals with each IR using used bedtools v2.27.1 [291]. For all features we considered the 20kb upstream and downstream of each IR respectively. For each histone modification, we computed the mean ChIP signal around the IRs. For ATAC-seq, we calculated the total number of peaks with the bedtools map count option. To look for TF motifs we counted the numbers of each motif for TFs expressed in K562s (see above) in each surrounding IR sequence using FIMO [269] (MEME suite 5.0.4). Because this resulted in a long list of TFs we further filtered the TFs to include only those with a significant correlation with MIN levels in the regression model. To determine the numbers of enhancers interacting with each IR we annotated the loops called from peakHiC above with chromHMM enhancer annotations using the GenomicInteractions R package [327] and counted the number of enhancers.

For the extrinsic features, we calculated the proportion of cells in the “stem-like” substate and “differentiated” substate and different cell cycle phases based on the barcodes that appeared in those substates. We removed IR locations that have less than 30 cells in any of the substates.

We used the glm function in R (version 3.6.3) to fit logistic regression models. We separated the IR locations into top 20% MIN and bottom 20% MIN and used logistic regression to classify locations. We first fit a model with just local sequence features (chromatin modifications, number of TF motifs, number of loops, whether the IR location is in a gene, GC content and the number of ATAC-seq peaks). We used data from one experiment for training the model and used data from another experiment as a holdout set of data to estimate the performance of the classifier. We next fit a model with cellular information for each IR location: proportion of cells with data for the IR location in S phase of the cell cycle, in G2 phase and the proportion of cells that are in the “stem-like” substate of K562 cells [313]. Lastly, we fit a model that incorporated the extrinsic features and the significant predictors from the intrinsic features model.
Transcriptome analyses associated with SARGENT

Processing the single-cell transcriptome data

The single-cell RNAseq data was processed with CellRanger 6.0.1 and SCANPY 1.9.1 [328]. Briefly, the raw reads were processed with the standard single-cell expression cell line pipeline. The resulting expression matrix was then imported into SCANPY for further visualization and clustering.

Identifying single cell clones

We identified the individual clones for Pool 4 which contained cells that grew out of 100 two-cell clones. Since most of the clones will have unique integrations into unique genomic locations, the cells that grew out from the same clone will have identical unique sets of locBCs. Due to the dropout rates associated with scRNAseq methods, not all barcodes will be present in all cells, nor will the cell barcodes be uniquely linked to correct sets of locBCs. To identify the barcodes belonging to the same clone, we first recorded locBCs that are linked by a given cell barcode. We then filtered the locBC list associated with a given cellBC based on the number of UMIs associated with these locBC. At this step, we used a knee point detection algorithm [329] that automatically detects the inflection point of the ordered UMI counts histogram. After filtering for locBCs that appear in more than 5 cells, we constructed a clonal graph by linking locBCs that co-occur in the same cells.

Validation of individual clones

We extracted gDNA from 16 clones that were grown out from Pool 4. We then amplified the barcodes from each clone using Q5 High Fidelity 2X Master Mix (NEB #M0492) with primers specific to our reporter gene (CAS P58-59). For each clone, we performed 4 PCRs and pooled the PCRs for purification. 4ng from each clone was then further amplified with 2 rounds of
PCR to add Illumina sequencing adapters (CAS P60-63). The barcodes were sequenced on the Illumina NextSeq platform.

**Estimating intrinsic vs extrinsic noise**

To understand how cellular environments affect IR expression, we computed the mean and standard deviation from all IR locations in the same cell. Since standard deviation is expected to increase with mean, we calculated the standard deviation/mean for each cell, which we termed the fluctuation index. To establish the null distributions, we randomly shuffled the cell labels for each clone and computed fluctuation indices for the shuffled cells.

Intrinsic and extrinsic noise were estimated using the statistical framework developed for the dual-reporter experiment [312]. In our experiment, single-cell expression differences among IR locations are treated as the intrinsic portion of the noise. We first extracted the pairwise expression level for IR locations in every single cell, then applied the above framework. The derivation is abbreviated and can be found in the original publication. Briefly, let $C$ denote the expression for the first locBC in the cell, $Y$ denote the expression for the second locBC in the cell and $n$ denote the number of cells.

Let $\eta_{\text{ext}}$ denote the extrinsic noise, and it can be calculated as:

$$
\eta_{\text{ext}} = \frac{1}{a\bar{C}\bar{Y}} \left( \sum_{i=1}^{n} C_i Y_i - n\bar{C}\bar{Y} \right)
$$

where

$$
a = (n - 1) \left( 1 + \frac{1}{n} \right) + \frac{1}{\rho^2}
$$

$$
\rho = \frac{\text{Cov}[C, Y]}{\sqrt{\text{Var}[C]\text{Var}[Y]}}
$$

Similarly, let $\eta_{\text{int}}$ denote the intrinsic noise, and it can be calculated as:

$$
\eta_{\text{int}} = \frac{1}{2a\bar{C}\bar{Y}} \left( \sum_{i=1}^{n} (C_i - Y_i)^2 - n(\bar{C} - \bar{Y})^2 \right)
$$
where
\[
a = \frac{2n^3 - 7n + 6}{2(n^2 - n)} + \frac{2 - n}{n^2 - n} \frac{\rho}{1 - \rho} + \frac{1}{2(n^2 - n)} \left( \frac{\rho}{1 - \rho} \right)^2
\]

\[
\rho = \frac{\text{Cov}[C, Y]}{\sqrt{\text{Var}[C]} \sqrt{\text{Var}[Y]}}
\]

**Cell substate impact on expression mean and noise**

To compute cell substate specific expression mean and noise at different genomic locations, individual cells were assigned a cell cycle phase of G1, S, or G2/M using a previously reported set of cell-cycle specific marker genes with SCANPY 1.9.1 [328]. For the stem-like substate analysis, we clustered cells based on their transcriptomes and assigned cells in the CD24 high cluster as CD24+ cells [313]. To ensure an accurate measurement of expression mean and noise, genomic locations with less than 15 cells in any phase were excluded from the cell cycle analysis. Based on this filtering criterion, 345 out of 939 genomic locations were used for this analysis. To determine the impact of cellular substates on gene expression noise, we calculated the proportion of cells in different cellular substates for each clone. For each clone, we also calculated the average mean and variance of all the IRs in that clone.

**Transgene integration analysis**

To examine whether the integration of a trans-gene alters endogenous gene expression, we first identified IR locations that were integrated into a gene body. Since the IR insertion only occurs in a single clone, we computed pseudobulk expression from cells in the clone using decouplerR 1.1.0 [330]. We then randomly sampled the same number of cells from all the other clones and used the pseudobulk expression from these cells as wild-type expression. To determine whether the expression in the IR clone is significantly different from wild-type expression, we computed the \( p \)-value of differential expression using a Fisher’s exact test.
4.6 Supplementary Figures

Figure S4.1: smFISH corroborates with SARGENT measurements. (A,B) Mean and noise levels of two IR locations (loc1 (A) and loc2 (B)) measured by SARGENT. Values were normalized (Z-scored) for comparison across different experiments. (C,D) Mean and noise levels of the same two IR locations (loc1 (C) and loc2 (D)) measured with smFISH. Error bars represent one std from two biological replicates.
Figure S4.2: Measurements of mean-independent noise (MIN) across different chromosomal environments. (A) IR locations are distributed all throughout the genome. Each black bar above the ideogram represents a separate integration. (B) IR locations are distributed across different chromatin types. (C,D) Expression mean is well correlated with fano factor (C) and CV² (D).
Figure S4.3: Model results for high mean vs low mean. Weights of features from the gene expression mean model using only intrinsic genomic features. Red bars: $p$-value < 0.05; Pink bars: $0.05 \leq p$-value < 0.1 from the logistic regression model.
Figure S4.4: Clonal identification allows for separation of intrinsic and extrinsic noise. (A) Histogram of number of integrations per clone. (B) Mock histogram showing the expected distributions if noise was either all intrinsic or all extrinsic. (C) Histogram of measured fluctuation indices (defined as the standard deviation/mean of all IRs in a cell) for 10 random clones. The shuffled distribution represents the distribution after cell labels have been shuffled, which simulates the case when all noise is intrinsic.
**Figure S4.5: Cell cycle and CD24 states partially explain extrinsic noise.** (A) UMAP clustering reveals CD24+ cell population. (B) CD24+ cells express stemness and proliferation marker gene GATA1. (C, D) Violin plots of MIN (C) and mean (D) levels in different phases of the cell cycle. *P*-values were calculated using the Mann-Whitney-Wilcoxon test. Legend: ∗: 0.01 < *p*-value ≤ 0.05, ∗∗: 0.001 < *p*-value ≤ 0.01, ∗∗∗: 0.0001 < *p*-value ≤ 0.001, ∗∗∗∗: *p*-value ≤ 0.0001. ns: not significant. (E) Gene expression mean model improves after the addition of extrinsic features. (F) Weights of features from the mean model using both intrinsic genomic and extrinsic features. Red bars: *p*-value < 0.05; Pink bars: 0.05 ≤ *p*-value < 0.1 from the logistic regression model.
Figure S4.6: IR integrations have little impact on endogenous expression. (A) Bar plot number of IRs in endogenous genes per clone. (B) Shuffling IR-endogenous gene labels results in no differentially expressed genes.
### 4.7 Supplementary Tables

**Table S4.1:** Primers used in this study

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAS P20</td>
<td>CTACACGACGCTCTCTTCCGATCT</td>
</tr>
<tr>
<td>CAS P32</td>
<td>GTCAGATGTGTATAAGAGACAG</td>
</tr>
<tr>
<td>CAS P41</td>
<td>GTGACTGGAGTTTCAGACGTTGCTCTTTCCGATCTTTTCTCGCTTC-GAGTCTAGA</td>
</tr>
<tr>
<td>CAS P42</td>
<td>GTGACTGGAGTTTCAGACGTTGCTCTTTCCGATCTTAGCTCTCGCTTC-GAGTCTAGA</td>
</tr>
<tr>
<td>CAS P45</td>
<td>TCTCGCTTCGAGTCTAGA</td>
</tr>
<tr>
<td>CAS P48</td>
<td>CAAGCAGAAGACGCGATACGAGATNNNNNNNNNGTGACTG-GAGTTCAGAC</td>
</tr>
<tr>
<td>CAS P55</td>
<td>GTGACTGGAGTTTCAGACGTTGCTCTTTCCGATCTTTAGGACCGGCTCT-CAAAGC</td>
</tr>
<tr>
<td>CAS P57</td>
<td>GGGGCACAAGCTTTAATTGAGAGGCTCTCCTGCTTGACTCTAGA-GANNNNNNNNNNNNNNCTCAGAGTGGTGGCAGGCCCCCTT</td>
</tr>
<tr>
<td>CAS P58</td>
<td>TGCCTGGGCTCTACTATGTC</td>
</tr>
<tr>
<td>CAS P59</td>
<td>AAGCGCCAGGGTTTCCCCAA</td>
</tr>
<tr>
<td>CAS P60</td>
<td>CTTTCCCCTACACAGCAGCTCTTTCCGATCT(N[1-4])CTCGCTTCGAGTCTAGA</td>
</tr>
<tr>
<td>CAS P61</td>
<td>GTGACTGGAGTTTCAGACGTTGCTCTTTCCGATCTGCCAGGTTTCC-CACA</td>
</tr>
<tr>
<td>CAS P62</td>
<td>AATGATACGGCGACCACCCGAGATCTACACNNNNNNNACACTCCTTCC-CTACACGACGCT</td>
</tr>
<tr>
<td>CAS P63</td>
<td>CAAGCAGAAGACGCGATACGAGATNNNNNNNNGTGACTG-GAGTTCAGACGTG</td>
</tr>
<tr>
<td>CAS P64</td>
<td>GTACCACATGCCGCGCATGATTATCTTTAAGCTACGTACCA</td>
</tr>
<tr>
<td>CAS P65</td>
<td>ACGACGCTCTCCTCCGATCTCTGCA(N[1-4])GTACGTCACAATATGATTATCTTTCTAG</td>
</tr>
<tr>
<td>CAS PP2</td>
<td>ATCTACACTCTTTTTTACAGACGCTCTTC</td>
</tr>
<tr>
<td>CAS PP4</td>
<td>AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACCA</td>
</tr>
<tr>
<td>CAS CS2</td>
<td>CGAGATCTACACTCGTGCCGAGTCAGATGTGTATAAGAGACAG</td>
</tr>
<tr>
<td>CAS CS4</td>
<td>AATGATACCGCGACCACCGAGATCTACACTCGTGCG</td>
</tr>
</tbody>
</table>
Chapter 5: Discussion

In this thesis, I focused on studying how genomic environments and cis-regulatory sequences work together to regulate gene expression. To learn the systems-level rules that govern gene expression, we need methods that can test large numbers of sequences in parallel in the context of the genome. Here I achieved this by adapting and developing new high-throughput methods to systematically integrate reporter genes into diverse locations across the genome and measure their activities in parallel. In Chapter 2, we tested the compatibility of core promoters with thousands of genomic environments. We find that core promoters are modular components of the genome that are broadly compatible with the CRSs and chromatin in all genomic environments. In Chapter 3, we developed a new method, MPIRE, to measure the activities of cis-regulatory sequences at defined locations across the genome. We use this method to assay the functions of insulators in diverse genomic environments. Finally, in Chapter 4, we developed a single-cell method, SARGENT, to measure the expression noise of reporter genes integrated in hundreds of genomic locations. We identified genomic and cellular features correlated with changes in gene expression noise and the utility of SARGENT to screen for “safe harbor” loci. This body of work contributes new methods and results that have broad implications in the gene regulation field.

5.1 Core promoters are broadly compatible with genomic environments

Since the discovery that core promoters contain diverse, wide-ranging sequence features beyond the TATA box, it has been hypothesized that core promoters with different features
respond to different sets of enhancers or chromatin environments. This compatibility helps determine E-P specificity in the genome and allows for cell-type specific expression of the same core promoter sequence. As discussed in the introduction, there are many examples from studies of individual core promoters that seem to support this idea. More recently, a study in *Drosophila* showed that core promoters can be separated into two classes, housekeeping (hk) and developmental (dev), using STARR-seq (self-transcribing active regulatory region sequencing), a type of plasmid-based MPRA [130]. These two classes of core promoters respond to distinct sets of enhancers containing different sequence features, and these different enhancers also distinct sets of core promoters [132]. The same group further showed that these classes of core promoters in humans respond to distinct sets of cofactors [30]. Together, these studies provided strong support for the core promoter compatibility hypothesis.

When we started our study, the high-throughput studies of promoter compatibility had only been performed on episomal plasmids. Thus, we sought to test the promoter compatibility hypothesis in human cells using patchMPRA, a method previously developed by our lab to integrate large libraries of CRSs into fixed, landing pad locations in the genome [243]. We reasoned if each core promoter only responded to specific enhancers, then the core promoter activities would differ in different genomic environments containing different enhancers and chromatin features. Surprisingly, we found that core promoter activities were independent of the genomic environment, indicating that promoters are compatible with all tested environments. We supported our findings by showing that the activities of six core promoters (containing different sequence features) are also scaled by thousands of locations in the genome. This genomic environment scaling effect was non-linear and depended on the strengths of the core promoters and genomic environments. Overall, our results contradicted the core promoter compatibility hypothesis and provided instead supports a model where the strength, not sequence, of the core promoter determines its activity in the genome.

After our results were published, two other groups performed plasmid-based MPRAs to
explicitly test the enhancer-promoter hypothesis [187, 188]. Both groups focused on subsets of human [187] or mouse [188] promoters and putative enhancers and tested all pairwise combinations for compatibility. While the results were approached slightly differently by each group, the data from both papers support the idea that enhancers and promoters are broadly compatible and showed very little evidence of biochemical specificity, again in contrast to the earlier studies in flies. Indeed, a simple multiplicative model of enhancer × promoter was sufficient to explain 80-90% of the variance observed in both experiments [187]. These results support our model that core promoters are modular, independent CRSs in the genome.

Notably, Bergman et al. [187] separated the enhancers and promoters into two classes each to explain the residual variance. The authors showed that the E1 class of enhancers activated P1 promoters more strongly than P2 promoters, while E2 enhancers activated all promoters similarly, and interpreted this to mean that there was some, albeit subtle, compatibility between E1 and P1 promoters. However, a simpler explanation for their results is that P1 promoters are weaker and inherently more “activatable” by strong promoters, and E1 promoters are stronger than E2 promoters. Using this framework, their results agree well with our model that interactions between enhancers and promoters are simply a function of strength Figure 2.5E rather than biochemical compatibility, which was reassuring since our study did not explicitly test the effects of enhancers. The concordance between our results and Bergman et al. also suggest that our assumption that different genomic environments represent diverse sets of enhancers is valid. Additionally, by modeling the response of promoters to E-P interactions, Xiao et al. showed that the sigmoidal relationship between E-P contact and promoter output (also proposed by our results in Figure 2.5A) can give the illusion of E-P specificity. A weak promoter could require a strong enhancer to drive detectable gene expression, while a stronger promoter already driving high basal expression would respond to weaker enhancers. Together, these results potentially reconcile the observations of E-P specificity in individual cases with the general lack of specificity in high-throughput studies.
However, the high-throughput fly and mammalian studies still contradict each other. At first glance, these differences could simply be due to species-specific differences in gene regulation. Perhaps, at some point, flies evolved to develop compatibility between enhancers and promoters. This could be because fly genomes are more compact and each promoter may come into contact with many more non-specific enhancers relative to the mammalian genome. However, a different possibility may also be the differences in assays used to study E-P compatibility. The fly experiment was done with STARR-seq, where the genome is sheared into random fragments of 500-700bp in length and tested for activity [130]. Activity in STARR-seq is determined by calling peaks across the random fragments. On the other hand, the mammalian assays are done with curated sets of enhancers of 264bp [187] or 450bp [188], and activity is measured directly as read counts derived from each E-P pair. While similar, these differences could be sufficient to lead to the divergent observations. Moving forward it will be important to test more pairs of enhancers and promoters in the same system to distinguish between these two possibilities.

5.2 Insulators organize the genome to modulate specific E-P interactions

Unlike core promoters, insulators are a less well-studied class of CRSs. The one defining feature that connects all insulators is that they can alter gene expression without directly influencing promoter activity. Beyond that, insulators have been found to protect against chromosome position effects, block enhancers, act as a barrier against heterochromatin and regulate 3D genome structure. These functions are not mutually exclusive, but it is unclear which function(s) insulators perform in different genomic contexts. Moreover, individual perturbations of insulator function in the genome has given rise to all kinds of effects on gene expression [90, 91], making it difficult to come up with a consistent model of how insulators impact gene expression.
To understand insulator activity at scale, it is imperative to measure the impact of insulators on gene expression across diverse genomic contexts. However, there was no existing method that would allow us to measure reporter genes with and without insulator sequences at the same locations across the genome. We developed MPIRE to do exactly that. We show that MPIRE is a robust and reproducible method for measuring CRS activity across the genome. In our study, we tested the effects of cHS4, A2 and ALOXE3, three insulators that have been shown to have insulator activity in the canonical EB assay [56, 70, 71]. cHS4 and ALOXE3 have also been shown to act as barriers against heterochromatin silencing and protect against position effects [70, 92]. Our results show that insulator activity is exquisitely context specific and reconcile the differences between genome-wide analyses and analyses of individual insulators. We found that insulators only function in small numbers of genomic locations, which explains why depletion of insulator-binding proteins such as CTCF or Cohesin have minimal changes on gene expression. At the locations that insulators do function in, different insulators preferentially block specific enhancers, and only ALOXE3 protects against heterochromatin silencing. These results explain why one-at-a-time experiments can sometimes show that insulators have large and specific effects on gene expression.

We also showed that at most locations where the insulators are functioning, they are dependent on either the CTCF (cHS4, A2) or the B-box (ALOXE3) motifs, as we expected from previous studies. However, one mystery that we never quite solved is that the mutated insulators appear to also be gaining function at other locations in the genome. Besides the technical caveat that these might be false positives called from our analyses, a biological explanation might be that losing CTCF or PolIII binding somehow changes the properties of the insulator and causes it to be active at other sites in the genome. The small number of locations that showed such gain of activity did not provide enough power for us to figure out what the mutated insulator was doing, but screening more locations across the genome might help us solve this puzzle. For ALOXE3 in particular, it will also be interesting to separate the contributions of PolIII binding
and transcription. Will replacing the B-box with a PolII promoter restore ALOXE3’s activity? More systematic mutational scans will help us define the sequence features that drive insulator activity.

While we were conducting this study, Ribeiro-dos-Santos et al. performed a variant of the TRIP assay with different insulators, but could not test the constructs at the same locations and could only rely on genome-scale correlative observations [272]. Reassuringly, they also showed that insulator activity is very context-dependent. Further studies applying MPIRE to more insulator constructs combined with other types of functional experiments would be important to learn the rules governing insulator activity in the genome. Eventually, our goal would be to be able to predict where and how insulators would function in the genome, and understand the sequence features contributing to insulator activity to further predict how variants would impact insulator activity and downstream gene expression.

We acknowledge that a key caveat with using MPIRE to test insulator activity is the lack of the insulator at the 3’ end of the reporter construct. In trying to avoid having the insulator in the 3’UTR of the reporter gene we could not perfectly recapitulate the position effect blocking assays. In future experiments, it might be possible to include the second insulator with a few modifications. First, assuming that the insulator have the same impact of all mRNA copies regardless of where in the genome it is transcribed, we can measure the effect of the additional insulator on mRNA stability and normalize for it when calculating expression levels. Second, we can attempt to flank the second insulator with splice sites, so that the insulator will only be present in the DNA but not the mRNA. These modifications will take some optimization but could contribute to a more thorough understanding of insulator effects across the genome.

While we developed MPIRE to study insulator activity, we envision that MPIRE can be used for a multitude of other questions. For example, MPIRE can be used to study differences in TF binding across different chromatin contexts, or to study how splicing or DNA damage changes at different genomic locations. Additionally, MPIRE can be adapted to other readouts.
Instead of measuring the RNA barcodes, we can perform chromatin immunoprecipitation (ChIP) or Cut&Tag and enrich for and sequence the MPIRE barcodes to get the chromatin modifications/TF binding across all LPs in parallel. Another particularly exciting direction that would be relevant to insulator activity is to perform a 4C-type experiment using the LPs as bait. Since the MPIRE barcodes would also be in the sequencing read, we can determine the interacting partners of all the LPs at the same time. Finally, MPIRE can be used as a highly efficient tool to screen genomic locations for transgene expression. Thus, we believe that MPIRE will be an important tool for the gene regulation and synthetic biology community.

More broadly, the results from these studies contribute to ongoing discourse about how promoters are specifically expressed in the genome. Our lab had previously shown that enhancer activity is independent of the genomic environment, indicating that enhancers are also modular components of the genome. Given that core promoters also seem to be broadly compatible with genomic environments, our results suggest that insulators might play a more important role in directing specificity. Complementing our studies with genome architecture and TF binding assays will help us gain a more complete picture of how CRSs and chromatin environments function to regulate gene expression.

### 5.3 Regulation of gene expression noise

Studying cell-to-cell variability in gene expression is fascinating for two reasons. Noise in gene expression can be exploited to generate diverse cell populations from the genome, which is necessary for cell type specification in development [295]. On the other hand, quantifying and measuring noise requires single-cell measurements and analyses, which can reveal features of gene expression that could not be accessed with bulk analyses [235]. By understanding how sequence and chromatin features impact gene expression noise, our goal is to eventually identify the transcriptional steps controlled by different features for a mechanistic understanding of gene regulation.
To achieve this goal we require methods that will allow us to measure the impact of different genomic features on a large number of genes. scRNA-seq methods can in theory achieve this, but the measurements are confounded by the different endogenous promoters. In this study we develop SARGENT, which extends scRNA-seq to measure the noisiness of integrated reporter genes that contain identical promoters. The only variable that is changing is the genomic position of the reporter gene, allowing us to directly quantify the impact of genomic position. Using SARGENT, we find that there there are chromatin and sequence features that are correlated with differences in gene expression noise, and that extrinsic cellular features also contribute significantly. However, as previously shown, most of the variance is explained by mean levels of expression, which suggests that specifically modulating noise independently of the mean is not a widespread phenomenon at least in K562. Further studies in other less differentiated cell types where noise might be more important for cell type specification and differentiation may lead to more interesting results. Our results provide a method and a starting point for dissecting the contributions of various genomic features to noise.

SARGENT has a few limitations that can be improved upon. First, the low capture efficiencies of mRNA molecules means that we are only quantifying a small subset of the total expressed mRNAs. Despite our efforts to specifically enrich for the reporter gene, we are significantly undercounting the number of mRNA molecules per cell when compared to smFISH methods. As a consequence, we were unable to apply rigorous, quantitative models to fit our data. Second, we are limited by the number of cells that can be sequenced, which also limits the number of locations that can be tested in one experiment. In our study, we measured ~900 locations, which does not quite cover enough genomic environments to get high statistical power. As single-cell technologies improve, we will be able to gain a more quantitative understanding of gene regulation with SARGENT.
5.4 Conclusion

Overall, this work contributes to our understanding of how the different CRSs and chromatin modifications that comprise genomic environments regulate gene expression. The methods introduced in this study will also help study the interplay between different genomic features. The widespread compatibility between genomic environments and enhancers/core promoters suggest that gene regulation is not as complex as we might have thought. A better understanding of how insulators generate different outcomes in different contexts will allow us to learn generalizable rules and eventually develop better predictive models of gene expression.
References


Appendix A: Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mESCs

This work was performed in collaboration with Dana King, James Shepherdson, David Granas, Brett Maricque and Barak Cohen. I designed and cloned the spacer library and helped with the spacer library MPRA. I also performed the analysis for that section. This study was published in 2018 in *eLife* (doi: 10.7554/eLife.41279).
Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells

Dana M King¹,²†, Clarice Kit Yee Hong¹,², James L Shepherdson¹,², David M Granas¹,², Brett B Maricque¹,²‡, Barak A Cohen¹,²*

¹Edison Center for Genome Sciences and Systems Biology, Washington University in St. Louis, St. Louis, United States; ²Department of Genetics, Washington University in St. Louis, St. Louis, United States

Abstract In embryonic stem cells (ESCs), a core transcription factor (TF) network establishes the gene expression program necessary for pluripotency. To address how interactions between four key TFs contribute to cis-regulation in mouse ESCs, we assayed two massively parallel reporter assay (MPRA) libraries composed of binding sites for SOX2, POU5F1 (OCT4), KLF4, and ESRRB. Comparisons between synthetic cis-regulatory elements and genomic sequences with comparable binding site configurations revealed some aspects of a regulatory grammar. The expression of synthetic elements is influenced by both the number and arrangement of binding sites. This grammar plays only a small role for genomic sequences, as the relative activities of genomic sequences are best explained by the predicted occupancy of binding sites, regardless of binding site identity and positioning. Our results suggest that the effects of transcription factor binding sites (TFBS) are influenced by the order and orientation of sites, but that in the genome the overall occupancy of TFs is the primary determinant of activity.

Introduction

Independence versus interaction of transcription factor binding sites

Enhancers are composed of combinations of transcription factor binding sites (TFBS). An important question is: to what extent do TFBS act independently within enhancers and to what extent do specific interactions between transcription factors (TF) underlie enhancer function? Independence suggests a modular genome in which the effects of multiple binding sites are predictable from their individual effects. Interactions, such as cooperativity between TFs, cause the effect of multiple TFBS to be more (or less) than the combination of their individual effects. Constructing models that predict the expression of genes based on the TFBS composition of their surrounding regulatory DNA will require understanding the degree to which sites function independently and how interactions between sites contribute to the activity of regulatory sequences.

Regulatory grammar

The extent to which TFs function either independently or through interactions should be reflected in the cis-regulatory grammar of TFBS, defined as the ways that the order, orientation, spacing, and affinity of binding sites impact the activity of enhancers. If TFs function independently then we do not expect strong constraints on the positioning of their binding sites within regulatory elements. If TFs function mostly through interactions with other TFs that require a precise geometry, then we expect strong biases in the positioning of TFBS within regulatory elements. At least three models make predictions of how grammar might influence enhancer activity, the billboard model, the...
**elife digest** Transcription factors are proteins that flip genetic switches; their role is to control when and where genes are active. They do this by binding to short stretches of DNA called *cis*-regulatory sequences. Each sequence can have several binding sites for different transcription factors, but it is largely unclear whether the transcription factors binding to the same regulatory sequence actually work together.

It is possible that each transcription factor may work independently and there only needs to be a critical mass of transcription factors bound to throw the genetic switch. If this is the case, the most important features of a *cis*-regulatory sequence should be the number of binding sites it contains, and how tightly the transcription factors bind to those sites. The more transcription factors and the more strongly they bind, the more active the gene should be. An alternative option is that certain transcription factors may work better together, enhancing each other’s effects such that the total effect is more than the sum of its parts. If this is true, the order, orientation and spacing of the binding sites within a sequence should matter more than the number.

One way to investigate to distinguish between these possibilities is to study mouse embryonic stem cells, which have a core set of four transcription factors. Looking directly at a real genome, however, can be confusing and it is difficult to measure the effects of different *cis*-regulatory sequences because genes differ in so many other ways. To tackle this problem, King et al. created a synthetic set of *cis*-regulatory sequences based on the four core transcription factors found in mouse stem cells.

The synthetic set had every combination of two, three or four of the binding sites, with each site either facing forwards or backwards along the DNA strand. King et al. attached each of the synthetic *cis*-regulatory sequences to a reporter gene to find out how well each sequence performed. This revealed that the *cis*-regulatory sequences with the most binding sites and the tightest binding affinities work best, suggesting that transcription factors mainly work independently.

There was evidence of some interaction between some transcription factors, because, of the synthetic sequences with four binding sites, some worked better than others, and there were patterns in the most effective binding site combinations. However, these effects were small and when King et al. went on to test sequences from the real mouse genome, the most important factor by far was the number of binding sites.

Synthetic libraries of DNA sequences allow researchers to examine gene regulation more clearly than is possible in real genomes. Yet this approach does have its limitations and it is impossible to capture every type of *cis*-regulatory sequence in one library. The next step to extend this work is to combine the two approaches, taking sequences from the real genome and manipulating them one by one. This could help to unravel the rules that govern how *cis*-regulatory sequences work in real cells.
importance the precise arrangements of TFBS play in setting the activities of enhancers, and control of gene expression likely incorporates aspects of all three models. Quantifying the extent to which grammar influences activity in different contexts is an important step toward producing more predictive models of gene expression.

We and others have used mouse embryonic stem cells (mESCs) as a system for studying cis-regulatory grammar and cooperative interactions between the pluripotency factors POU5F1 (OCT4), SOX2, ESRRB, and KLF4 (Dunn et al., 2014; Fiore and Cohen, 2016; Williams et al., 2004). The pluripotency factors are a core set of TFs that maintain pluripotency in mESCs and are sufficient to induce pluripotency in terminally differentiated cells (Feng et al., 2009; Liu et al., 2008; Niwa, 2014; Takahashi and Yamanaka, 2006; Zhang et al., 2008). The pluripotency TFs activate self-renewal genes and repress genes that promote differentiation (Chambers and Tomlinson, 2009). Based on known physical and genetic interactions, as well as genome-wide binding assays, multiple interacting TFs specify target gene expression in mESCs (Huang et al., 2009; Niwa, 2014; Reményi et al., 2003; Williams et al., 2004). However, it remains unclear how pluripotency TFs collaborate to drive-specific patterns of gene expression in ESCs, and what role, if any, is played by TFBS grammar in determining specificity in the genome (Chambers and Tomlinson, 2009; Chen et al., 2008b). Understanding how these factors combine to regulate their target genes is central to understanding the establishment and maintenance of the pluripotent state.

We previously addressed these questions by assaying a set of synthetic cis-regulatory elements that represent a small fraction of the possible arrangements of pluripotency TFBS. We identified some evidence for a grammar that is constrained by TFBS arrangement, including OCT4-SOX2 interactions. However, our previous study lacked sufficient power to detect other interactions (Fiore and Cohen, 2016). Here, we explore the role of grammar for pluripotency TFBS by assaying an exhaustive set of synthetic cis-regulatory elements, composed of TFBS for SOX2, OCT4, KLF4 and ESRRB, as well as a limited set of genomic regulatory sequences with comparable configurations of binding sites. The pattern of expression of synthetic regulatory elements is well predicted by a model that incorporates binding site position. However, despite all genomic sequences overlapping ChIP-seq peaks for at least one of the four pluripotency factors, only about a third of sequences drove reporter gene activity above background levels. Additionally, the positional grammar learned from synthetic sequences performed poorly in predicting the activity of genomic sequences. Genomic sequences appear to also include sequence features that recruit additional TFs, either directly through TF-DNA interactions or possibly indirectly through TF-TF interactions. Our results suggest that in the genome the overall occupancy of TFs is the best predictor of binding site activity. Our results with synthetic elements suggest that other aspects of grammar (order, orientation) can tune the activity of sites, but these effects are difficult to observe without direct experimental manipulations. In the genome only the number and affinity of sites shows a correlation with activity.

Results

Rationale and description of enhancer libraries

We designed two reporter gene libraries to explore the role of grammar in regulatory elements controlled by the pluripotency TFs. The first library, synthetic (SYN), contains a set of synthetic combinations of consensus TFBS for OCT4 (O), SOX2 (S), KLF4 (K), and ESRRB (E). We did not include sites for NANOG in our libraries as its position weight matrix (PWM) has low information content and is not amenable to a synthetic binding site approach. Nanog also appears to be dispensable for reprogramming terminal cells to a pluripotent state (Wang et al., 2013; Wang et al., 2012; Jauch et al., 2008; Pan and Thomson, 2007; Takahashi and Yamanaka, 2006). We did not incorporate MYC-binding sites in our libraries because MYC often acts independently of the core pluripotency TFs (Chen et al., 2012; Chen et al., 2008; Liu et al., 2008).

We designed the SYN library to test how interactions between different TFs (heterotypic interactions) determine the activities of regulatory elements. If heterotypic interactions depend on the geometry of TF binding, then the order, orientation, and spacing of sites should influence activity. To test this prediction, we designed the SYN library to assay different orders and orientations of the pluripotency binding sites. The SYN library includes all possible 624 unique combinations of two, three, and four TFBS (2-mers, 3-mers, and 4-mers, respectively), with each TFBS in either the forward
or reverse direction (Supplementary file 1A). Each synthetic element in the SYN library contains no more than one copy of a given TFBS. We chose this library design to focus on heterotypic interactions and to avoid the confounding effects of homotypic interactions, which we examined in detail in a previous study (Fiore and Cohen, 2016). We embedded each TFBS in a constant 20 bp sequence with fixed spacing between sites to ensure that all the sites sit on the same side of the DNA helix. We avoided varying the length of the spacer sequence between sites because increasing the length of spacer sequences risks introducing cryptic binding sites that confound the results. For each TF, we used a consensus binding site based on its position weight matrix (PWM) in the JASPAR database (Sandelin, 2004; Fiore and Cohen, 2016). We did not vary the predicted affinity of the sites in the SYN library because we could not assay a library large enough to vary the affinity of sites while still testing all possible arrangements of sites. Our rationale was to retain the maximum power to detect the effects of the order and orientation of sites, and this required us to compromise on our ability to detect the effects of the spacing and affinity of sites. The highly controlled nature of the SYN library provides maximum power to detect interactions mediated by the order and orientation of sites.

The second library includes sequences from the mouse genome that match, as best as possible, members of the SYN library. Using the same PWMs used to design the SYN library, we scanned the mouse genome for combinations of the TFBS for O, S, K, and E within 100 bp of regions bound by any of the four pluripotency TFs in E14 mESCs as measured by ChIP-seq (Fiore and Cohen, 2016; Bailey et al., 2009; Chen et al., 2008c). We chose genomic sequences that contain one and only one binding site that scores above the PWM threshold for each factor to mimic the composition of the SYN library. We identified few clusters that included all four binding sites (<70). We therefore selected 407 genomic sequences with three pluripotency TFBS that could be compared to the exhaustive set of synthetic 3-mer elements. The resulting genomic wild-type library (gWT) is composed of 407 unique genomic sequences with combinations of any three of the four TFBS, with each site represented no more than once per sequence (Materials and methods, Supplementary file 1E-F). Although these sequences differ from SYN elements in the individual site affinities, spacings between TFBS, as well as intervening sequence composition, our expectation was that the gWT sequences would test how well interactions learned from the SYN library apply to genomic sequences. To confirm that the activity of the gWT sequences depends on the presence of pluripotency TFBS, we generated matched genomic mutant sequences (gMUT) in which all three of the identified pluripotency TFBS were mutated by changing two positions in each TFBS from the highest information content base to the lowest information base according to the PWM (Figure 1—figure supplement 1). The final gMUT sequences lack detectable TFBS for O, S, K, or E when rescanned with the threshold used to select the gWT sequences. The combined gWT/gMUT library allows us to quantify the contributions of the pluripotency sites to regulatory activity, as well as sample configurations of pluripotency TFBS from the genome that may provide insight into grammar for these sequences.

**MPRA of reporter gene libraries**

We assayed the cis-regulatory activity of the SYN and gWT/gMUT libraries in mESCs using a plasmid-based Massively Parallel Reporter Assay (MPRA) (Kwasnieski et al., 2012). Each unique library member described above is present eight times with a different unique sequence barcode (BC) in its 3’ UTR (Fiore and Cohen, 2016). The elements were placed directly upstream of a minimal promoter, mirroring classical tests of enhancer activity. The assay does not, however, test whether elements can function as long-range enhancers. To determine the relative activity of each sequence compared to the minimal promoter included in each construct, we included copies of plasmids with only the minimal promoter paired with over a hundred unique BCs in each library (Materials and methods). Our measurements were highly reproducible between biological replicates, with $R^2$ between 0.98 and 0.99 for replicates of the SYN library and 0.96–0.98 for the gWT/gMUT library, and are not driven by abundance biases in the library (Figure 1—figure supplement 2). After thresholding on DNA and RNA counts, we recovered reads for 100% (624/624) of our SYN elements and 99% (403/407) of paired gWT/gMUT sequences. The high concordance between replicates and simultaneous sequencing of the two libraries allowed us to make quantitative comparisons, both within and between libraries.
Synthetic and genomic libraries support different grammar models
TFBS in synthetic regulatory elements make strong independent contributions to expression. Most synthetic elements drive expression over basal activity regardless of the number, order, or orientation of sites within the element (Figure 1A). Of all SYN elements, 77% (6% of 2-mers, 66% of 3-mers, 92% of 4-mers) were statistically different from basal levels in all three replicates after correcting for multiple hypothesis testing (Wilcoxon rank-sum test; Bonferroni correction, n = 637; p-values reported in Supplementary file 1C). In most cases, three or four consensus binding sites are sufficient to increase expression above basal levels, which suggests strong independent contributions of TFBS to the activity of synthetic elements. Synthetic elements with more binding sites generally drive higher expression than elements with fewer binding sites, supporting the idea that TFBS can contribute to expression in an independent and additive manner. However, the wide range of expression levels observed from different 4-mer elements must be due to the arrangement of the TFBS, as site number, identity, and affinity are fixed. The strong positive effect of adding sites demonstrates an independent effect of TFBS, while the diversity of expression among elements with the same number of sites reveals that grammar can quantitatively modulate activity.

In contrast to the synthetic elements, most genomic sequences in the gWT library did not exhibit regulatory activity above basal levels. Only 28% (113/403) of wild type genomic sequences were statistically different from basal levels in all three replicates (p<0.05, Wilcoxon rank-sum test; Bonferroni correction, n = 403; p-values reported in Supplementary file 1H). This low fraction of active gWT sequences is consistent with observations from functional tests of genomic sequences bound by key

Figure 1. Activity of synthetic elements and genomic sequences. (A) The activity of synthetic elements with different numbers of binding sites. Expression is the average log of the ratio of cDNA barcode counts/DNA barcode counts for each synthetic element normalized to basal expression (dotted line). (B) The activity of genomic sequences is largely dependent on the presence of pluripotency binding sites. Normalized expression of wild type (gWT) sequences is plotted against expression of matched sequences with all three pluripotency TFBS mutated (gMUT sequences). Red indicates sequences with significantly different expression between matched gWT and gMUT sequences. The diagonal solid line is the expectation if mutation of TFBS had no impact on expression level. Expression of both gWT and gMUT sequences are normalized to basal controls, but basal expression is only plotted for gWT sequences on the y-axis (dotted line).

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Pluripotency motif substitutions for gMUT sequences.
Figure supplement 2. MPRA data quality.

King et al. eLife 2020;9:e41279. DOI: https://doi.org/10.7554/eLife.41279
TFs in other cell types (Fisher et al., 2012; Grossman et al., 2017; White et al., 2013). The difference between the SYN and gWT libraries is that the surrounding sequence context in which the pluripotency sites occur in the gWT library varies much more than in the SYN library, and these contextual differences appear to have strong effects on the pluripotency sites. In most cases, the effect of sequence context in the gWT library was strong enough to suppress the independent contributions of the binding sites to activity. For genomic sequences that were statistically different from basal, 99% (112/113) have a significant difference between matched gWT and gMUT sequences (Figure 1B; p<0.05, Wilcoxon rank-sum test; Bonferroni correction, n = 403; p-values reported in Supplementary file 1H), indicating that the activity of these genomic sequences depends on one or more of the pluripotency TFBS. Our observation that the presence of high-quality pluripotency TFBS is generally insufficient to drive expression demonstrates that binding sites must be presented in the proper surrounding sequence context in order to generate a functional regulatory element.

**Synthetic elements support a positional grammar**

While the overall pattern of expression of SYN elements supports strong independent contributions from binding sites, direct comparisons of different TFBS configurations also support a role for interactions between factors. Pairwise comparisons between 3-mers and their matched 4-mers that include one additional site at either the 5’ or 3’ end, reveal that the position of the extra site can strongly influence expression. For example, the O-K-E 3-mer and the matched O-K-E-S 4-mer drive indistinguishable expression, while the matched S-O-K-E 4-mer drives one of the highest expression levels in the SYN library (Figure 2A). Other examples are consistent with either strong position dependence or both position and orientation dependence (Figure 2—figure supplement 1A–B). Taken together, these results show that when an additional TFBS is added to an existing synthetic element, the position and orientation of the new site can have large effects on activity.

Synthetic elements appear to follow a grammar that includes some position specific interactions between TFBS. The ten highest expressing elements in the SYN library all have S and O sites next to each other and in the first two positions (Figure 2B), while the ten lowest expressing 4-mers have a strong bias for O and S in the last two positions (Figure 2C). The 10 highest expressing 4-mers all have K followed by E in the last two positions, while the lowest expressing 4-mers tend to have K and E in the first two positions. The fourth position can have an especially large effect on expression. In the highest 25% of 4-mers S is depleted (0/96) in the fourth position (Figure 2D), while in the lowest 25% E is virtually depleted (1/96) in the fourth position (Figure 2E). Conversely, in the fourth position, S is overrepresented in the top 25% (64/96) while S is overrepresented in the bottom 25% (48/96). These patterns also hold for comparisons of the strongest and weakest 3-mer and 2-mer elements (Figure 2—figure supplement 1C–F). These patterns indicate a grammar that includes a bias for S and O sites positioned upstream of K and E sites. This positioning may favor interactions between these factors and the basal transcriptional machinery or TFs recruited by the minimal promoter. As specifying a site at a given position restricts possible sites in neighboring positions, these patterns could also represent favorable interactions between factors. These data show that the precise arrangement of TFBS influences the activities of synthetic elements.

**Modeling supports a role for TFBS positions in setting expression level for synthetic elements but not for genomic sequences**

While the grammar of O, S, K, and E sites influences the relative activities of the SYN elements, their order and orientation does not appear to contribute to the activity of genomic sequences. We compared the SYN and gWT libraries for elements with configurations of OKE, OSE, OSK, and SKE TFBS. Unlike SYN 3-mer elements, all four classes of gWT sequences span the full range of expression levels observed for the entire library, with only OSK sequences having a higher average expression (Figure 3—figure supplement 1A). Thus, in genomic sequences, the same arrangement of sites embedded in different genomic contexts can either fail to drive detectable activity or drive expression higher than the highest SYN library member. To quantify the divergence in activities between genomic and synthetic elements directly, we matched gWT sequences with pluripotency TFBS-dependent activity to SYN elements with the corresponding order of TFBS. We observed no correlation in regulatory activity between matched site configurations, \( R^2 = 0.001; \) (Figure 3—figure supplement 1B). These data indicate that other variables contribute to the cis-regulatory activity of...
gWT sequences, such as the spacing and affinities of the sites, or the presence of TFBS for additional factors in flanking sequences that are held constant in the SYN library.

To identify additional sequence features that might be contributing to activity, we used a variation of the Random Forest (RF) model, an unsupervised machine learning technique. RF models can be applied for either simple classification, assigning observations to group predictions, or classifying individual observations into semi-continuous bins to make quantitative, regression-case predictions. The accuracy of predictions are assessed over a large number of decision trees trained on random subsets of the data, which allows the contribution or ‘variable importance’ of specific features to be measured. As RFs are prone to biases from early random splits in the decision trees for unbalanced data, we used iterative Random Forests (iRF) as a tool for feature selection as well as for predicting activity (Basu et al., 2018).

We first trained a regression-case iRF model on the data from the SYN library. We initialized the models with four features (Supplementary file 2A), representing only the presence or absence of each of the four pluripotency TFBS. This ‘independent’ iRF model had an $R^2$ of 0.56 between observed and predicted observations when tested on held-out data for the final iRF iteration (Figure 3—figure supplement 2). However, the independent iRF model cannot account for the differences in activities between 4-mers, because all 4-mers have identical TFBS composition (4-mers $R^2 = 0.00$). To identify features that might distinguish between the activities of 4-mers, we trained an additional regression-case iRF model, ‘independent + position’, initialized with 20 features.

Figure 2. Non-additivity in synthetic elements. (A) Comparison of synthetic 3-mer elements with matched 4-mer elements containing one additional site in the first or fourth position. Mean expression of elements across barcodes (black dot) is plotted +/- SEM (black whiskers). Green line for comparison to expression of 3-mer; Green transparency highlights SEM of 3-mer shown. Capital letter represents binding site in forward orientation and lower-case letter represents binding site in reverse orientation. Activity of the ten highest (B) and ten lowest (C) expressing 4-mers. Red line represents average expression of all synthetic 4-mer elements. Case represents binding site orientation as in (A) Mean expression of each element across barcodes (black dot) +/- SEM (black whiskers). Activity logos for the top 25% (n = 96) (D) and bottom 25% (E) of 4-mer synthetic elements. Height of letter is proportional to frequency of site in indicated position. Positions organized from 5′ end (Position 1) to 3′ end (Position 4) of elements.

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Additional examples of non-additivity in synthetic elements.

King et al. eLife 2020;9:e41279. DOI: https://doi.org/10.7554/eLife.41279

7 of 24
Research article
Computational and Systems Biology
representing both the presence and position of the four TFBS in each SYN element (Supplementary file 2A). The 20-term positional model performs well in predicting SYN expression, with an overall $R^2$ of 0.87 for the last model iteration on a held-out test set (Figure 3A). The positional iRF model highly weights the presence/absence of the sites, as expected from the performance of the independent iRF model, but also has contributions from the presence of E in the 4th position and S in the first and second positions (Figure 3B). These results reinforce the conclusion that the activity of synthetic sequences depends both on the composition and positioning of TFBS.

iRF models trained on the SYN library failed to predict or classify the expression of genomic sequences. While synthetic elements had a range of activities, elements in the gWT library are predominantly inactive, and the small number of active gWT sequences drive expression across an order of magnitude of activity levels (Figure 3—figure supplement 1A). Having such a large number of inactive sequences in the pool makes it difficult to train a model that predicts the relative activities of genomic sequences. Retraining iRF regression models to predict gWT expression fails during the training step and has no correlation with the observed expression data (independent: $R^2 = 0.03$; independent + position: $R^2 = 0.001$). In all subsequent analyses of genomic sequences, we limited

---

**Figure 3.** Positional grammar in synthetic elements. (A) Iterative random forest (iRF) regression model that includes features for presence and position of pluripotency TFBS predicts relative expression of synthetic elements. Number of binding site per element is indicated in pink (2-mers), green (3-mers), and blue (4-mers). Observed and predicted expression are both plotted in log$_2$ space. (B) Ranking of variables in synthetic iRF model. Variable importance is estimated by Increased Node Purity (IncNodePurity), the decrease in node impurities from splitting on that variable, averaged over all trees during training.

The online version of this article includes the following figure supplement(s) for figure 3:

**Figure supplement 1.** Comparison of synthetic and genomic patterns of transcription factor binding sites (TFBS).

**Figure supplement 2.** Additive effects in synthetic elements.

**Figure supplement 3.** Effect of spacer sequences between TFBS on synthetic 4-mer expression.
ourselves to models that attempt to distinguish between active and inactive genomic sequences, without predicting the relative differences in activity among active sequences. However, our first attempt to produce a classifier failed. Training a classification model to distinguish between active and inactive gWT sequences (top 25%, n = 102; bottom 75%, n = 305) using either only independent or independent + position features also fails to perform better than chance (Independent: Area Under the Receiver Operator Curve (AUROC) = 0.52, Area Under the Precision Recall Curve (AUPRC) = 0.22; Positional: AUROC = 0.47, AUPRC = 0.25; Supplementary file 2B). Genomic and synthetic elements with the same pattern of sites can drive drastically different expression levels (Figure 3—figure supplement 1B). Other sequence features present in the flanking genomic sequences and absent from the synthetic elements must therefore play a role in setting activity levels, in addition to the identity and position of the individual pluripotency TFBS.

Our results with genomic elements suggested that the sequences flanking the pluripotency TFBS play a role in determining cis-regulatory activity. We tested the effect of changing spacer sequences that flank the TFBS in six 4-mer elements from the SYN library. We tested four different spacer sequences, for a total of 30 library members, which includes the original spacer sequence. The new spacers sequences were designed to match the nucleotide content of the original spacers and minimize the creation of new TFBS (Supplementary file 1J). To ensure the dynamic range of the library, we mixed this ‘mini spacer library’ library with a small portion of the SYN library and performed an MPRA.

We found that changing the spacer sequences in the SYN library had small, but significant effects on the activities of the 4-mers. The activities of all six 4-mers in the mini spacer library tested with all four spacer sequences remained in the original range of expression for 4-mers (Figure 3—figure supplement 3A). On average, the spacer sequences modified expression by 6% (0.3–25%, Figure 3—figure supplement 3B). Although the overall effects of spacer sequences were small, the rank order of the 4-mers did change for different spacers (Figure 3—figure supplement 3C), supporting the idea that sequence features flanking the binding sites do affect gene expression. These results are consistent with the differences between the SYN and gWT libraries.

Site affinity contributes to the activity of genomic sequences

We attempted to identify other sequence features that might differentiate active and inactive gWT sequences. Sequence-based support vector machines (kmer-SVMs) are powerful tools to predict the activity of regulatory elements (Fletez-Brant et al., 2013; Chaudhari and Cohen, 2018). To identify sequence features that explain the differences between genomic elements, we trained a gapped kmer SVM (gkm-SVM) (Ghandi et al., 2016; Ghandi et al., 2014). The best performing gkm-SVM classified our positive and negative sets with AUROC of 0.75 and AUPRC of 0.77 (k = 8, gap = 2; Figure 4A). Although all sequences in the gWT library were selected to contain TFBS for the four pluripotency factors, many of the discriminative 8-mers (29/50) have motif matches that include at least one pluripotency family member (Fletez-Brant et al., 2013; Bailey et al., 2009; Supplementary file 2D). This suggests that the differences between active and inactive genomic sites could be due to the primary pluripotency sites or secondary occurrences of these sites in the intervening sequences that scored below the scanning threshold.

Sequences with higher predicted affinity pluripotency TFBS may drive higher expression. To determine if differences in the primary pluripotency sites are part of the signal identified by the SVM, we annotated gWT sequences with PWM-based scores for each TFBS present (Grant et al., 2011). For SOX2, we found no difference in scores between high and low sequences (Figure 4B; p=0.07, Welch’s t-test). For OCT4, we found a modest difference between the average scores for high and low sequences and a broader but also a significant difference for KLF4 and ESRRB PWM scores (Figure 4C–E). Summing the PWM scores for all of the TFBS further separates high and low sequences (Figure 4F–G). These patterns suggest that the quality of the primary sites contributes to the activity differences observed among gWT sequences.

We then asked if secondary sites for the pluripotency TFs might contribute to cis-regulatory activity by calculating predicted occupancy for both gWT sequences and gMUT sequences that lack the primary binding sites (Materials and methods). Predicted occupancy is a metric that includes contributions from any primary, well-scoring TFBS plus contributions from weaker sites that might be missed with traditional motif scanning (White et al., 2016; White et al., 2013; Evans et al., 2012; Segal et al., 2008; Zhao et al., 2009). We found evidence for additional low predicted affinity sites
Figure 4. Sequence features separate active and inactive genomic sequences. (A) Performance of gkm-SVM for genomic sequences supports contribution of sequence-based features to activity. Word length of 8 bp with gap size of 2 bp was used for training with threefold cross validation. ROC curve (left panel) and PR curve (right panel) is plotted for the average across threefold cross-validation sets +/- standard deviation. (B–E) Primary (O,S,K,E) site affinities across gWT sequences, as output during motif scanning plotted for high genomic sequences (top 25% as ranked by expression, Figure 4 continued on next page)
for SOX2 and OCT4 in both high and low sequences, making it unlikely that low-affinity sites strongly contribute to expression differences (Figure 4—figure supplement 1). Together, these results suggest that the affinities of the primary sites in genomic sequences, which are fixed in synthetic elements, contribute to the regulatory activity of genomic sequences more than the presence of additional sites with low predicted affinity.

We also analyzed whether the spacing between binding sites correlated with the activity of cis-regulatory elements. Using the same annotations used to determine the predicted affinities of SOX2, OCT4, ESRRB, and KLF4 binding sites, we calculated the edge-to-edge distance between every possible pair of binding sites and plotted the frequency of each spacing for high and low activity sequences (Figure 4—figure supplement 2). We observed a preference in high activity sequences for closely spaced sites for OCT4 and SOX2 reflecting a known interaction between these TFs. We also observed preferences in high activity genomic sequences for closely spaced KLF4 and OCT4 sites, and for ESRRB and OCT4 sites. Binding site spacing may therefore play a role in setting the relative activities of genomic sequences.

**Contributions from sites for other transcription factors**

A major difference between the synthetic and genomic elements is the presence of sites for TFs besides the pluripotency factors. While the synthetic elements were designed to keep the sequences between pluripotency sites constant, genomic sequences differ in both the length and composition of sequences between the pluripotency sites. The presence of binding sites for additional transcription factors may contribute to the activity of genomic sequences. To identify sites for other factors that could contribute to differences between high and low activity gWT sequences, we examined the top discriminative 8-mers from the gkm-SVM, looking at possible PWM matches for additional TFs (Supplementary file 2D). We then used PWMs for these additional TFs to identify instances of sites for other factors in the genomic sequences (see Materials and methods) (Grant et al., 2011; Sandelin, 2004). We found significant enrichment for FOXA1 sites (Figure 4H). We also found that FOXA1 and NANOG had higher total PWM scores in the high activity sequences (Figure 5—figure supplement 1A). While FOXA1 is likely not present in mESCs, other family members (FOXA2, FOXD1, FOXP1) are expressed in ESCs and have been shown to contribute to the pluripotent regulatory network, and therefore could be acting on the gWT sequences through these binding sites (Pan and Thomson, 2007; Mulas et al., 2018; Gabut et al., 2011).

Genomic sequences with higher occupancy by TFs in the genome, as measured by ChIP-seq, have higher average expression in our assay. We annotated the gWT intervals with publicly available ChIP-seq data for additional TFs and with ATAC-seq data from E14 mESCs to determine if differences in accessibility explained the difference between high and low activity sequences (Supplementary file 2B). Both high and low activity gWT sequences were accessible in the genome showing that accessibility does not necessarily correlate with high activity sequences. High activity sequences had a small but significant overlap with NANOG peaks (Figure 5—figure supplement 1B). However, for the 328 genomic sequences with a NANOG ChIP-seq signal, only 16% had an underlying TFBS as determined by motif scanning. Therefore, NANOG might be recruited by other pluripotency TFs to these sequences independent of high-quality TFBS for this factor. If we compare expression levels to the number of overlapping ChIP-seq peaks, including O,S,K,E and these additional TFs, we see that gWT sequences with higher occupancy in the genome have higher average expression in our assay (Figure 5), which has been previously observed in HepG2 cells (Ulirsch et al., 2016). This result supports a model where cumulative occupancy sets activity level.
To understand the relative contributions of the sequence features that were enriched individually, we trained iRF models with different subsets of these sequence features and compared their performance on a held-out test set (Supplementary file 2B). None of these models accurately predicted the activity of genomic sequences, likely because most genomic sequences in our collection had no activity above basal levels. Therefore, we attempted to classify active from inactive genomic sequences.

We trained an iRF model initialized with 58 features that capture differences between gWT sequences and SYN elements. These features include predicted affinity and preferred spacings between the pluripotency TFBS, the predicted occupancy for the pluripotency TFs, the presence of binding sites for additional TFs, plus chromatin accessibility (ATAC-seq) and ChIP-seq peaks for both TFs and histone marks, as well as summary features such as the total primary site affinities for each sequence (Supplementary file 2B). This gWT iRF model classified active from inactive on a held out test set with AUROC = 0.67, and AUPRC = 0.46 (Figure 6A–B, model ‘All’). Models that only included subsets of features — the spacing between elements (model ‘Spacing’), the strength of the

Figure 5. Activity of genomic sequences scales with increased occupancy in the genome. Expression of elements binned by number of intersected ChIP-seq peak signals for different factors. Number of sequences in each bin indicated in center of boxplot. All gWT sequences overlapped at least one ChIP-seq peak as per library design.

The online version of this article includes the following figure supplement(s) for figure 5:
Figure supplement 1. Genomic sequences show signatures for other factors.

To understand the relative contributions of the sequence features that were enriched individually, we trained iRF models with different subsets of these sequence features and compared their performance on a held-out test set (Supplementary file 2B). None of these models accurately predicted the activity of genomic sequences, likely because most genomic sequences in our collection had no activity above basal levels. Therefore, we attempted to classify active from inactive genomic sequences.

We trained an iRF model initialized with 58 features that capture differences between gWT sequences and SYN elements. These features include predicted affinity and preferred spacings between the pluripotency TFBS, the predicted occupancy for the pluripotency TFs, the presence of binding sites for additional TFs, plus chromatin accessibility (ATAC-seq) and ChIP-seq peaks for both TFs and histone marks, as well as summary features such as the total primary site affinities for each sequence (Supplementary file 2B). This gWT iRF model classified active from inactive on a held out test set with AUROC = 0.67, and AUPRC = 0.46 (Figure 6A–B, model ‘All’). Models that only included subsets of features — the spacing between elements (model ‘Spacing’), the strength of the

Figure 5. Activity of genomic sequences scales with increased occupancy in the genome. Expression of elements binned by number of intersected ChIP-seq peak signals for different factors. Number of sequences in each bin indicated in center of boxplot. All gWT sequences overlapped at least one ChIP-seq peak as per library design.

The online version of this article includes the following figure supplement(s) for figure 5:
Figure supplement 1. Genomic sequences show signatures for other factors.

To understand the relative contributions of the sequence features that were enriched individually, we trained iRF models with different subsets of these sequence features and compared their performance on a held-out test set (Supplementary file 2B). None of these models accurately predicted the activity of genomic sequences, likely because most genomic sequences in our collection had no activity above basal levels. Therefore, we attempted to classify active from inactive genomic sequences.

We trained an iRF model initialized with 58 features that capture differences between gWT sequences and SYN elements. These features include predicted affinity and preferred spacings between the pluripotency TFBS, the predicted occupancy for the pluripotency TFs, the presence of binding sites for additional TFs, plus chromatin accessibility (ATAC-seq) and ChIP-seq peaks for both TFs and histone marks, as well as summary features such as the total primary site affinities for each sequence (Supplementary file 2B). This gWT iRF model classified active from inactive on a held out test set with AUROC = 0.67, and AUPRC = 0.46 (Figure 6A–B, model ‘All’). Models that only included subsets of features — the spacing between elements (model ‘Spacing’), the strength of the
pluripotency sites (‘PrimarySites’), or the overlapping ChIP signal (‘ChIPSignals’) — did not perform as well (Figure 6A–B). The features that best separate active from inactive sequences were related to attributes of the pluripotency sites with the top feature being the summed pluripotency factor predicted affinity per sequence (‘OSKE_TotalAffinity’, Figure 6C). Taken together, our data suggest that genomic sequences drive higher expression when they contain strong binding sites with preferred spacing and are embedded in sequences that can mediate the recruitment of other TFs or cofactors.

Discussion
In this study, we sought to understand how pluripotency factors collaborate to drive specific levels of expression by testing both an exhaustive set of synthetic arrangements of TFBS for OCT4, SOX2, KLF4, and ESRRB and comparable genomic sequences. The experimental design allowed for direct
comparisons between the regulatory grammar of synthetic and genomic sequences. The strongest similarity between synthetic and genomic elements is that in both cases activity depends heavily on the number and affinity of binding sites. These results are most consistent with a model in which the overall occupancy of a sequence by its cognate TFs is the primary determinant of that element’s activity. Consistent with this hypothesis, the predictive power of our trained genomic model derived primarily from summing over the number and affinity of binding sites. We also observed correlation between the occupancy of sites as measured by ChIP-seq and their activity in MPRA assays. While there are many steps involved in activating gene expression, the occupancy model posits that the strength of a regulatory element is primarily controlled by its fractional occupancy by TFs.

The occupancy model might also explain the surprising result that the activity of genomic elements in our plasmid MPRA experiments do not correlate with experimental measurements of how accessible the chromatin is in their native locations. Plasmid assays might not capture regulation by chromatin, but in many cases plasmid assays do recapitulate the activity of chromosomally integrated elements (Maricque et al., 2019; Inoue et al., 2017). Alternatively, accessible regions may be bound by transcription factors but may not necessarily drive activity, such as in the case of ‘poised’ regulatory elements (Cruz-Molina et al., 2017). Nucleosome exclusion is important for regulatory activity (Khoeiery et al., 2010) and may reflect TF binding, but accessibility itself may not be sufficient for regulatory activity. Another possibility is that open chromatin may not be a direct reflection of the occupancy of an element by its cognate TFs. Other factors besides occupancy by TFs also determine the openness of chromatin, such as chromosome topology, the proximity of origins of replication, and nucleotide composition. This may explain why some genomic sequences with binding sites that reside in open chromatin do not drive high activity in MPRA assays. The prediction is that these regions are open for reasons other than occupancy by cognate TFs. That the activity of genomic elements correlates with TF occupancy as measured by ChIP-seq, but not necessarily open chromatin measurements by ATAC-seq, supports the occupancy model.

While TF occupancy was the best predictor of activity, the AUROC and AUPRC analyses show that we are still missing important features that underlie the activity of genomic sequences. Indeed, two-thirds of genomic sequences that contain consensus motifs and reside under a ChIP-seq peak for one of the pluripotency TFs had no activity in our assay. Why don’t all sequences occupied by TFs have strong regulatory activity? The sequence context in which occupied binding sites occur must contribute heavily to their activity. We attempted to address this issue by examining the regulatory grammar of synthetic elements.

Synthetic elements provide a highly controlled system for exploring whether TFBS are constrained by a regulatory grammar. With synthetic elements we found clear evidence that their activity depends on the position and orientation of pluripotency binding sites. Synthetic elements with the same number and affinity of TFBS had different levels of activity depending on the order and orientation of the sites. This result suggests that active regulatory elements in the genome are defined not only by the presence of TF occupied motifs, but also by cues in the surrounding DNA sequences. However, our models that captured the specific regulatory grammar of synthetic elements failed to predict the activity of genomic sequences.

Why don’t models that robustly predict the activity of synthetic elements also predict the activity of genomic sequences? With synthetic elements, each sequence differs from others in the library by only a small number of sequence features. In synthetic libraries, there are many pairs of elements that differ by only a single sequence feature, which provides power to observe experimentally the effect of a single variable. In contrast, libraries of genomic elements are much more diverse, and the analysis of genomic sequences relies on detecting correlations between elements that share sequence features. However, it is difficult to isolate the effect of a single sequence feature because genomic elements that share a certain sequence feature will always be very different in terms of other features. The strength of the synthetic approach is the power it provides to isolate the effects of specific sequence features or pairs of sequence features. The weakness of the synthetic approach is that genomic elements are subject to many context specific constraints, all of which cannot be captured in a single synthetic library. When we changed the spacer sequences in our synthetic library, we found small but reproducible effects on expression. Our interpretation of this result is that changing the spacer sequences did not have large effects on the independent contribution of each TFBS, but did have effects on the interactions between sites (i.e. the regulatory grammar). In the future, we plan to use the regulatory grammar derived from synthetic elements to design
experiments that manipulate single features of genomic elements. If the grammar that is learned from synthetic elements reflects real constraints in the cell, then models of synthetic elements should predict the relative effects of single perturbations of genomic elements even if they cannot predict the absolute expression of genomic sequences. A combined approach that leverages both synthetic and genomic sequences should continue to help unravel the rules that govern cis-regulation of expression in cells.

**Materials and methods**

### Key resources table

<table>
<thead>
<tr>
<th>Reagent type (species) or resource</th>
<th>Designation</th>
<th>Source or reference</th>
<th>Identifiers</th>
<th>Additional information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell line (Mus musculus mouse)</td>
<td>RW4</td>
<td>other</td>
<td>RRID: CVCL_6442</td>
<td>Gift from Mitra Lab, CGS, Department of Genetics, Washington University School of Medicine in St. Louis. The cell line tested negative for mycoplasma contamination by the Genome Editing and iPSC core at Washington University in St. Louis.</td>
</tr>
<tr>
<td>Commercial assay or kit</td>
<td>PureLink RNA Mini Kit</td>
<td>ThermoFisher Scientific/Invitrogen</td>
<td>Cat#: 12183018A</td>
<td>Followed manufacturer's protocol</td>
</tr>
<tr>
<td>Commercial assay or kit</td>
<td>PureLink DNase Set</td>
<td>ThermoFisher Scientific/Invitrogen</td>
<td>Cat#: 12185010</td>
<td>Followed manufacturer's protocol</td>
</tr>
<tr>
<td>Commercial assay or kit</td>
<td>TURBO DNA-free</td>
<td>ThermoFisher Scientific/Invitrogen</td>
<td>Cat#: AM1907</td>
<td>Followed manufacturer's protocol</td>
</tr>
<tr>
<td>Commercial assay or kit</td>
<td>SuperScript III Reverse Transcriptase</td>
<td>ThermoFisher Scientific/Invitrogen</td>
<td>Cat#: 18080044</td>
<td>Followed manufacturer's protocol</td>
</tr>
<tr>
<td>Commercial assay or kit</td>
<td>anti-Alkaline Phosphatase (AP) staining</td>
<td>System Biosciences</td>
<td>Cat.# AP100R-1</td>
<td>Followed manufacturer's protocol</td>
</tr>
<tr>
<td>Recombinant DNA reagent</td>
<td>SYN</td>
<td>this paper</td>
<td></td>
<td>Recombinant plasmid library of synthetic (SYN) elements upstream of a minimal Pou5f1 promoter and dsRed/SV40 UTR reporter element</td>
</tr>
<tr>
<td>Recombinant DNA reagent</td>
<td>GEN</td>
<td>this paper</td>
<td></td>
<td>Recombinant plasmid library of sequences identified in the mouse genome (GEN) upstream of a minimal Pou5f1 promoter and dsRed/SV40 UTR reporter element</td>
</tr>
<tr>
<td>Recombinant DNA reagent</td>
<td>miniSpacer</td>
<td>this paper</td>
<td></td>
<td>Recombinant plasmid library of synthetic elements with swapped spacer (miniSpacer) sequences upstream of a minimal Pou5f1 promoter and dsRed/SV40 UTR reporter element</td>
</tr>
<tr>
<td>Software, algorithm</td>
<td>iRF v2.0.0</td>
<td><a href="https://cran.r-project.org/web/packages/iRF/index.html">https://cran.r-project.org/web/packages/iRF/index.html</a></td>
<td>DOI: 10.1073/pnas.1711236115</td>
<td></td>
</tr>
</tbody>
</table>

*Continued on next page*
Library design

To generate a library that contained both synthetic and genomic elements, we ordered a custom pool of 13,000 unique 150 bp oligonucleotides (oligos) from Agilent Technologies (Santa Clara, CA) through a limited licensing agreement. Each oligo in the SYN pool was 150 bp in length with the following sequence:

\[
\text{CTTCTACTACTAGGGCCCA[SEQ]AAGCTT[FILL]GAATTCTCTAGAC[BC]TGAGCTCTACATGC}
\]

where [SEQ] is a 40–80 bp synthetic element comprised of concatenated 20 bp building blocks of pluripotency sites, as described previously, with the fifth position of the KLF4 site changed to ‘T’ to facilitate cloning (Fiore and Cohen, 2016). [FILL] is a random filler sequence of variable length to bring the total length of each sequence to 150 bp, and [BC] is a random 9 bp barcode. The oligonucleotide pool contained all possible combinations of the pluripotency binding sites in both orientations, with no more than one of each site per sequence in lengths of two, three, and four building blocks. The sequence of each of the element is listed in Supplementary file 1B. In total, the SYN library has 624 unique synthetic elements. Each synthetic element is present in the pool eight times, each time with a different unique BC. There are also 112 oligos in the pool for cloning the basal promoter without any upstream element, each with a unique BC.

Genomic sequences were represented in the pool by 150 bp oligos with the following sequences:

\[
\text{GACTTACATTAGGGCCGT[SEQ]AAGCTT[FILL]GAATTCTCTAGAC[BC]TGAGCTCTACATGC}
\]

where [SEQ] is either a reference (gWT) or mutated (gMUT) genomic sequence of 81–82 bps. Reference gWT sequences were selected by choosing regions of the genome within 100 bps of previously identified ChIP-seq peaks for these four pluripotency factors (Chen et al., 2008b). After excluding poorly sequenced and repetitive regions (ENCODE Project Consortium, 2012; Waterston et al., 2002), we scanned the remaining regions using FIMO with the four PWMs used previously to design the synthetic building blocks, with a p-value threshold of $1 \times 10^{-3}$ (Grant et al., 2011; Bailey et al., 2009; Fiore and Cohen, 2016). Regions that contained more than one overlapping site identified by FIMO were excluded. Binding sites that were located less than 20 bp from each other were then merged into a single genomic element using Bedtools (Quinlan and Hall, 2010). Elements with no more than one of each site per element were then selected and expanded to 81–82 bp centered on the motifs. Expanded sequences were rescanned to confirm the presence of only three binding sites with the same threshold as used to originally scan the sequences. Sequences that contained restriction sites for were then removed from the library, leaving 407 genomic sequences with combinations of the OCT4, SOX2, KLF4, and/or ESSRB TFBS.

We generated matched mutated sequences (gMUT) for each of the 407 gWT sequences by changing two positions in each motif from the highest information content base to the lowest information base for that position (Figure 1—figure supplement 1). The reverse complement position and substitution was made for the reverse orientation of each motif. The mutated sequences were rescanned with all four original PWMs to confirm that no detectable pluripotency TFBS remained, using FIMO with the same p-value threshold ($1 \times 10^{-3}$) as above.

In total, the pool of oligos representing genomic sequences contained 407 wild-type sequences (gWT) and the corresponding 407 gMUT sequences. The sequence of each element is listed in Supplementary file 1G. Each of these 814 sequences were associated with eight unique BCs. The primers for gWT and gMUT sequences were identical so all subsequent steps for this library was performed in a single pool. There are also 112 oligos in the pool for cloning the basal promoter without

<table>
<thead>
<tr>
<th>Reagent type (species) or resource</th>
<th>Designation</th>
<th>Source or reference</th>
<th>Identifiers</th>
<th>Additional information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software, algorithm</td>
<td>BEEML</td>
<td><a href="http://stormo.wustl.edu/beeml/">http://stormo.wustl.edu/beeml/</a></td>
<td>DOI: 10.1371/journal.pcbi.1000590</td>
<td></td>
</tr>
</tbody>
</table>
any upstream element, each with a unique BC (Supplementary file 1F). The rest of the array contained sequences not used in this study.

**Cloning of plasmid libraries**

For a full list of primers, see Supplementary file 3. The synthesized oligos were prepared as previously described (Kwasnietski et al., 2012; Fiore and Cohen, 2016), except using primers Synthetic_FW-1 and Synthetic_Rev-2 with an annealing temperature of 55°C for the SYN library and primers Genomic_FW-1 and Genomic_Rev-1 with an annealing temperature of 53°C for the gWT/gMUT libraries. PCR products were purified from a polyacrylamide gel as described previously (White et al., 2013). Each library was cloned as described previously (Fiore and Cohen, 2016), with an SYN element (SYN library) or either a gWT or gMUT sequence (gWT/gMUT library) cloned into the ApaI and SacI sites of plasmid pCF10.

The *pou5f1* basal promoter and dsRed reporter gene were amplified from pCF10 using primers CF121 and CF122, and inserted into the plasmid library pools from the previous step at the XbaI and HindIII sites. Digestion of the libraries with Spel and subsequent size selection was omitted as the SYN library had less than 2% background and the combined gWT/gMUT library had less than 1% background in the final cloning step.

**Spacer library**

For the mini spacer library, we ordered an oligo pool containing 4-mer elements with different spacer sequences from Integrated DNA Technologies (Coralville, IA). Each oligo in the mini library was 161 bp in length with the following sequence:

```
GACATCAAGATCTGCGCTCGGGC[SEQ]AAGCTTGAATTCTCTAGAC[BC]TGAGCTCTCGCTTCGAGCAGACATGAT
```

where [SEQ] represents an oligo sequence described below and [BC] is a random 9 bp barcode. We picked six 4-mer oligos from the original synthetic library to span the 4-mer expression range and swapped out the spacer sequences in the oligos for four other sequences, generating a total of 30 constructs, including the original spacers. Each construct was represented in the pool with five unique barcodes. The sequence of each element is in Supplementary file 1K.

The mini spacer library was cloned into the same backbone as the previous libraries. Briefly, pCF10 was digested with ApaI and SacI, and the single-stranded oligo pool was directly assembled into the backbone using HiFi DNA assembly. The *pou5f1* basal promoter and dsRed reporter gene were amplified from pCF10 using CF121 and CF122, then ligated into the mini spacer library following the same approach as the SYN, gWT, and gMUT libraries.

**Cell culture and transfection**

RW4 mESCs were cultured as described previously (Xian et al., 2005; Chen et al., 2008a) on 2% gelatin coated plates in standard media (DMEM, 10% fetal bovine serum, 10% newborn calf serum, nucleoside supplement, 1000 U/ml leukemia inhibitory factor (LIF), and 0.1 μM B-mercaptoethanol). Approximately 1 million cells at 100% estimated viability were seeded into six-well plates 24 hr prior to transfection. The SYN library and combined gWT/gMUT were transfected in parallel using 10 μL Lipofectamine 2000 (Life Technologies, Carlsbad, CA), 3 μg of plasmid library, and 0.3 μg CF128 (a GFP control plasmid) per well, as described previously (Fiore and Cohen, 2016). Four biological replicates of each library pool, the SYN plasmid pool or combined gWT/gMUT plasmid pool, were transfected and the plates were passaged 6 hr post-transfection. For three replicates of each library pool, RNA was extracted 24 hr post-transfection from approximately 9 million cells per replicate, using the PureLink RNA mini kit (Life Technologies, Carlsbad, CA) with the fourth transfection replicate reserved for estimating transfection efficiency via fluorescent microscopy and staining for alkaline phosphatase (AP) activity, a universal pluripotency marker (Singh et al., 2012).

**Massively parallel reporter assay**

Massively parallel reporter gene assays were used to measure the activity of each element as described previously (Fiore and Cohen, 2016; Mogno et al., 2013). Briefly, we used Illumina Next-Seq (San Deigo, CA) sequencing of both the RNA and original plasmid DNA pool, removing excess
DNA from the RNA pool using TURBO DNA-free kit (Life Technologies, Carlsbad, CA). cDNA was then prepared using SuperScript RT III (Life Technologies, Carlsbad, CA) with oligo dT primers. Both the cDNA and the plasmid DNA pool were amplified using primers CF150 and CF151b, for 13 cycles. The PCR amplification products were digested using XbaI and Xhol (New England Biolabs, Ipswich, MA), ligating the resulting digestion products to custom Illumina adapter sequences, P1_XbaI_X (where X is 1 through 8, with in-line multiplexing BC sequences) to the 5’ overhang and PE2_SIC69_SalI on the 3’ Xhol overhang, each of which is comprised of annealed forward (F) and reverse (R) strands. An enrichment PCR with primers CF52 and CF53 was then used, and the resulting products were mixed at equal concentration and sequenced on one NextSeq lane.

Sequencing reads were filtered to ensure that the BC sequence perfectly matched the expected sequence. For the SYN library, this resulted in 40 million reads combined for the three demultiplexed RNA samples (P1_XbaI_1, P1_XbaI_2, P1_XbaI_3; 12.7–13.5 million each), and 19.7 million reads for the DNA library sample (P1_XbaI_7). For the combined gWT/gMUT libraries, this resulted in approximately 37 million reads combined for the three demultiplexed RNA samples (P1_XbaI_4, P1_XbaI_5, P1_XbaI_6; 9.4–16 million each), and 19.6 million reads for the DNA library sample (P1_XbaI_8). For each library, BCs that had less than three raw counts in any RNA replicate or less than 10 raw counts in the DNA sample were removed before proceeding with downstream analyses.

Expression normalization was performed by first calculating reads per million (RPM) per BC for each replicate for both the SYN library and the combined gWT/gMUT library. For each BC, expression was calculated by dividing the RPMs in each RNA replicate by the DNA pool RPMs for that BC. Normalizing by DNA RPMs successfully removed the impact of the representation of the construct in the original pool as the calculated expression has no correlation with the DNA counts for both the SYN library and the combined gWT/gMUT. Within each biological replicate, the BCs corresponding to each synthetic element (SYN) or genomic sequence (gWT/gMUT) were averaged and then normalized by basal mean expression in that replicate. These normalized expression values were then averaged across biological replicates. All downstream analyses were performed in R version 3.3.3 and plotted with ggplot2 version 2.2.1. Expression summaries per replicate are reported in Supplementary file 1C for the SYN library, Supplementary file 1H for the gWT/gMUT library and Supplementary file 1L for the ‘mini spacer’ library.

**Predicted occupancy**

Custom code, based on Zhao and Stormo’s BEEML algorithm (Zhao et al., 2009), was used to compare sequences of interest to a provided Energy Weight Matrix (EWM) at a set protein concentration (mu) and output a predicted occupancy for that TF as in White et al. (2013). Briefly, an energy landscape (EWM score) is calculated by comparing all n-mers of each sequence, where n = length of provided motif, to the matrix to generate an array of individual base scores for the forward and reverse orientation of the sequence. Occupancy is then predicted using equation 3 for binding probability at equilibrium, \( P(OCC) = \frac{1}{1 + e^{\Delta G - \mu}} \). Position Frequency Matrices equivalent to the PWMs used for both SYN building block design and for scanning the mouse genome were used to generate EWMs, using the formula \( RT = \ln \left( \frac{Freq(Base consensus)}{Freq(Base i)} \right) \) to convert the frequency of each base at each position i to a pseudo \( \Delta G \) values for each factor (White et al., 2013). Predicted occupancy (P(Occ)) for the 3-mer SYN elements was calculated for different assumed protein concentrations (\( mu = 0.5, 1, 2, 4, 5, 8, 10, 12 \)) to determine at what point the SYN elements are predicted to be saturated, where P(Occ) \( \geq 3 \) for each SYN element, that is: approaching one for each TFBS in the sequence. SYN elements were saturated by each of the four pluripotency factors at \( mu = 8 \) with the exception of the shorter Oct4 motif, which reached saturation at \( mu = 10 \). Occupancy of gWT and gMUT sequences was predicted for gWT and gMUT at an assumed high protein concentration of \( mu = 8 \) for Sox2, Klf4, Esrb, and \( mu = 10 \) for Oct4, consistent with the role of these factors in mESCs. The predicted occupancy of each factor for matched gMUT sequences are reported in Supplementary file 2F as a feature of gWT sequences. iRF models:

We built iterative Random Forest (iRF) models to classify our data using the R package iRF (version 2.0.0) (Basu et al., 2018). To run the software a model is initialized with 1/p weights for each of p features to be included in fitting the model. In each iteration, p features are reweighted by their Gini Importance (\( w^i \)), a measure that is calculated by how purely a node, split by feature, separates the classes (Menze et al., 2009; Louppe et al., 2013). Default settings were used for model training,
with four iterations of reweighting $p$ features specified for each model as indicated in Supplementary files 2A and 2B.

Synthetic data was split into training and test sets by randomly subsetting 50% of the total SYN elements (total $n = 407$). Mean normalized expression was the response variable for model fitting for the synthetic models (see Supplementary file 2E for feature annotations for SYN elements). Four iterations of model fitting on training data was used.

Genomic data was split into training and test sets by randomly subsetting 50% of the total gWT/gMUT intervals (total $n = 624$). Classification as ‘active’, 1, if mean normalized gWT expression was greater than or equal to the 3rd quartile and ‘inactive’, 0, if mean normalized gWT expression was less than the 3rd quartile (cutoff value = 1.983), was the response variable for model fitting (see Supplementary file 2F for feature annotations and response values for gWT sequences). Four iterations of model fitting on training data was used. gkm-SVM:

We used a gapped $k$-mer Support Vector Machine (gkm-SVM) to search for gapped $k$-mers that distinguish between highly active and inactive genomic sequences (Ghandi et al., 2016). We subset sequences from the gWT library into top 25% (high) and bottom 25% (low) based on expression data for a total of 101 positive and 101 negative intervals for the training set. FASTA sequences were then generated from the mm10 reference genome (Bioconductor, BioMart) for each region (Supplementary file 4). We then used the gkm-SVM R package to classify high vs. low sequences (Ghandi et al., 2016). Word length ($L$) values of 6 ($gap = 2$), 8 ($gap = 2$), and 12 ($gap = 6$), were tested with cross validation. Default settings were used for other function options. Three-fold cross validation was chosen due to the the amount of structure in the data, with combinations of OSK binding sites overrepresented in positive training sequences (Figure 3—figure supplement 1). The best average performance on training data as evaluated by AUCs was the model trained with parameters of $L = 8$ and $gap = 2$ (See Supplementary file 2G for output scores). The final gkm-SVM model includes approximately 1 million unique $k$-mers (See Supplementary file 2C for full kmer list and weights).

Other analysis and data sources

All genome coordinates from previous mouse genome builds were converted to mm10 using the UCSC liftover tool (Kuhn et al., 2013). Binding matrices for SOX2, OCT4, KLF4, ESRRB were as previously reported (Fiore and Cohen, 2016). The Bedtools suite (version 2.20) was used for manipulations and analysis of bed files (Quinlan and Hall, 2010). Statistical tests were chosen based on expectations of normalcy, with Wilcoxon rank-sum test used for comparisons of BC expression as these distributions were observed to be skewed for some library members, Welch’s t-test used where sample sizes were equal and roughly normal, and Fisher’s 1-sided tests used for testing for enrichment in small sample sizes.

Data access

Raw sequencing data for SYN library and gWT/gMUT library can be found under SRA accession number SRR7515851. Processed sequencing data, specifically demultiplexed barcode counts per replicate, can be found under GEO accession number GSE120240. Additionally, a table of normalized reads per million (RPMs) across replicates for all barcodes are included as Supplementary file 1D for the SYN library, Supplementary file 1I for the gWT/gMUT library, and Supplementary file 1M for the MiniSpacer library.

Acknowledgements

We thank members of the Cohen Lab for critical reading and feedback, particularly Michael White, Max Staller and Hemangi Chaudhari for helpful discussion over the course of the project, Jessica Hoisington-Lopez from the DNA Sequencing Innovation Lab for assistance with high-throughput sequencing, and Karl Kumbier for modeling discussions. This work is supported by a grant from the National Institutes of Health, R01 GM092910 to BAC.
Additional information

**Funding**

<table>
<thead>
<tr>
<th>Funder</th>
<th>Grant reference number</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Institutes of Health</td>
<td>R01 GM092910</td>
<td>Barak Cohen</td>
</tr>
</tbody>
</table>

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

**Author contributions**

Dana M King, Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Methodology; Clarice Kit Yee Hong, Data curation, Formal analysis, Methodology, Writing - review and editing; James L Shepherdson, Investigation, Writing - review and editing; David M Granas, Investigation; Brett B Maricque, Data curation, Formal analysis, Methodology; Barak A Cohen, Conceptualization, Supervision, Funding acquisition

**Author ORCIDs**

Dana M King [https://orcid.org/0000-0003-4635-5272](https://orcid.org/0000-0003-4635-5272)
Clarice Kit Yee Hong [https://orcid.org/0000-0002-9485-1425](https://orcid.org/0000-0002-9485-1425)
James L Shepherdson [https://orcid.org/0000-0002-3288-7000](https://orcid.org/0000-0002-3288-7000)
Barak A Cohen [https://orcid.org/0000-0002-3350-2715](https://orcid.org/0000-0002-3350-2715)

**Decision letter and Author response**

Decision letter [https://doi.org/10.7554/eLife.41279.sa1](https://doi.org/10.7554/eLife.41279.sa1)
Author response [https://doi.org/10.7554/eLife.41279.sa2](https://doi.org/10.7554/eLife.41279.sa2)

**Additional files**

**Supplementary files**

- Supplementary file 1. Composition of libraries and expression measurements.
- Supplementary file 2. Features, weights and motifs of iRF and gkmSVM models.
- Supplementary file 3. Sequences of primers used in this study.
- Supplementary file 4. FASTA-format input file of GEN library sequences for gkmSVM.
- Transparent reporting form

**Data availability**

Sequencing data has been deposited in GEO under accession code GSE120240. Any additional data generated during this study are included in the manuscript and supporting files.

The following dataset was generated:

| Author(s)                          | Year | Dataset title                                                                 | Dataset URL                                                                 | Database and Identifier                           |
|------------------------------------|------|-------------------------------------------------------------------------------|----------------------------------------------------------------------------|

The following previously published datasets were used:

| Author(s)                        | Year | Dataset title                                                                 | Dataset URL                                                                 | Database and Identifier                           |
|----------------------------------|------|-------------------------------------------------------------------------------|----------------------------------------------------------------------------|


References


Chen CT, Gottlieb DI, Cohen BA. 2008a. Ultraconserved elements in the Olig2 promoter. PLOS ONE 3:e3946. DOI: https://doi.org/10.1371/journal.pone.0003946, PMID: 19079603


Evans NC, Swanson CI, Barolo S. 2012. Sparkling insights into enhancer structure, function, and evolution. Current Topics in Developmental Biology 98:97–120. DOI: https://doi.org/10.1016/B978-0-12-386499-4.00004-S, PMID: 22305160


Hare EE, Peterson BK, Eisen MB. 2008a. A careful look at binding site reorganization in the even-skipped enhancers of Drosophila and sepsids. PLOS Genetics 4:e1000268. DOI: https://doi.org/10.1371/journal.pgen.1000268

Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. 2008b. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLOS Genetics 4:e1000106. DOI: https://doi.org/10.1371/journal.pgen.1000106, PMID: 18584029


Zhao Y, Granas D, Stormo GD. 2009. Inferring binding energies from selected binding sites. PLOS Computational Biology 5:e1000590. DOI: https://doi.org/10.1371/journal.pcbi.1000590, PMID: 19997485
Appendix B: A single-cell massively parallel reporter assay detects cell type-specific gene regulation

This work was performed in collaboration with Siqi Zhao, David Granas, Connie Myers, Mike White, Joe Corbo and Barak Cohen. I helped provide the initial library and some core promoter analyses, as well as manuscript editing and figure preparation. This work was accepted for publication in 2022 in *Nature Genetics* and is currently on *Biorxiv* (doi: 10.1101/2021.11.11.468308) [331].
A single-cell massively parallel reporter assay detects cell type specific cis-regulatory activity

Siqi Zhao¹,², Clarice KY Hong¹,², Connie A Myers³, David M Granas¹,², Michael A White¹,², Joseph C Corbo³, Barak A Cohen¹,²

¹Edison Family Center for Systems Biology and Genome Sciences, ²Department of Genetics, Washington University School of Medicine, USA. ³Department of Pathology, Washington University School of Medicine, USA.

Massively parallel reporter gene assays are key tools in regulatory genomics, but cannot be used to identify cell-type specific regulatory elements without performing assays serially across different cell types. To address this problem, we developed a single-cell massively parallel reporter assay (scMPRA) to measure the activity of libraries of cis-regulatory sequences (CRSs) across multiple cell-types simultaneously. We assayed a library of core promoters in a mixture of HEK293 and K562 cells and showed that scMPRA is a reproducible, highly parallel, single-cell reporter gene assay that detects cell-type specific cis-regulatory activity. We then measured a library of promoter variants across multiple cell types in ex vivo mouse retinas and showed that subtle genetic variants can produce cell-type specific effects on cis-regulatory activity. We anticipate that scMPRA will be widely applicable for studying the role of CRSs across diverse cell types.

Introduction

The majority of heritable variation for human diseases maps to the non-coding portions of the genome¹–⁶. This observation has led to the hypothesis that genetic variation in the cis-regulatory sequences (CRSs) that control gene expression underlies a large fraction of disease burden⁷–¹⁰. Because many CRSs function only in specific cell types¹¹, there is intense interest in high-throughput assays that can measure the effects of cell-type-specific CRSs and their genetic variants.

Massively Parallel Reporter Assays (MPRAs) are one family of techniques that allow investigators to assay libraries of CRSs and their non-coding variants en masse¹²–¹⁸. In an MPRA experiment, every CRS drives a reporter gene carrying a unique DNA barcode in its 3’ UTR, which allows investigators to quantify the activity of each CRS by the ratio of its barcode abundances in the output RNA and input DNA. This approach allows investigators to identify new CRSs, assay the effects of non-coding variants, and discover general rules governing the functions of CRSs¹²,¹⁹–²³. One limitation of MPRAs is that they are generally performed in monocultures, or as bulk assays across the cell types of a tissue. Performing cell-type specific MPRAs in tissues will require methods to simultaneously readout reporter gene activities and cell type information in heterogeneous pools of cells.

To address this problem, we developed scMPRA, a procedure that combines single-cell RNA sequencing with MPRA. scMPRA simultaneously measures the activities of reporter genes in single cells and the identities of those cells using their single-cell transcriptomes. The key
component of scMPRA is a two-level barcoding scheme that allows us to measure the copy number of all reporter genes present in a single cell from mRNA alone. A specific barcode marks each CRS of interest (CRS barcode, “cBC”) and a second random barcode (rBC) acts as a proxy for DNA copy number of reporter genes in single cells (Fig. 1a). The critical aspect of the rBC is that it is complex enough to ensure that the probability of the same cBC-rBC appearing in the same cell more than once is vanishingly small. In this regime, the number of different cBC-rBC pairs in a single cell becomes an effective proxy for the copy number of a CRS in that cell. Even if a cell carries reporter genes for multiple different CRSs, and each of those reporter genes is at a different copy number, we can still normalize each reporter gene in each individual cell to its plasmid copy number. With this barcoding scheme, we can measure the activity of many CRSs with different input abundances at single-cell resolution, which allows us to measure the activity of CRSs simultaneously across different populations of cells.

Results

scMPRA enables single-cell measurement of CRS activity

As a proof of principle, we first used scMPRA to test whether different classes of core promoters show different activities in different cell types. Core promoters are the non-coding sequences that surround transcription start sites, where general cofactors interact with RNA polymerase II\(^{24,25}\). Core promoters are divided into different classes by the functions of their host genes (housekeeping vs developmental), as well as by the sequence motifs they contain (TATA-box, downstream promoter element (DPE), and CpG islands). We selected 676 core promoters that we previously tested\(^{24,25}\) and cloned them into a double-barcoded MPRA library. Given the complexity of the library (>1x10\(^7\) unique cBC-rBC pairs), we calculated that the probability of plasmids with the same cBC-rBC pair occurring in the same cell is less than 0.01 with our transfection protocols. Given this low likelihood, the number of rBC per cBC in a cell represents the copy number of a CRS in that cell. Knowing the copy number of CRSs in single cells allows us to normalize reporter gene expression from each CRS to its copy number in individual cells.

We performed a cell mixing experiment to test whether scMPRA could measure cell type specific expression of reporter genes. We transfected K562 (chronic myelogenous leukemia) and HEK293 (human embryonic kidney) cells, and performed scMPRA on a 1:1 mixture of those cell lines (Fig. 1b). The mRNA from single cells was captured, converted to cDNA, and sequenced. The resulting cBC-rBC abundances and transcriptome of each single cell are linked by their shared 10x cell barcode.

We recovered a total of 3112 cells (97%) that could be unambiguously assigned to one of the two cell types (Fig. 2a, Supplementary Fig. 1a,b) and computed the mean expression of each core promoter in the library in each cell type (Methods). The measurements were reproducible in both cell types (K562: Pearson’s R = 0.89, Spearman’s \(\rho = 0.57\), HEK293: Pearson’ R = 0.96, Spearman’s \(\rho = 0.92\) (Figs. 2b,c, Supplementary Table 1), and we obtained measurements for 99.5% of core promoters in K562 cells and 100% in HEK293 cells, highlighting the efficiency of scMPRA. The median number of cells in which each core promoter was measured was 76 for K562 cells and 287 for HEK293 cells (Figs. 2d,e). We also tabulated the
Fig. 1: scMPRA measures CRS at single-cell resolution. (a) Each CRS reporter construct is barcoded with a cBC that specifies the identity of the CRS, and a highly complex rBC. The complexity of the cBC-rBC pair ensures that the probability of identical plasmids being introduced into the same cell is extremely low. (b) Experimental overview for scMPRA using the mixed-cell experiment as an example. K562 cells and HEK293 cells are transfected with the double-barcoded core promoter library. After 24 hours, cells were harvested and mixed for 10x scRNA-seq. Cell identities were obtained by sequencing the transcriptome, and single-cell expression of CRSs were obtained by quantifying the barcodes. The cell identity and CRSs expression (as measured by the cBC-rBC abundances) were linked by the shared 10x cell barcodes.

We asked whether the data allowed us to detect core promoters with differential activity between K562 and HEK293 cells. While different classes of core promoters generally had similar activities in both cell lines (Fig. 2h), our differential analysis using DEseq2 identified a small number of promoters (11 out of 669) that are upregulated in K562 cells, and 59 promoters that are downregulated in K562 cells (adjusted p-value < 0.01, log2 fold change > 0.3, Fig. 2i, Supplementary Table 2). Among the down-regulated promoters, 48 out of 59 core promoters belong to housekeeping genes (p=1.08x10^-11, Fig. 2j), and 46 out of 59 core promoters are CpG-island-containing core promoters (p=2.18x10^-6, Fig. 2k). This result is not due to differences in the quality of measurements between housekeeping and developmental promoters.
Fig. 2: scMPRA detects cell type specific CRS activity. (a) UMAP of the transcriptome from the mixed-cell scMPRA experiment. 3312 out of 3417 cells are assigned to either K562 or HEK293 cells. (b,c) Reproducibility of replicate measurements of the mean expression from each core promoter in both K562 and HEK293 cells. (d,e) Histogram of the number of cells in which each core promoter was measured for HEK293 and K562 cells. (f,g) Correlations between scMPRA and bulk MPRA using mRNA abundances (cBC counts per cell) to make the two methods comparable. (h) Boxplot of the activities of core promoters from different categories in K562 (orange) and HEK293 (blue) cells. The promoter categories are taken from Haberle et al. Because the average expression of all promoters was different between K562 and HEK293, we plotted each category according to its deviation from the average expression (z-score) of all promoters in each cell type. (i) Volcano plot for differential expression (DE) of core promoters in K562 and HEK293 cells. Significant DE reporters (red dots) have adjusted p-value <0.01 and log2 fold change greater than 0.3. (j) Venn diagram of the functional characterization (housekeeping vs developmental) of down-regulated core promoters in K562 cells. Housekeeping promoters are enriched (p-value = 1.08x10^{-11} from hypergeometric test). (k) Pie chart of the sequence features (CpG, DPE, TATA) of down-regulated core promoters. CpG promoters are enriched (p=2.18x10^{-6}, from hypergeometric test).

These results demonstrate the ability of scMPRA to detect CRSs with cell-type specific activities.
scMPRA detects cell sub-state specific CRS activity

Single-cell studies have revealed heterogeneity in cell states even within isogenic cell types\textsuperscript{27–30}. Therefore, we asked if scMPRA can identify CRSs with cell-state specific activity. We repeated scMPRA on K562 cells alone and obtained a total of 5141 cells from two biological replicates. Measurements of each library member were again highly correlated between replicates and agree well with independent bulk measurement (Supplementary Figs. 2a,b).

Because the phases of the cell cycle represent distinct cell-states, we asked whether scMPRA could identify reporter genes with differential activity through the cell cycle. We assigned cell cycle phases to each cell using their single cell transcriptome data (Fig. 3a) and calculated the mean expression of each reporter gene in different cell cycle phases. We found that most core promoters in our library are upregulated in the G1 phase of the cell cycle, and that some housekeeping promoters are highly expressed through all cell cycle phases (Fig. 3b, Supplementary Table 3). We also identified core promoters with different expression dynamics through the cell cycle. For example, we found that the core promoter of \textit{UBA52} remains highly expressed in the S phase, whereas the core promoter of \textit{CXCL10} is lowly expressed throughout (Supplementary Fig. 2c). This analysis illustrates the ability of scMPRA to identify CRSs whose expression naturally fluctuates with cellular dynamics.

We then asked whether scMPRA could detect reporter genes with activities that were specific to other cell-states in K562 cells, after normalizing for cell cycle effects. We focused on two specific sub-states that have been reported and experimentally validated for high proliferation rates in K562 cells\textsuperscript{31,32}. The first is the CD34\textsuperscript{+}/CD38\textsuperscript{−} sub-state that has been identified as a leukemia stem cell subpopulation, and the second is the CD24\textsuperscript{+} sub-state that is linked to selective activation of proliferation genes by bromodomain transcription factors\textsuperscript{28,29}. To identify these sub-states in our single-cell transcriptome data, we first regressed out the cell cycle effects and confirmed that the single cell transcriptome data no longer clustered by cell cycle phase (Supplementary Fig. 2d). We then identified clusters within K562 cells that have the CD34\textsuperscript{+}/CD38\textsuperscript{−} expression signature, or the CD24\textsuperscript{+} signature (Fig. 3c). Although the CD34\textsuperscript{+}/CD38\textsuperscript{−} cells represent only 9.3\% of the cells, scMPRA revealed two distinct classes of core promoters that are upregulated and downregulated in these cells relative to the CD24\textsuperscript{+} and “differentiated” clusters (Fig. 3d). Conversely, the expression patterns of promoters are similar between the CD24\textsuperscript{+} and “differentiated” clusters (Fig. 3d, Supplementary Table 4). Motif analysis of the up/down regulated classes of promoters in CD34\textsuperscript{+}/CD38\textsuperscript{−} cells showed that different core promoter motifs are enriched in each class, with the TATA box and Motif 5 being enriched in the upregulated class and MTE and TCT motifs being enriched in downregulated class (Fig. 3e). This result suggests that differences in core promoter usage might be driving the differences between CD34\textsuperscript{+}/CD38\textsuperscript{−} and the other clusters. Because the TATA box is mostly found in developmental core promoters, the CD34\textsuperscript{+}/CD38\textsuperscript{−} subpopulation likely reflects the more “stem-like” cellular environment in these cells. Our analysis highlights the ability of scMPRA to identify CRSs with differential activity in rare cell populations.
Fig. 3: scMPRA detects sub-state-specific CRS activity. (a) PCA plot of K562 cells classified by their cell cycle scores. (b) Heatmap of core promoter activities in different cell cycle phases (Color bar indicates housekeeping (blue) vs developmental (red) promoters). Core promoter activities have been normalized within each cell cycle phase to highlight the differences between housekeeping and developmental promoters. (c) UMAP embedding of K562 cells with high proliferation sub-states (CD34+/CD38- and CD24+). (d) Hierarchical clustering showing two clusters ("up" and "down") based on expression patterns in the three substates. The promoter activities are plotted as their z-score from the average across cell states to highlight the difference between cell states. (e) Proportion of promoters in the up and down clusters that contain the indicated core promoter motif. * represents significant enrichment in one cluster over the other (p < 0.05, Fisher's exact test).

scMPRA is reproducible and accurate in murine retinas

To demonstrate that scMPRA is applicable in a complex tissue with multiple cell types, we performed experiments in explanted murine retinas. Intact retina from newborn mice can be cultured and transfected ex vivo. This system has been useful for bulk MPRA experiments\textsuperscript{13,19,33}, but the results from those experiments report the aggregate expression of library members across the cell types of the retina. Performing scMPRA in ex vivo retina provided a chance to assay an
MPRA library in a living tissue with multiple cell types in their proper three-dimensional organization.

For this analysis we designed a library consisting of two independently synthesized wild-type copies, and 113 variants, of the full-length Gnb3 promoter (115 library members, Supplementary Table 5). We chose the Gnb3 promoter because it has high activity in photoreceptors and bipolar cells, but lower expression in other interneurons (i.e. amacrine cells) and Müller glia cells. The library contains mutations in the known transcription factor binding sites (TFBSs) in the Gnb3 promoter as well as mutations that scan across two phylogenetically conserved regions of the promoter (details in Fig. 6 below). We constructed this library of Gnb3 promoter variants using the double barcoding strategy described above, with one key modification that we now describe.

In the Gnb3 promoter library we addressed the inability of scMPRA to measure silent library members. In the first iteration of scMPRA, when a library member produces no mRNA barcodes its corresponding plasmid cannot be detected, and thus, a cell containing a silent plasmid is indistinguishable from a cell without a plasmid. To avoid this potential problem in our retina experiments, we included an additional cassette on the Gnb3 promoter library that allows us to detect the presence of plasmids carrying silent promoter variants. We included a cassette in which the U6 promoter drives the expression of a second copy of the cBC coupled to the 10x Capture Sequence (Fig. 4a). The U6 promoter drives strong RNA Polymerase III-dependent transcription, and is independent of the activity of the Gnb3 promoter. While we do not expect interference between the pol III-dependent U6 promoter and the pol II-dependent Gnb3 promoter variants, we minimized this possibility by putting the U6 cassette downstream of the Gnb3 variants and placing a polyA signal between the cassettes. The Capture Sequence is a specific sequence that is typically used to identify gRNAs in Perturb-seq experiments, but we use it here to isolate U6 expressed cBCs (Fig. 4b). When a cell contains a U6 cBC without the corresponding Gnb3 promoter cBC, it indicates the presence of a silent library member.

We introduced the Gnb3 promoter variant library into newborn mouse retinas and assessed the cell types into which the library entered by scRNA-seq (Fig. 4c). We obtained a total of 22,161 cells from two replicate experiments with a mean of 22,528 reads per cell and 1,642 genes per cell. The scRNA-seq data showed that we recovered rod photoreceptors (87.3%), bipolar cells (3.5%), interneurons (i.e. amacrine cells) (5.2%), and Müller glia cells (3.9%) (Fig. 4d, Supplementary Figure 3 a,b).

We then computed the expression of each Gnb3 promoter variant in each cell type by sequencing the Gnb3-expressed barcodes and the U6 barcodes from single cells. Cells with U6-expressed cBC counts, but no Gnb3-expressed cBC counts, represented cells in which that promoter variant was silent. On average, Gnb3 promoter variants were silent in 22% of cells, but this number varied widely (Fig. 4e) and was linearly related to the strength of the promoter, with stronger promoters expressing in a larger fraction of cells (Supplementary Fig. 3c). Using both the Gnb3-driven and U6-driven counts allowed us to compute the average expression of a promoter variant across all the cells of a given cell type, while still accounting for cells in which that promoter variant is silent (Methods).

Biological replicates measurements of the Gnb3 promoter variant library were reproducible in all four cell types (Fig. 4f, Supplementary Table 6). Reproducibility was highest in rod cells (Spearman’s p: 0.97, Pearson’s R: 0.98) because rod cells are the most abundant
Fig. 4: scMPRA design and workflow in mouse retina. (a) Schematic of Gnb3 promoter library constructs. In addition to the cBC and rBC barcodes, the Gnb3 promoter library contains an additional cassette in which the constitutive U6 promoter expresses a second copy of the cBC with a capture sequence for isolating these transcripts on gel beads. (b) Two different types of transcripts produced from the Gnb3 promoter library to measure promoter expression and detect unexpressed promoters respectively. The two types of transcripts originating from the same cell share the same 10x cell barcodes. (c) Experimental workflow for scMPRA in ex vivo mouse retinas. (d) UMAP of all cells measured in scMPRA with four major cell types identified. (e) For each Gnb3 variant in the library, we determined the proportion of cells that contain barcoded poly(A) transcripts out of all the cells that contained the variant. (f) Reproducibility of promoter activities between biological replicates in each of the four major cell types.

cell type in the mouse retina. The reproducibility was slightly lower in the rarer cell types (bipolar cells: Spearman’s ρ: 0.88, Pearson’s R: 0.92, Müller glia: Spearman’s ρ: 0.93, Pearson’s R: 0.95, and interneurons: Spearman’s ρ: 0.95, Pearson’s R: 0.98), but remained high enough to assess the expression of individual library members. We determined how reproducibility scales with the number of cells in scMPRA by subsampling the expression data. The minimum number of cells required for reproducible measurements (Spearman’s ρ > 0.75) of mean reporter gene levels is
Fig. 5: scMPRA recapitulates Gnb3 expression patterns. (a) The expression of the wild-type Gnb3 promoter in scMPRA reflects endogenous expression levels of Gnb3 in the respective cell types (Error bar denotes 95% C.I from two biological replicates). (b) The expression of the entire Gnb3 library in different cell types also follows endogenous Gnb3 expression (****: p-value < 0.0001 for Mann-Whitney U test). (c) scMPRA recapitulates the effects of a known Gnb3 variant, where the CRX3\textsuperscript{Q50}/CRX5\textsuperscript{Q50} variant reduces expression in bipolar cells specifically (*: p-value < 0.05 for Welch’s t-test).

75 cells (Supplementary Fig. 3d). Our results show that scMPRA works well for measuring reporter gene levels across cell types in complex tissues using small numbers of cells.

Two additional observations suggest that scMPRA measurements are accurate in ex vivo retinas. First, the expression of the wild type Gnb3 reporter, as well as the average expression of all Gnb3 promoter variants, correlates with endogenous Gnb3 expression in the corresponding scRNA-seq data (Fig. 5a,b). Second, our scMPRA data reproduced the known effect of a cell type specific Gnb3 promoter variant. Murphy et al.\textsuperscript{42} showed that altering two of the K50 homeobox sites in the Gnb3 promoter to Q50 sites reduces expression in bipolar cells while leaving expression in rod cells relatively unaffected. We observed the same reduction in bipolar cells when compared with rod cells for this same mutant (Fig. 5c). In addition, scMPRA also revealed that this mutant shows increased activity in Müller glia and interneurons. Taken together, these observations demonstrate that scMPRA is reproducible and accurate when applied to cell types in a complex tissue.

scMPRA reveals cell-type specific promoter variants.

The Gnb3 library was designed to probe components of the promoter including five binding sites for K50-type homeodomain TFs, an E-box binding site, and two evolutionarily conserved regions (Fig. 6a). In this experiment we define the effect of a mutation as its relative fold-change to the WT Gnb3 promoter in each cell type because the Gnb3 promoter is expressed at different levels across cell types (Fig. 5a). We labeled the homeobox sites as Cone Rod Homeobox (CRX) sites because CRX is a K50-homeodomain protein which plays an important role in rods and bipolar cells and is required for Gnb3 expression. Inactivating mutations in any individual CRX site decreased Gnb3 reporter expression in bipolar and rod cells, but deletion of either CRX1 or CRX5 also resulted in increased expression in interneurons (Fig. 6b). The CRX2 disruption had the largest effect on expression, and mutating the CRX2 site in combination with any other CRX site also caused large reductions in expression in rods and bipolar cells. Single and double swaps
Mutations in the *Gnb3* promoter display cell-type specific effects. (a) Schematic of the *Gnb3* promoter showing the location of the five CRX binding sites and the E-box. (b) Effects of individual and pairwise deletions of CRX binding sites. (c) Effects of individual and pairwise mutations of CRX K50 binding sites to Q50 binding sites. (d) Effects of changing CRX binding site affinities. (e) Effects of saturation mutagenesis of the E-box. (f) Effects of shuffle mutants in conserved regions of the *Gnb3* promoter. Each region was split into 5bp windows and the nucleotides in each window were shuffled. Labels above the heatmap indicate locations where the mutations impact CRX or E-box binding sites. All plots show log2 fold changes of the mutant relative to WT *Gnb3* expression in that cell type. Stars above the plot indicate a significant cell-type specific effect by one-way ANOVA.

of K50 CRX binding sites with Q50 binding sites tended to yield cell-type specific effects, primarily because interneurons displayed larger responses to the Q50 swaps compared with rod and bipolar cells (Fig. 6c). Increasing the affinity of CRX sites tended to have mild effects on expression in rods and bipolar cells, but increased expression significantly in interneurons (Fig.
The results from modifying CRX sites demonstrated that perturbations to single binding sites can produce cell-type specific effects.

We next examined the effects of single nucleotide changes in the E-box binding site (Fig. 6e). Helix-Loop-Helix (bHLH) transcription factors, which bind E-box motifs, are critical for the development of multiple retinal cell types\(^3\). Several single-nucleotide substitutions in the E-box resulted in strong effects on expression, although only one substitution produced significant cell-type specific effects. While the E-box is critical for strong expression of the Gnb3 promoter, subtle changes to its sequence do not generally result in cell-type specific changes to its activity.

To examine the effects of more severe sequence changes, and to assess the effects of perturbations outside the known TFBSs, we tiled mutations through the two evolutionarily conserved regions shuffling 5 bp at a time (Fig. 6f). Mutations in all six TFBSs resulted in cell-type specific changes in expression, but several mutations in the Gnb3 promoter outside of the known TFBS also resulted in cell-type specific changes in expression. Thus, other information in the Gnb3 promoter provides important cell-type context for the functioning of the CRX and E-box motifs.

Our analysis of the Gnb3 promoter shows that single-binding site and single-nucleotide variants can result in cell-type specific changes to cis-regulation, and that scMPRA is a powerful tool for identifying these changes across cell-types in mammalian tissues. The cis-regulatory logic of the Gnb3 promoter keeps it expressed at high levels in rods and bipolar cells in the early postnatal period, and at much lower levels in interneurons, which we speculate is why most cell-type specific perturbations result in effects of different sizes in interneurons when compared with rods and bipolar cells.

**Discussion**

We have presented a single-cell MPRA method to measure the cell-type and cell-state specific effects of CRSs. We demonstrated that scMPRA detects cell-type specific reporter gene activity in a mixed population of cells as well as in living retinas, and cell-state specific activity in isogenic K562 cells. The assay is reproducible and reports accurate mean levels of reporter gene activity in as few as 75 cells in a complex tissue. New methods that increase the number of single cells measured per experiment\(^3\) will increase the size of libraries that can be assayed by scMPRA. The dynamic range was relatively small in this study (8-fold between the strongest and weakest Gnb3 variants), which may reflect the activity of these specific sequences, but may also arise from the low efficiency of mRNA capture in single cells. scMPRA will therefore benefit from continuing improvements of methods to capture and recover mRNA from single cells.

With the burgeoning of Adeno-associated viral delivery systems\(^3\)–\(^4\), we anticipate that scMPRA will be widely used to study cis-regulatory effects in a variety of complex tissues. Given the hypothesis that non-coding variants with cell-type specific effects underlie a large fraction of human disease, an important application of scMPRA will be to identify genetic variants with cell-type specific cis-regulatory effects.
Methods

Cell Culture
K562 cells were cultured in Iscove’s Modified Dulbecco’s Medium (IMDM) + 10% Fetal Bovine Serum (FBS) + 1% non-essential amino acids + 1% pen/strep at 37°C with 5% CO₂. HEK293 cells were cultured in Eagle’s Minimum Essential Medium (EMEM) + 10% Fetal Bovine Serum (FBS) + 1% pen/strep at 37°C with 5% CO₂.

Core Promoter Library Cloning
We developed a two-level barcoding strategy to enable single-cell normalization of plasmid copy number. We applied this strategy to a library of core promoters we previously tested by bulk MPRA²⁴. That core promoter library contains 676 core promoters, each with a length of 133bp. The library cloning was done in three steps: First, we synthesized a library of 676 core promoters each barcoded with 10 different cBCs and cloned this library into a backbone²⁴. In a second step, a dsRed fluorescent reporter cassette was cloned between each core promoter and its associated cBCs as described²⁴. Thirdly, we modified this library for scMPRA by adding random barcodes downstream of the cBCs, but upstream of the polyA site.

To add the random barcodes (rBCs) we synthesized a single-stranded 90 bp DNA oligonucleotide (oligo) containing a 25 bp random sequence (the rBC), a restriction site, and 30 bp homology to the library vector on each side of the rBC region. We used NEBuilder® HiFi DNA Assembly Master Mix (E2621) to clone this oligo into the core promoter library. 4μg of the plasmid library was split into four reactions and digested with 2μl of SalI for 1.5 hours at 37°C. The digested product was run at 100V for 2 hours on a 0.7% agarose gel. The correct size band was cut and purified with the Monarch Gel Extraction Kit (New England BioLabs T1020L). The insert single-stranded DNA was diluted to 1 μM with H2O. Three assembly reactions were pooled together, each reaction containing 100 ng of digested library backbone, 1μM of insert DNA, 1μl of NEBuffer 2, 10 μl of 2X HiFi assembly mix, and H2O up to 20 μl. The reaction was incubated at 50°C for 1 hour. The assembled product was purified with the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030L) and eluted in 12 μl of H2O.

The assembled plasmid was transformed using Gene Pulser Xcell Electroporation Systems by electroporation (BIO-RAD 1652661) into 50 μl of ElectroMax DH10B electrocompetent cells (Invitrogen 18290015) with 1 μl of assembled product at 2 kV, 2000 Ω, 25 nF, with 1 mm gap. 950 μl of SOC medium (Invitrogen 15544034) was added to the cuvette and then transferred to a 15 ml Falcon tube. Two transformations were performed, and each tube was incubated at 37 °C for 1 hour on a rotator with 300 rpm. The culture was then added to pre-warmed 150 μl LB/Amp medium and grown overnight at 37°C. 1 μl of the culture was also diluted 1:100 and 50 μl of the diluted cultured was plated on a LB agar plate to estimate the transformation efficiency. For the core promoter library, we prepared DNA from more than 4x10⁸ colonies. Shallow sequencing of this library (below) showed that the majority of library members encoded unique cBC-rBC combinations.
**Gnb3 Promoter Variant Library Design and Cloning**

The Gnb3 library was designed to probe components of the promoter including five binding sites for K50-type homeodomain TFs, an E-box binding site, and two evolutionarily conserved regions. We labeled the K50 homeobox sites as Cone Rod Homeobox (CRX) sites because CRX is a K50-homeodomain protein required for Gnb3 expression and a key-lineage determining factor in retina, even though other K50-type homeobox proteins are also expressed in retinas. To test whether the disruption of CRX sites in the Gnb3 promoter has cell-type specific effect, we made the following three types of mutations: (1) All individual and pairwise deletions of the CRX binding sites by mutating the CRX sites to 5'-CTACTCCC-3'. (2) All individual and pairwise mutations of CRX binding sites from K50 homebox to Q50 homeobox motifs: 5'-CTAATTAC-3'. (3) All individual mutations of CRX binding sites to high (5'-CTAATCCC-3'), medium (5'-CTAAAGCCC-3') and low affinity (5'-CTTATCCC-3') K50 homeobox sites. Our unpublished data suggested that the E-box is important for the Gnb3 promoter activity and E-box motif is bound by many neuronal specific TFs, hence we mutated each base pair in the E-box to every other base pair and made pairwise mutations of the two core base pairs in the E-box motif. Lastly, we took an unbiased approach to screen for potential cell-type-specific mutations by shuffling mutations across the two conserved regions in the Gnb3 promoter. Each conserved region was tiled into 5bp windows and the nucleotides within each window were shuffled. All library sequences and the corresponding cBCs can be found in Supplementary Table 5.

The library of Gnb3 promoter variants was constructed in four steps. In the first step, we cloned the Gnb3 promoter variant library into the core promoter library vector backbone. We ordered double-stranded DNA fragments from Integrated DNA Technologies (Coralville, Iowa) encoding the varying part of the (520 bp) Gnb3 promoter and 113 promoter variants. The wild type Gnb3 promoter sequence was included twice, each time fused to a different cBC. The DNA fragments were manually pooled and cloned together as a library. In the second step, we cloned the remaining Gnb3 promoter (300bp) and an mEmerald reporter cassette between the Gnb3 promoter variants and the first cBC copy using HiFi assembly. In the third step, we used NEB HiFi DNA Assembly Master Mix (New England Biolabs E2621) to insert the U6 promoter between the two copies of the cBCs where it drives expression of the downstream copy of the cBC. In the fourth step, we introduced high-complexity rBCs between the first cBC and the U6 promoter. We synthesized a DNA oligo containing a 25 bp random sequence (the rBC), a restriction site, and 30 bp homology to the library vector on each side of the rBC barcode region. We then used HiFi Assembly to clone the rBC oligos into the Gnb3 promoter variant library. In this final library each plasmid contains a Gnb3 promoter variant driving mEmerald with a unique cBC-rBC combination in its 3’ UTR, which is followed by a polyA signal and the U6 promoter driving a second copy of the cBC, a capture sequence, and a termination signal. A total of eight HiFi Assembly reactions were pooled together to increase the library complexity. This library was transformed and amplified in E. coli as described above, and DNA was prepared from 2x10^9 colonies.

**Estimating Library Complexity**

To estimate the complexity of the core promoter library, we sequenced the DNA library using a nested PCR-based Illumina library preparation protocol. Briefly, we first used Q5 polymerase (New England BioLabs M0515) to amplify the region containing the two barcodes with SCARED...
17 and SCARED P18 (primer sequences can be found in Supplementary Table 8). The total reaction volume was 50μl using 50ng of plasmid library with 2.5 μl of 10uM primer each. After 25 cycles of amplification the product was purified with the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030), and eluted with 20 μl of ddH2O. For the second round of PCR we used the primers SCARED P19 and SCARED P20 in a 25 μl reaction with 0.25 μl product from the previous step. After 10 rounds of amplification the product was purified using the Monarch PCR&DNA Cleanup kit. For the last PCR we added the P5 and P7 Illumina adapters with SCARED P5, SCARED P7 with 10 cycles of amplification in a 25 μl reaction with 2 μl of purified product. This final product was sequenced on an Illumina MiSeq, and we obtained a total of 1,693,933 reads. After filtering out reads without a cBC or rBC of the correct length, we obtained a total of 1,359,176 reads (80% of the total reads) and 99.5% represented unique cBC-rBC pairs. For the Gnb3 library, we performed shallow sequencing, and obtained a total of 1,939,479 reads. After filtering out reads without correct cBC or rBC, we obtained a total of 1,838,415 reads (94.7% of the total reads). Among the 1,838,415 correct reads, 99.5% represented unique cBC-rBC pairs.

Estimating the Probability of Identical cBC-rBC Pairs in the Same Cell

We estimated the probability that more than one copy of a plasmid carrying the same cBC-rBC pair would be transfected into the same cell. We call this probability the collision rate. If the library is transfected into n cells, and a specific cBC-rBC pair is present at m copies in the library, then the expected number of collisions per experiment is given by:

\[ n^{-m} \sum_{k=0}^{n} \binom{n}{k} \left( \sum_{q=0}^{n-k} \binom{n-k}{q} \binom{m}{q} q! \left\{ \binom{m-q}{n-k-q} \right\} \right)_{n \geq 2} (n-k-q)!(m-q) \]

where k denotes the number of cells that received no plasmid, q denotes the number of cells transfected with exactly one plasmid, parentheses denote the binomial coefficient, and brackets denote the partition function. The above expression was simplified by substituting with the bivariate generating function, and the expected number of collisions per experiment is:

\[ m(1 - \left( \frac{n-1}{n} \right)^{m-1}) \]

The expected number of collisions per cell (\( \lambda \)) is given by,

\[ \lambda = \frac{m(1 - \left( \frac{n-1}{n} \right)^{m-1})}{n} \]

And, assuming collisions are a Poisson process, the probability of at least one collision in a cell is:

\[ P(\text{Collision}) = 1 - e^{-\lambda} \]
Using this framework, we can estimate the probability of a collision in our experiment. We assume one million cells (\( n \)) are transfected using 10 \( \mu g \) of plasmid DNA, and that the effective number of plasmids that enter the nucleus is 10% of that input amount (1 \( \mu g \))\(^{43}\) 1 \( \mu g \) of plasmid DNA is \( 2.3 \times 10^{11} \). Thus, the value of \( m \) in the nucleus is \( 2.3 \times 10^{11} \) divided by the number of unique members of the library. This allows us to calculate \( P(\text{Collision}) \) for a library of any given size. This framework shows that we require a library with \( 1.6 \times 10^6 \) unique members to achieve \( P(\text{Collision}) = 0.01 \). To be 99% sure that a library has at least \( 1.6 \times 10^8 \) unique members requires preparing that library 4.5 times as many independent colonies (\( 7.2 \times 10^8 \)), assuming a Poisson distributed library. The core promoter library was prepared from \( 4 \times 10^8 \) colonies, 55 times more than required for \( P(\text{Collision}) = 0.01 \), and the \( GnB3 \) variant library was prepared from \( 2 \times 10^9 \) colonies, 277 times more than required for \( P(\text{Collision}) = 0.01 \).

**Cell Line Transfections**

K562 cells were transfected with the core promoter library using electroporation with the Neon transfection system (Invitrogen MPK5000). One million cells were transfected with 2 \( \mu g \) of plasmid DNA (mixed-cell experiment) or 10 \( \mu g \) of plasmid DNA (K562 sub-state experiment), with 3 pulses of 1450 V for 10 ms.

HEK293 cells were transfected with the core promoter library using the Lipofectamine3000 protocol. 4 \( \mu l \) of p3000 reagent, 4 \( \mu l \) of Lipofectamine, and OptiMEM were mixed with 2 \( \mu g \) of plasmid DNA to a volume of 250 \( \mu l \). The lipofectamine reagents and plasmid were mixed and incubated at room temp for 15 minutes and then added dropwise to 1 million cells. We harvested K562 and HEK293 cells 24 hours after transfections for scMPRA.

**Ex vivo culturing and transfection of mouse retinas**

Retinas from newborn mice were dissected, cultured, and electroporated as described in White et al.\(^ {44} \). For each replicate experiment, three retinas were electroporated with 0.5 \( \mu g/\mu L \) of the \( Gnb3 \) promoter variant library, and co-electroporated with 0.5 \( \mu g/\mu L \) of a plasmid in which the \( Rhodopsin \) promoter drives the dsRED fluorophore. Electroporated retinas were harvested and dissociated as in Murphy et al.\(^ {42} \) with modifications as outlined below.

Briefly, three retinas/replicate were washed 3x in cold Hanks’ Balanced Salt Solution (HBSS) (Thermo Fisher) and were then incubated in 400 \( ul \) of HBSS containing 0.65 mg papain (Worthington Biochem) for 10 min at 37°C. 600 ul of Dulbecco’s Modified Eagle Medium (DMEM) (Thermo Fisher) containing 10% fetal calf serum (FCS) (Gibco) was added and the tissue was gently triturated with P1000 to achieve single cells suspension. 100 units DNase1 (Roche) was added, the cell suspension and incubated an additional 5 min at 37°C. Cells were centrifuged at 400g for 4 min then resuspended in 600 ml of sorting buffer (2.5 mM EDTA, 25 mM HEPES, 1% BSA in HBSS) and passed through a 35um filter and used directly for Fluorescence Activated Cell Sorting (FACS).

Because the majority of cells in murine retinas are rod photoreceptors, we attempted to enrich other cell types using FACS. The co-electroporated Rhodopsin-DsRed construct marks rod cells specifically. Therefore, we used FACS to generate a 1:1 mixture of dsRED\(^{+} \) to dsRED\(^{-} \) cells from
dissociated retinas. This procedure should yield a mix of cells in which rod cells comprise 50% of the total cells. In practice, rod cells still comprised 87% of the cells that were analyzed by scMPRA.

**Bulk MPRA from Cell Lines**
For both K562 cells and HEK293 cells, we harvested cells 24 hours after transfection, extracted total mRNA, and performed reverse transcription using the Superscript IV Reverse Transcriptase kit (Invitrogen 18090010). A sequencing library was then constructed using a nested PCR strategy. Briefly, we first used Q5 (New England BioLabs M0515) polymerase to amplify the region containing the cBC-rBC barcodes with SCARED P17 and SCARED P18. The total reaction volume was 50μl with 50ng of backbone and 2.5 μl of 10uM primer each. After 25 cycles of amplification the product was purified with the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030L) and eluted with 20 μl of ddH2O. Sequencing adapters were then added with 2 rounds of PCR, 10 cycles each. 0.25μl of PCR product was used with SCARED P19 and P20, then purified using the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030L). Finally, we used 2 μl of this product with primers SCARED P5 and SCARED P7, and then sequenced the resulting product on an Illumina Mi-seq instrument. The activity of library members was computed as described previously.

**Single-cell RNA-seq for scMPRA**
To perform scMPRA we targeted 2000 cells from the HEK/K562 mixed pool per replicate for each mixed cell experiment, 2500 cells per replicate for the K562-only experiment, and 2500 cells (after sorting) per replicate for the retina experiment. The cells were prepared according to the 10x Chromium Single Cell 3’ Feature Barcode Library Kit (PN-1000079) protocol.

Our goal was two-fold: to quantify the cBC-rBC pairs from each single cell and to sequence the cellular mRNAs from those same single cells. We captured all polyadenylated RNAs (barcoded reporter RNAs and cellular mRNAs) from single cells following the manufacturer’s protocol up to the cDNA amplification step.

For the cellular mRNAs (transcriptome), we followed the 10x protocol, using 1/4 of the cDNA library to generate dual-indexed transcriptomes. To quantify the cBC-rBC pairs, we performed separate PCRs using primers specifically targeting the reporter gene to improve barcode recovery efficiencies. Because the 10x protocol only uses 1/4 of the generated cDNA, we separately amplified the barcodes from another 1/4 of the pellet cleanup. We first used Q5 (New England BioLabs M0515) polymerase to amplify the region containing the cBC-rBC pairs with SCARED P17 and SCARED P18 with 10 cycles. The sample was divided equally into eight PCR reactions, each with 50 μl of total volume to reduce possible jackpotting. The product was then purified with the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030L), and eluted with 20 μl of ddH2O. We then added sequencing adapters using an additional two rounds of PCR. The first adapter PCR was performed with SCARED P21 and SCARED PP2 with a total of 10 ng of product from the barcode PCR. Again, we pooled eight PCR reactions, each with 50 μl of total volume and 10 PCR cycles. The PCR product was purified using the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030L). For the last PCR, to add the P5 and P7 Illumina adapters, we
used the primers SCARED P45, SCARED PP3 with 10ng of product and pooled eight PCR reactions, each with 50 μl of total volume and 10 PCR cycles.

For the U6 promoter library construction, we followed Step 4 of the 10x feature barcoding library preparation protocol (Chromium Next GEM Single Cell 3’ Reagent Kits v3.1 (Dual Index) CG000316 Rev C).

The transcriptome and barcode libraries were combined and sequenced on the Illumina NextSeq 500 with 28x105 cycles. Read1 was limited to 28bp to avoid sequencing the constant poly(A) sequence.

**scRNA-seq data processing**

The single-cell RNAseq data was processed using Cellranger 6.0.1 and Scanpy 1.8.1 following the standard pipeline. Briefly, different sequencing runs from the same biological replicate were pooled together and processed with CellRanger 6.1.1; the final output expression matrix was then imported into Scanpy for further processing. We first removed cells with less than 1000 genes, genes that were present in less than three cells, and cells with high counts of mitochondrial genes. Next, we normalized the UMI counts to the total cell UMI counts. The normalized expression matrix was used for clustering and visualization with Scanpy.

**scMPRA Data Processing**

For each promoter library, paired-end reads generated from barcoded reporter RNAs were processed with custom scripts that can be found on Github ([https://github.com/barakcohenlab/scMPRA](https://github.com/barakcohenlab/scMPRA)). In each paired-end read, Read1 contains a 10x cell barcode and a UMI, while Read2 contains the cBC and rBC sequences. We define a “quad” as a 10x Cell Barcode, UMI, cBC, and rBC originating from the same individual paired-end read. To tabulate the cBC-rBCs we first matched the constant sequences flanking both barcodes, filtering out reads where either barcode was not the correct length. We performed this filtering using a stand-alone program ([https://github.com/szhao045/scMPRA_parsingtools](https://github.com/szhao045/scMPRA_parsingtools)). Second, we filtered out incorrect 10x Cell Barcodes based on the CellRanger output barcode list using error-correction with a maximum hamming distance of one. Third, to mitigate the effect of template-switching during the PCR steps, we plotted the rank read depth for each unique quad and identified an “elbow point” at a minimum depth of 1 read for the mixed-cell and the retina experiment, and 10 reads for the K562 alone experiment. We kept all reads above the minimum depth, and also kept a low-depth unique quad if it contained a cBC-rBC matching a high-depth pair with a hamming distance of at most one. Lastly, for the mixed-cell experiment and the K562 cell alone experiment, we removed any cell with less than 100 scMPRA-associated UMIs, since the scMPRA reads from those cells were poorly sampled. For the last step, because the retina experiment contains additional information from the U6 promoter, we did not threshold based on cells. Since U6 promoter data provides information on whether a given cBC in a given cell is sampled well, we removed all unique barcode pairs containing only 1 UMI for a cBC.
Calculating the Single-Cell Activities of Promoters

Once the high-confidence quads were identified, we computed $A$, the activity of a promoter in an individual cell using,

$$A = \frac{\sum_{i=1}^{n} \text{UMI count for } cBC_i}{\sum_{i=1}^{n} \text{rBC count for } cBC_i}$$

where $n$ is the number of unique cBCs that mark a single promoter in the library, and the UMI and rBC counts are summed over all quads with a given 10x cell barcode. We then compute $C$, the cell-type specific activity of a promoter as,

$$C = \frac{\sum_{j=1}^{m} A_j}{m}$$

where $m$ is the number of cells in a given cell type, and all 10x cell barcodes assigned to a given cell type are identified from their matched scRNA-seq profiles. For scMPRA data from the retina we modified the equation for cell-type specific activity as follows,

$$C = \frac{\sum_{j=1}^{p} A_j}{P + U}$$

where $P$ is the number of cells of a given cell type in which $Gnb3$-driven cBCs were detected and $U$ is the number of cells of that cell type for which a U6 promoter cBC was detected without detecting any corresponding $Gnb3$-driven cBC. This modification has the effect of adding activities of zero for all cells with U6-driven cBCs that did not express a $Gnb3$-driven cBC.

Cell cycle analysis

Cell cycle analysis for the scRNA-seq experiment was done with Scanpy 1.8.1 with cell cycle genes. The expression profile of each cell was projected onto a PCA plot based on the list of cell cycle genes using Scanpy.

Motif analysis

The core promoters were first clustered according to their expression levels in the different cell sub-state populations by hierarchical clustering. We categorized our data into up/down regulated clusters at the first branching point, aiming to preserve the large structure. We then identified core promoter motifs in each promoter according to the parameters in Zabidi et al using MAST v4.10.0 and plotted the proportion of promoters containing each motif in each promoter class.

Statistical Analyses

All statistical analyses were done using Python 3.9.6, Numpy 1.12.1, Scipy 1.6.3 and R 4.0.2.

Ethics

This study was performed in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. All of the animals were handled according to protocol # A-3381-01 approved by the Institutional Animal Care and Use Committee of Washington University in St. Louis. Euthanasia of mice was performed according to the recommendations of the American Veterinary Medical Association Guidelines on
Euthanasia. Appropriate measures are taken to minimize pain and discomfort to the animals during experimental procedures.

Data and Code Availability
Next-generation sequencing data that support the findings of the study are available in the Gene Expression Omnibus using accession code GSE188639.

The code that supports the findings of this study is available as a Github Repository (https://github.com/barakcohenlab/scMPRA).

Acknowledgements
We thank the members of the Cohen laboratory for their critical feedback on the manuscript. We thank Jess Hoistington-Lopez and MariaLynn Crosby for assistance with high-throughput sequencing. This work is supported by grants to B.A.C from the National Institutes of Health, R01 GM140711 and R01 GM092910, and by a grant to J.C.C from the National Institutes of Health R01 EY030075.

Author Contributions
S.Z. and B.A.C. conceived and designed the project. All experiments and analyses were performed by S.Z. with technical contributions from C.K.Y. H. and D.M.G., except for the electroporation and culturing of mouse retinas, which was performed by C.A.M. J.C.C and M.A.W provided critical input into the design of the Gnb3 promoter variant library. S.Z., C.K.Y.H., and B.A.C. wrote the manuscript with input and feedback from all authors.

Competing Interests
S.Z. and B.A.C. are inventors on a pending patent filed by Washington University in St. Louis which may encompass the methods, reagents, and data disclosed in this manuscript. B.A.C is on the scientific advisory board of Patch Biosciences.
Supplementary Figures

Supplementary Fig. 1: scMPRA measures cell-type specific CRS activity. (a) UMAP of the single-cell transcriptome from the mixed-cell experiment. 105 out of 3417 cells (3%) are labeled by both K562 and HEK293 cell genes. (b) UMAP of the mixed-cell experiment with cells marked by other representative markers for K562 and HEK293 cell expression. (c,d) Histogram of the number of plasmids (unique cBC-rBC pairs) transfected into K562 cells and HEK293 cells. (e,f) Histogram of the mean number of rBC per cBC (CRS) per cell for K562 cells and HEK293 cells. (g,h) Correlation of bulk MPRA versus scMPRA where only the scMPRA data has been UMI normalized. (i,j) Scatterplot of scMPRA reproducibility for housekeeping and developmental promoters in K562 cells and HEK293 cells.
Supplementary Fig. 2: scMPRA measures CRS activity in K562 cell substates. (a) Reproducibility for mean expression of core promoters in K562 cells. (b) Correlation of bulk and scMPRA (non-UMI corrected) in K562 cells (c) Different dynamics of expression. For *UBA52*, the promoter is most highly expressed in S phase; whereas for *CSF1*, the promoter is most highly expressed in G1 phase. For *CXCL10*, the promoter is expressed evenly through cell cycle (Stars indicate significance from Wilcoxon rank sum test, *: p < 0.05) (d) Cells no longer cluster together based on cell cycle genes after the effects of the cell cycle are removed.
Supplementary Fig. 3: Robust measurements of Gnb3 promoter library in ex vivo retina. (a) Expression of marker genes by scRNA-seq used to identify cell types in the retina. (b) Percentage of the total cells recovered represented by each retinal cell type. (c) Plot showing the relationship between the mean activity of a Gnb3 promoter variant in a given cell type (x-axis) and the proportion of cells in which that promoter variant is silent (y-axis). Individual cells in which a given Gnb3 variant is silent are identified as cells with U6-expressed cBC, but no Gnb3-expressed cBC. (d) The correlation between biological replicates is plotted as a function of the number of cells used in the analysis.
References


46. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell


Appendix C: High-throughput screening of cell-type specific brain enhancers with single-cell MPRA

This work was performed in collaboration with Yawei Wu, Jie Li and Barak Cohen in our lab, Din Selmanovic with help from Allen Yen and Lite Yang in Joe Dougherty’s lab and Constance Myers in Joe Corbo’s lab. Yawei Wu, Barak Cohen and I conceived and designed the project. I designed and cloned the pilot library, designed the enhancer library and performed the single-cell experiments together with Yawei Wu. This study is still in progress, so the following represents the work that we have completed so far.

Introduction

The brain is a highly complex organ that is composed of clusters of circuits with diverse molecular and functional properties. The latest Allen Brain Institute study estimated that there are approximately 388 cell types in the mouse isocortex alone [1]. We have only just begun to scratch the surface of understanding the diversity of cell types in the brain with single-cell transcriptomics, but we are far from understanding how these cell types develop and contribute to proper brain function. Proper cell-type specification is controlled by cis-regulatory sequences (CRSs) such as enhancers. Identifying cell-type specific enhancers will not only help us understand how they regulate cell-type specific gene expression, but will also provide us with much-needed tools to selectively label and perturb cell types for further functional characterization [2].

To identify cell-type specific enhancers in the brain, researchers have profiled brain tissues
using a variety of epigenomic assays. Several groups have used either bulk ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) or single-nuclei ATAC-seq (snATAC-seq) to identify the open chromatin regions in different cell types, reasoning that the identified regions act as a proxy for active CRSs [3–7]. On the other hand, large-scale epigenomic maps generated by the ENCODE and Roadmap consortiums have identified various histone modifications that correlate with enhancer activity [8–10]. Thus, other groups have developed Cut&Tag-based single-cell methods to measure the levels of various histone modifications in mouse brain tissues, including H3K27ac and H3K4me3, which are characteristic enhancer modifications [11, 12].

While epigenomic modifications are useful proxies for enhancer identification, they are not always well-correlated with actual enhancer activity [10, 13, 14]. Confidently calling enhancers requires direct measurement of the enhancer’s activity in a given cell type. The gold standard for measuring enhancer activity is using an reporter assay, where the enhancer is cloned either upstream or downstream of a minimal promoter driving the expression of a reporter gene such as luciferase or the green fluorescent protein (GFP). High-throughput versions of reporter assays, collectively referred to as Massively Parallel Reporter Assays (MPRAs), have been developed to enable testing of hundreds or thousands of putative enhancers in the same experiment [15–17]. While MPRAs have been successfully applied in bulk mouse brain, the measured enhancer activities represent the aggregate across all cell types in the brain [4, 18, 19]. Yet, individual enhancers in the brain can be exquisitely cell-type specific [5, 6, 20, 21], necessitating the development of methods to measure cell-type specific enhancer activity.

To enable cell-type specific measurements of enhancer activity, Hvartin et al. developed Paralleled Enhancer Single Cell Assay (PESCA) [22]. Putative enhancer libraries are barcoded and packaged into Adeno-Associated Viruses (AAVs) for delivery into the mouse brain, and the activity of each library member is measured by single-nuclei RNA-seq (snRNA-seq). Using PESCA, the authors successfully identified enhancers that are specifically expressed in SST+
interneurons. However, this method only measures RNA expression from the library without normalization for DNA counts. DNA libraries contain substantial variation in the abundance of individual elements, which will bias RNA counts towards high abundance library members. To solve the DNA normalization problem in single cells, our lab previously developed a single-cell MPRA (scMPRA) [23]. In scMPRA, a second, extremely diverse random barcode (rBC) is added to the library in addition to the barcode tagging each CRS (cBC). Due to the diversity of the rBC, only one copy of each rBC should be present in a single cell. Thus, the number of rBCs per cBC in a cell represents the number of DNA copies of the cBC in that cell, and both the RNA and DNA copy numbers can be obtained from by single-cell RNA-sequencing (scRNA-seq). This allows us to normalize the activity of each CRS in each cell in parallel.

We previously showed that scMPRA can work in mouse retinal tissues by ex vivo electroporation of a plasmid library followed by 10x scRNA-seq [23]. Here, we adapt scMPRA by packaging the library into AAVs for in vivo delivery into the mouse cortex. Using a pilot library consisting of a single CRS with random barcodes, we tested two different modes of AAV delivery and validated that we could retrieve barcodes from the cortex by both 10x and sci-RNA-seq (Single cell Combinatorial Indexing RNA sequencing) [24, 25]. We also designed a library of putative cell-type specific enhancers to screen with scMPRA. This study expands the toolbox of methods for screening cell-type specific CRSs in vivo.

Results

in vivo targeting of mouse cortex for scMPRA with AAVs

Researchers are increasingly turning to the use of AAVs for delivery of transgenes to tissues in vivo due to its safety, ease of use and potential to target specific tissues [26]. Recently, directed evolution methods been used to engineer AAVs that target the peripheral or central nervous systems with high efficiencies [27, 28]. To adapt scMPRA for use with AAVs, we first cloned a
pilot construct consisting of the CAG promoter driving the expression of a GFP reporter gene barcoded with a diverse library of rBCs in an AAV vector. The CAG promoter was selected to ensure high levels of expression across all brain cell types. We estimate that the library contains millions of rBCs. We then packaged the library into recombinant AAVs with the PhP.eB capsid to specifically target the central nervous system Figure C.1A.

The pilot library was designed to answer a few questions. First, what method of viral delivery should be used? The PhP class of AAVs have previously been shown to penetrate the blood-brain barrier, allowing for efficient in vivo delivery from tail vein injections [29]. However, the transduction efficiencies are variable and differ between different mouse strains [30]. On the other hand, we can use a more invasive delivery method, where P0 mice are subject to direct intracranial injections of the AAV library [31]. We sought to compare the efficiencies of both methods of AAV delivery using this pilot library. We delivered the same PhP.eB library
to mice either by tail vein injections or intracranial injections. After ~21 days, we dissected the cortex from the mouse brain and extracted nuclei from flash-frozen tissues to perform sci-RNA-seq. We modified the sci-RNA-seq protocol to be compatible with scMPRA by adding an additional PCR step to enrich for our library barcodes (Methods). Because millions of nuclei can be processed in parallel with sci-RNA-seq, we were able to profile two tail-vein injected and two intracranial injected mice cortexes in one experiment. Because we only wanted to get an estimate of the number of cell types targeted, we did not sequence the library very deeply. Nevertheless, we recovered 37,265 cells, ranging from 8209-11372 per cortex Figure C.1B-C. We then quantified the number of cells where we recovered at least one barcode from our pilot library. We find that using tail vein injections, we were only able to transduced about 5-10% of all cells, but the intracranial injections reached 30-40% Figure C.1D. Reassuringly, the numbers were similar between different replicates in different mice. Thus, we proceeded with the intracranial injections for downstream experiments.

Next, we wanted to test the tropism of PhP.eB in our hands. It was previously shown that PhP.eB has a preference for astrocytes over other cell types, though neurons (GABAergic and glutamatergic) and oligodendrocytes are also targeted at lower efficiencies [32, 33]. Knowing which cell types can be recovered is important to ensure that our enhancer library is designed for the correct cell types. For this experiment, we introduced our library by intracranial injections at P0, and waited 10 days for the virus to mature and express in brain cells. Because our sci-RNA-seq experiment showed that only ~40% of the cells would be transduced, we decided to sort for GFP+ cells by fluorescence-activated cell sorting (FACS) after cortex dissection to ensure that all the sequenced cells would contain a library member. Our FACS data also showed that 20-40% of the cells contained GFP, agreeing with our sci-RNA-seq data above Figure C.2A. We then performed scMPRA with the sorted cells using the 10x scRNA-seq kit using our previously established protocol [23]. In total, we confidently assigned cell types to ~11,700 cells using the associated transcriptomes Figure C.2B-C. While we were able to recover
Figure C.2: Quantification of AAV tropism and efficiencies. (A) Percentage of GFP-positive cells sorted by FACS. (B) t-SNE plot of all cell types recovered from 10x scRNA-seq after pilot. (C) Number of cells recovered for each cell-type. (D) Violin plot of number of barcodes recovered per cell. n represents the median number of barcodes for each cell type.

most of the major cell types in the cortex, there was a noticeable absence of glutamatergic neurons. This could be because glutamatergic neurons are not well-transduced by our AAVs, or because our gating strategy during FACS somehow excluded these neurons. Surprisingly, we also did not recapitulate the bias towards astrocytes. Instead, we obtained a large number of oligodendrocytes and microglia. Regardless, our results indicate that we would be well-powered to detect cell-type specific enhancers in the major cell types of the cortex.

Finally, we wanted to examine how many rBCs we could obtain per cell. In previous scMPRA experiments, we were able to recover hundreds of library members per cell via transfection or
electroporation [23]. However, a study using retro-orbital injections of PhP.eB with single-cell sequencing suggested that the copy number of plasmids per cell was on the order of magnitude of 1-10 [32], while a different study using imaging instead of sequencing suggested that the copy number was on the order of hundreds per cell [33]. The size of the library that we can test in one experiment will be constrained by the number of plasmids per cell, given our previous estimate that each CRS should be present ~30-50 cells in order to obtain a reproducible measurement. We recovered at least one rBC per cell in 97% of the cells. We then tabulated the number of rBCs that we observed per cell for each of the major cell types recovered, and found that they were all transduced at roughly the same efficiencies (median per cell is ~10), with numbers ranging up to about 100 rBCs per cell Figure C.2D. This is more than was previously measured by scRNA-seq, suggesting that we are getting good recovery of barcodes. Note that because the barcode was located ~900bp away from the 3’ UTR, the capture efficiencies were not optimal and these barcode estimates likely represent the lower bound. Given these numbers, we conservatively estimate that we can test a library size of about 400 elements in one experiment.

**Enhancer library design**

We designed a library of putative CRSs based on a combination of previous epigenomic experiments. Because we are only able to recover sufficient numbers of cells from the major cell types (neurons, astrocytes, oligodendrocytes and microglia), we focused only on these cell types for this library. We first downloaded scATAC-seq data from CATlas [7] and merged the peaks from the same major cell types. We then selected peaks that were only present in one major cell type and not any of the other cell types. This resulted in approximately 4000-18000 putative enhancers for each cell type. Because we could not test such a large library, we further filtered the putative enhancers using a variety of datasets. We identified enhancers that had a positive association with changes in RNA expression [7], high levels of H3K27ac and H3K4me3 from Paired-Tag experiments [11], large numbers of predicted TFBSs, overlap with BRD4 binding
Figure C.3: Overview of brain enhancer library for in vivo scMPRA. (A) Barplot of number of library members in each class. (B) Distribution of enhancer and promoter lengths in enhancer library. (C) Schematic of vector that the enhancer library is cloned into.

(only for astrocytes and neurons) [34]. We then ranked the enhancers by the number of features they overlapped with and selected the top 70 enhancers from each cell type for our library for a total of 280 putative enhancers. We also included 30 putative human enhancers from astrocyte-specific open chromatin regions obtained from human scATAC-seq data. In addition, we selected a 10 positive control enhancers that have previously been shown to be cell-type specific and 15 promoters from genes will cell-type specific expression in our pilot scRNA-seq data (details in Table C.1). The full library composition is summarized in Figure C.3A, and the enhancers range from about 400-800bp Figure C.3B. The library members were cloned into an AAV vector upstream of a minimal $\beta$-globin promoter, and a U6 promoter driving the expression of the same CRS barcode was included after the poly(A) tail, similar to the our previous scMPRA vector Figure C.3C [23]. The library was then packaged into AAVs and will be used for scMPRA.
### Table C.1: Positive control enhancers and promoters.

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>CRS Type</th>
<th>Target</th>
<th>Genome Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>mscCRE4</td>
<td>Enhancer</td>
<td>Neuron</td>
<td>Mouse</td>
<td>Graybuck <em>et al.</em> [5]</td>
</tr>
<tr>
<td>mscCRE10</td>
<td>Enhancer</td>
<td>Neuron</td>
<td>Mouse</td>
<td>Graybuck <em>et al.</em> [5]</td>
</tr>
<tr>
<td>mscCRE13</td>
<td>Enhancer</td>
<td>Neuron</td>
<td>Mouse</td>
<td>Graybuck <em>et al.</em> [5]</td>
</tr>
<tr>
<td>mscCRE16</td>
<td>Enhancer</td>
<td>Neuron</td>
<td>Mouse</td>
<td>Graybuck <em>et al.</em> [5]</td>
</tr>
<tr>
<td>GRE12</td>
<td>Enhancer</td>
<td>Neuron</td>
<td>Mouse</td>
<td>Hrvatin <em>et al.</em> [22]</td>
</tr>
<tr>
<td>GRE44</td>
<td>Enhancer</td>
<td>Neuron</td>
<td>Mouse</td>
<td>Hrvatin <em>et al.</em> [22]</td>
</tr>
<tr>
<td>hDLX156i</td>
<td>Enhancer</td>
<td>Neuron</td>
<td>Human</td>
<td>Dimidschstein <em>et al.</em> [21]</td>
</tr>
<tr>
<td>eHGT_022h</td>
<td>Enhancer</td>
<td>Neuron</td>
<td>Human</td>
<td>Mich <em>et al.</em> [6]</td>
</tr>
<tr>
<td>eHGT_023h</td>
<td>Enhancer</td>
<td>Neuron</td>
<td>Human</td>
<td>Mich <em>et al.</em> [6]</td>
</tr>
<tr>
<td>eHGT_140h</td>
<td>Enhancer</td>
<td>Neuron</td>
<td>Human</td>
<td>Mich <em>et al.</em> [6]</td>
</tr>
<tr>
<td><em>Gfap</em></td>
<td>Promoter</td>
<td>Astrocyte</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>Rps18</em></td>
<td>Promoter</td>
<td>All cells</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>Gapdh</em></td>
<td>Promoter</td>
<td>All cells</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>Aldh111</em></td>
<td>Promoter</td>
<td>Astrocyte</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>Tmem119</em></td>
<td>Promoter</td>
<td>Microglia</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>Slc2a5</em></td>
<td>Promoter</td>
<td>Microglia</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>Cx3cr1</em></td>
<td>Promoter</td>
<td>Microglia</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>Sp8</em></td>
<td>Promoter</td>
<td>Neuron</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>Stmn2</em></td>
<td>Promoter</td>
<td>Neuron</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>NeuroD2</em></td>
<td>Promoter</td>
<td>Neuron</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>Sox10</em></td>
<td>Promoter</td>
<td>Oligodendrocyte</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>Mag</em></td>
<td>Promoter</td>
<td>Oligodendrocyte</td>
<td>Mouse</td>
<td>-</td>
</tr>
<tr>
<td><em>SLC1A2</em></td>
<td>Promoter</td>
<td>Astrocyte</td>
<td>Human</td>
<td>-</td>
</tr>
<tr>
<td><em>GJB6</em></td>
<td>Promoter</td>
<td>Astrocyte</td>
<td>Human</td>
<td>-</td>
</tr>
</tbody>
</table>
Discussion

Our genomes encode for a surprising diversity of cell types that perform specific functions in the brain. However, most of the cells have not be functionally characterized due to a lack of tools to access and manipulate specific cell types. In this study, we adapt and improve scMPRA to profile putative enhancers in the mouse brain in vivo. This study is still ongoing, precluding any important biological findings. However, our preliminary results indicate that we can successfully perform scMPRA in an in vivo setting using AAVs. We designed a library that, to the best of our knowledge, is maximally powered for the detection of cell-type specific enhancers. We further show that scMPRA can be adapted to be performed with sci-RNA-seq in addition to 10x scRNA-seq, which will allow us to screen larger libraries and profile more cellular subtypes. We anticipate that this study will pave the way for performing scMPRA in other tissues that can be targeted with AAVs.

Methods

Pilot library cloning

The AAV plasmid was obtained from Addgene (#104061). The vector was digested with EcoRV and random barcodes (rBCs) were added by NEBuilder HiFi DNA Assembly (NEB #E2621). The library was then packaged into PhP.eB AAVs by the Hope Center Viral Vectors Core at Washington University in St. Louis.

sci-RNA-seq pilot experiment

We dissected the cortexes from mouse pups injected with the pilot library at P10. The cortexes were flash frozen and stored at -80°C until nuclei extraction. The tissues were homogenized with a glass douncer in hypotonic lysis buffer (2ml 10x-PBS-hypotonic stock solution,
2.28g sucrose, 30 μl of 2M MgCl$_2$, 50μl 10% IGEPAL, 200μl DEPC and nuclease-free water up to 20ml) [25]. The 10x-PBS-hypotonic stock solution contains 5.45g Na$_2$HPO$_4$ (dibasic), 3.1g NaH$_2$PO$_4$·H$_2$O, 1.2g KH$_2$PO$_4$, 1g KCl, 3g NaCl in nuclease-free water (final volume 500ml). The cells were spun at 500g for 5min at 4°C, and the nuclei were resuspended in 25% Iodixanol gradient and layered on 35% Iodixanol (OptiPrep Density Gradient Medium, Sigma #D1556). The nuclei were spun at 10,000g for 30min at 4°C, and the nuclei were extracted from the middle of the Iodixanol layers. Finally, the nuclei were washed in 0.3M SPBSTM buffer (28.5g sucrose, 25ml 10x PBS, 25ml 10% TritonX-100, 275μl 2M MgCl$_2$ in nuclease-free water up to 250ml) [25], then snap-frozen until sci-RNA-seq3 was performed.

We followed the sci-RNA-seq3 protocol as published [25] with the following modifications. First, we added a second oligo in the reverse transcription step. The second oligo is closer to the rBC and should capture the rBC better than the oligo(dT) primer. Second, we performed the first library amplification step with shorter primers (without sequencing adapters) and purified the reactions individually with Ampure beads. We then performed two separate PCRs to amplify the transcriptome and barcodes separately. For the transcriptome, we performed PCR with primers containing sequencing adapters. For the random barcodes, we first enriched for the barcode with primers specific to the reporter genes, then performed a third PCR with primers containing sequencing adapters. Both pools were then pooled, purified with Ampure beads and sequenced the library on the NextSeq 500.

10x pilot experiment

We dissected the cortexes from mouse pups injected with the pilot library at P10. The cortexes were then dissociated into single cells with the Adult Brain Dissociation Kit, mouse and rat (Miltenyi Biotech #130-107-677). The cells were then sorted for GFP+ cells on the CytoFLEX SRT Benchtop Cell Sorter (Beckman Coulter) and the 10x protocol was performed as previously described.
**Estimating library size**

We estimated the size of library based on the assumption that we would recover 10,000 cells in one experiment with a median 10 rBCs per cell. Assuming each major cell type is present in roughly equal numbers, after excluding the other smaller populations of cells and the cells that we cannot recover rBCs from, we conservatively expect to obtain about 2000 cells per cell type. Given that we will need about 50 cells to obtain a good estimate of CRS activity, we expect that we will be able to measure a library of about 400 elements with high confidence.

**Enhancer library design**

snATAC-seq peaks (candidate CREs, cCREs) were first downloaded from CATlas ([http://catlas.org/mousebrain]) [7]. The cCREs from sub-cell types were then merged using bedtools v2.27.1 [35] according to the cell types in (Table C.2). To find cell-type specific cCREs, merged cCREs were compared to find cCREs that are only present in one cell type. For neurons, we further filtered for cCREs that are present in >70% of sequenced nuclei. After filtering the cCREs to ensure they can be synthesized and do not contain the restriction enzymes needed for cloning, we grouped the cCREs by high or low H3K27ac and H3K4me1 levels (from CATlas, [http://catlas.org/pairedTag] [11]), high and low number of TF motifs (determined by FIMO [36]), association with gene expression (from CATlas) and BRD4 binding (data from [34]). cCREs were then ranked by scores based on the number of overlaps with each feature. A cCRE with high levels of histone modifications, high number of motifs, positive association with gene expression and BRD4 binding would have a score of 4, while a cCRE with only one feature would have a score of 1. We selected 70 cCREs with the highest scores from each category to maximize the chances of the CRE being active in at least one tissue.

The positive control enhancers were curated from papers that had previously reported cell-type specific activities Table C.1. The positive control promoters were selected from genes found using the FindMarkers function in Seurat [37]. The transcription start site (TSS) of each
**Table C.2: Sub-cell types merged for major cell types**

<table>
<thead>
<tr>
<th>Major cell type</th>
<th>Sub-cell type</th>
<th>Major cell type</th>
<th>Sub-cell type</th>
<th>Major cell type</th>
<th>Sub-cell type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC</td>
<td>ASCW</td>
<td>Neuron</td>
<td>SSTGA6</td>
<td>Neuron</td>
<td>NPGL4</td>
</tr>
<tr>
<td>ASC</td>
<td>ASCG</td>
<td>Neuron</td>
<td>SSTGA7</td>
<td>Neuron</td>
<td>NPGL5</td>
</tr>
<tr>
<td>OGC</td>
<td>MFOL</td>
<td>Neuron</td>
<td>SSTGA8</td>
<td>Neuron</td>
<td>PTGL1</td>
</tr>
<tr>
<td>OGC</td>
<td>MOL</td>
<td>Neuron</td>
<td>SSTGA9</td>
<td>Neuron</td>
<td>PTGL2</td>
</tr>
<tr>
<td>MGL</td>
<td>MGL</td>
<td>Neuron</td>
<td>SSTGA10</td>
<td>Neuron</td>
<td>PTGL3</td>
</tr>
<tr>
<td>Neuron</td>
<td>LAMGA1</td>
<td>Neuron</td>
<td>ITL23GL1</td>
<td>Neuron</td>
<td>PTGL4</td>
</tr>
<tr>
<td>Neuron</td>
<td>LAMGA2</td>
<td>Neuron</td>
<td>ITL23GL2</td>
<td>Neuron</td>
<td>PTGL5</td>
</tr>
<tr>
<td>Neuron</td>
<td>LAMGA3</td>
<td>Neuron</td>
<td>ITL23GL3</td>
<td>Neuron</td>
<td>PTGL6</td>
</tr>
<tr>
<td>Neuron</td>
<td>LAMGA4</td>
<td>Neuron</td>
<td>ITL4GL1</td>
<td>Neuron</td>
<td>PTGL7</td>
</tr>
<tr>
<td>Neuron</td>
<td>VIPGA1</td>
<td>Neuron</td>
<td>ITL4GL2</td>
<td>Neuron</td>
<td>PTGL8</td>
</tr>
<tr>
<td>Neuron</td>
<td>VIPGA2</td>
<td>Neuron</td>
<td>ITL5GL1</td>
<td>Neuron</td>
<td>CTGL1</td>
</tr>
<tr>
<td>Neuron</td>
<td>VIPGA3</td>
<td>Neuron</td>
<td>ITL5GL2</td>
<td>Neuron</td>
<td>CTGL2</td>
</tr>
<tr>
<td>Neuron</td>
<td>VIPGA4</td>
<td>Neuron</td>
<td>ITL5GL3</td>
<td>Neuron</td>
<td>CTGL3</td>
</tr>
<tr>
<td>Neuron</td>
<td>PVGA1</td>
<td>Neuron</td>
<td>ITL6GL1</td>
<td>Neuron</td>
<td>CTGL4</td>
</tr>
<tr>
<td>Neuron</td>
<td>PVGA2</td>
<td>Neuron</td>
<td>ITL6GL2</td>
<td>Neuron</td>
<td>CTGL5</td>
</tr>
<tr>
<td>Neuron</td>
<td>PVGA3</td>
<td>Neuron</td>
<td>ITL6GL3</td>
<td>Neuron</td>
<td>CTGL6</td>
</tr>
<tr>
<td>Neuron</td>
<td>PVGA4</td>
<td>Neuron</td>
<td>ITL6GL4</td>
<td>Neuron</td>
<td>L6bGL1</td>
</tr>
<tr>
<td>Neuron</td>
<td>PVGA5</td>
<td>Neuron</td>
<td>ITL6GL5</td>
<td>Neuron</td>
<td>L6bGL2</td>
</tr>
<tr>
<td>Neuron</td>
<td>PVGA6</td>
<td>Neuron</td>
<td>ITL6GL6</td>
<td>Neuron</td>
<td>L6bGL3</td>
</tr>
<tr>
<td>Neuron</td>
<td>PVGA7</td>
<td>Neuron</td>
<td>ITHGL1</td>
<td>Neuron</td>
<td>L6bGL4</td>
</tr>
<tr>
<td>Neuron</td>
<td>SSTGA1</td>
<td>Neuron</td>
<td>ITHGL2</td>
<td>Neuron</td>
<td>L6bGL5</td>
</tr>
<tr>
<td>Neuron</td>
<td>SSTGA2</td>
<td>Neuron</td>
<td>ITHGL3</td>
<td>Neuron</td>
<td>L6bGL6</td>
</tr>
<tr>
<td>Neuron</td>
<td>SSTGA3</td>
<td>Neuron</td>
<td>NPGL1</td>
<td>Neuron</td>
<td>CLAGL1</td>
</tr>
<tr>
<td>Neuron</td>
<td>SSTGA4</td>
<td>Neuron</td>
<td>NPGL2</td>
<td>Neuron</td>
<td>CLAGL2</td>
</tr>
<tr>
<td>Neuron</td>
<td>SSTGA5</td>
<td>Neuron</td>
<td>NPGL3</td>
<td>Neuron</td>
<td>CLAGL3</td>
</tr>
</tbody>
</table>
gene was manually determined from the RefSeq database [38], and we used the sequence 500bp upstream and 100bp downstream of the TSS from the mm10 genome as the putative promoter. Finally, we selected 30 putative astrocyte human enhancers that contained astrocyte specific peaks from human snATAC-seq data.

**Enhancer library cloning**

The library was ordered from Integrated DNA Technologies (IDT) as eBlocks with overhangs required for cloning. An example library member is shown here:

```
CCTGCGGCGCCACGCGTAAC-SpeI-enhancer/promoter-XbaI-AsiSI-EcoRI-cBC-
TCGCTTCGAGTGATGTCAGC-XhoI-PmeI-NheI-cBC-SacI-GTTTAAGAGCTAAGCTGGAA
```

The eBlock fragments were pooled in equal amounts, with the putative enhancers and promoters kept in separate pools because only final constructs with the enhancer will contain a minimal promoter.

We digested the AAV2 pilot library plasmid with SpeI and SacI and assembled either the enhancer or promoter library pools using HiFi DNA Assembly (NEB #E2621). We then digested the respective libraries with XhoI and NheI and dropped in a construct containing the poly(A) signal and U6 promoter with T4 DNA ligase (NEB #M0202). Next, we digested the libraries with XbaI and EcoRI. For the enhancer library, we dropped in a construct containing the minimal β-globin promoter, mEmerald and WPRE with T4 DNA ligase (NEB #M0202). For the promoter library, we dropped in a construct containing just the mEmerald and WPRE. At this stage, we pooled the enhancer and promoter libraries at a ratio of 15:320 (promoter:enhancer libraries) to ensure that all library members are present in equal amounts. Finally, the pooled enhancer/promoter library was digested with XhoI, and a random barcode was dropped in using HiFi DNA Assembly (NEB #E2621).
References


