


Winter 12-15-2018

# Grammar and Variation: Understanding How cis-Regulatory Information is Encoded in Mammalian Genomes

Dana Michele King  
*Washington University in St. Louis*

Follow this and additional works at: [https://openscholarship.wustl.edu/art\\_sci\\_etds](https://openscholarship.wustl.edu/art_sci_etds)

 Part of the [Biodiversity Commons](#), [Biostatistics Commons](#), [Evolution Commons](#), and the [Genetics Commons](#)

---

## Recommended Citation

King, Dana Michele, "Grammar and Variation: Understanding How cis-Regulatory Information is Encoded in Mammalian Genomes" (2018). *Arts & Sciences Electronic Theses and Dissertations*. 1701.  
[https://openscholarship.wustl.edu/art\\_sci\\_etds/1701](https://openscholarship.wustl.edu/art_sci_etds/1701)

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](mailto:digital@wumail.wustl.edu).

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences  
Molecular Genetics and Genomics

Dissertation Examination Committee:

Barak A. Cohen, Chair  
Joseph Dougherty  
Kristen M. Naegle  
Tim Schedl  
Cristina de Guzman Strong

Grammar and Variation: Understanding How *cis*-Regulatory Information is Encoded in  
Mammalian Genomes

by  
Dana M. King

A dissertation presented to  
The Graduate School  
of Washington University in  
partial fulfillment of the  
requirements for the degree  
of Doctor of Philosophy

December 2018  
St. Louis, Missouri

© 2018, Dana King

# **Table of Contents**

<b>List of Figures</b>	iv
<b>List of Tables</b>	v
<b>Acknowledgments</b>	vi
<b>Abstract</b>	x
<b>Chapter 1: Introduction</b>	<b>3</b>
Mechanisms of eukaryotic transcription	4
Transcription factors bind specific but degenerate DNA sequences	5
The problem of specificity	7
Three models for the requirements for TFBSs in regulatory elements	9
Interpreting noncoding variation in context of human disease and evolution	11
Focus of Dissertation	13
<i>Determining the grammar of pluripotency factors in mouse embryonic stem cells</i>	13
<i>Prioritizing non-coding variants in the human genome that may have functional impact</i>	14
<b>Chapter 2: Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in Mouse Embryonic Stem Cells</b>	<b>17</b>
Introduction	18
Results	20
Discussion	30
Methods	32
Data Access	39
Disclosure Declaration	39
Acknowledgements	39
Author Contributions	40
Figures	41
Supplemental Materials	47
<b>Chapter 3: Evaluating fitness predictions for identification of cis-regulatory variants using massively parallel reporter assays</b>	<b>64</b>
Introduction	66

Results	69
Discussion	75
Methods	77
Data Access	82
Disclosure Declaration	82
Acknowledgements	82
Figures	83
<b>Chapter 4: Conclusion and Future Directions</b>	<b>89</b>
Synthetic and Genomic grammar of pluripotency factors	89
Identifying regulatory variants using fitness predictions	91
Evidence for role of repression in mammalian cis-regulation	94
Conclusions	95
<b>References</b>	<b>97</b>

## **List of Figures**

Figure 2.1. Activity of synthetic elements and genomic sequences	39
Figure 2.2. Non-additivity in synthetic elements	40
Figure 2.3. Positional grammar in synthetic elements	41
Figure 2.4. Sequence features separate active and inactive genomic sequences	42
Figure 2.5. Activity of genomic sequences scales with increased occupancy in the genome	43
Figure 2.6. Performance of iRF classification models that include features specific to genomic sequences	44
Figure 3.1. Activity of reference sequences in GM12878	81
Figure 3.2. Activity of reference sequences versus probability of selection	82
Figure 3.3. Allelic skew of 1000 genome variants in GM12878	83
Figure 3.4. Allelic skew versus probability of selection and deleteriousness score	84
Figure 3.5. Allelic skew versus allele frequency	85

## **List of Tables**

Table 2.1: SYN library composition	53
Table 2.2: gWT site composition	53
Table 2.3: gWT/gMUT library composition	53
Table 2.4: Primer sequences	54
Table 2.5: iRF SYN feature matrix	56
Table 2.6: iRF gWT Feature Matrix	57

## Acknowledgments

My graduate work would not have been possible without the guidance and support of the people around me. Growing and developing as a scientist in the Cohen Lab has been an incredible experience. I am grateful to Barak for being a fearless captain as we've navigated the unknown waters that is pursuing research. Although it took us both some time to figure out when we were arguing in disagreement or agreement, his rigor and insights shaped my outlook on science and the world.

I must also acknowledge the many members of the lab who have supported me and challenged me in equal measure. Lab meetings and informal discussions were key to developing my research questions, addressing pitfalls and celebrating the hard earned victories. I thank Mike White for his insights on new ways to look at my data and for generously offering his time to discuss how to work through coding or project challenges. I thank Chris Fiore for laying the foundation of work that allowed me to develop my project on grammar and for leading by example in his coding elegance, as well as Jamie Kwasnieski and Illaria Mogno for their careful development of our MPRAs, and Marc Sherman, Kimberly Lorenz, and Priya Sudarsanam for their scientific feedback and welcoming me into the lab at the very beginning. Hemangi Chaudhari, my baymate and friend for the majority of my time in the lab, for her tirelessness in pursuing her work and for her help troubleshooting my analyses and my experiment. Devjane Swain-Lenz for her evolutionary perspective and her laughter. Max Staller for his advice, enthusiasm and empathy. Brett Maricque for his devil-may-care attitude while being ready to



step in when I needed help. Linda for sharing her extensive experience and excitement. Nicole Rockweller for her outside perspective, depth of knowledge of human genetics, and kindness. Avi for his patience explaining statistical concepts to me and for contributing to the fitness project. Clarice Hong for letting me vent while she pushes the lab down new avenues of research. Siqi Zhao, Ryan Friedman, Jeff Hansen, and Kai Lowell for their energy, insights, and humor. I owe every member a debt for making the Cohen lab an open, equitable, and fearless place to do science.

The Center for Genome Sciences (CGS) has been an incredible community to develop my research. I am grateful to Jess Hoisington-Lopez for single-handedly assuring the quality and speed of my sequencing. Sabrina Wagoner in the Gordon Lab for always answers my questions and helped me navigate the bureaucracy of repairing equipment. Xuhua Chen in the Mitra Lab for helping me solve cell culture crises as well as the many other members of the CGS family that were always willing to chat about my data or science as a whole.

I would like to thank the members of my thesis committee - Tim Schedl, Kristen M. Naegle, Cristina de Guzman Strong, Jim Havranek, and Joseph Dougherty - for their generosity with their time and for the crucial advice and feedback on my projects. I am also grateful for the scientists I worked with prior to coming to Washington University for starting me down of the path of being an independent researcher, Sudhir Nayak at The College of New Jersey and Christopher Hammell at Cold Spring Harbor Laboratories, as well as the amazing science education I received in public school that fed my inquisitive mind and sparked my love of biology.

Through graduate school, I also met the wonderful Kevin McAlister, whom I thank for his love, support, and patience in this process that has taken so many unexpected twists and turns. Kevin's warmth and intelligence has helped me grow personally and professionally, and I am so grateful to have a partner that challenges me and pursues his passions in stride. Kevin's family has also been incredible supportive of my work and I'm grateful to have been so completely welcomed by such a kind and insightful group of people. I am also thankful to those in my family who have championed me from the beginning, especially Sandy, Debbie, and my sisters Anita & Laurel.

My research was funded by NIH grant R01 GM092910.

Dana M. King

*Washington University in St. Louis*

*December 2018*

Dedicated to all the strong women who mentored and supported me, but most of all, you  
'Grandma Dottie'.

## ABSTRACT OF THE DISSERTATION

Grammar and variation, understanding how cis-regulatory information is encoded in mammalian genomes

by

Dana Michele King

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2018

Barak A. Cohen, Chair

Understanding how genotype leads to phenotype is key to understand both the development and dysfunction of complex organisms. In the context of regulating the gene expression patterns that contribute to cell identity and function, the goal of my thesis research is to how changes in genome sequence may impact impact gene expression by determining how sequence features contribute to regulatory potential. To accomplish this goal, I first leveraged the key regulatory role of pluripotency transcription factors (TFs) in mouse embryonic stem cells (mESCs) and tested synthetically generated and genomic identified combinations of binding site for four TFs, OCT4, SOX2, KLF4, and ESRRB. I found that although the position of binding sites explained 87% of the variation in expression observed for synthetic elements, the position of binding sites did not explain the expression of tested genomic sequences despite roughly similar binding site composure. Instead, for genomic sequences I found that the quality and spacing of the binding sites contribute more to distinguishing active sequences, suggesting that the arrangements of binding sites are less important for controlling expression in mESCs.

In a separate set of experiments, I tested regions of the human genome assigned a regulatory function based on chromatin features and predicted to have high to low probabilities of being under selection in a commonly used human immune progenitor cell culture model,

GM12878. Although only a quarter of the library was assigned as ‘Repressive’ according to chromatin marks, 45% of tested sequences showed repressive activity. Sequences predicted to have high probabilities of being under selection have a small but significant higher average level of activation, but not a higher likelihood of either repression or activation. By making single substitutions found at those loci in human populations for a subset of sequences, I tested the predictive power of two independent programs that aim to integrate both functional annotations and evolutionary signals. I found that neither sets of predictions enriched for variants that impacted regulatory activity. This suggests that although we can survey human genotypes for impacts on regulation, it may be difficult to separate organismal level selection from other processes that contribute to the proper control of gene expression.

These results demonstrate that in mESC, the fixed affinity and fixed spacing found in synthetic combinations of binding sites are unlikely to predict the activity of genomic sequences. Furthermore, testing sequences from the human genome in GM12878 shows that repression may be more prevalent than estimated by chromatin features alone and that predictions of selection do not enrich for human variants that impact regulatory activity. Together, these experiments demonstrate that the relationship between genotype and proper regulatory function is complex and that understanding this relationship is important to understand both subtle and severe impacts to phenotype.

# **Chapter 1: Introduction**

Understanding the mechanisms that govern transcriptional regulation is critical to our understanding of biology. Protein coding sequences compose only 2.94% of the human genome and are largely conserved over evolutionary time, while up to a third of the genome is estimated to be involved in a range of regulatory processes (Mathelier, Shi, and Wasserman 2015; ENCODE Project Consortium 2012; Y. Pan et al. 2010; Levine and Tjian 2003; Castillo-Davis 2005; Rubinstein and de Souza 2013). Genome Wide Association Studies (GWAS) studies indicate that the vast majority of disease associated polymorphisms lie within non-coding regions, with some fraction likely impacting *cis*-regulatory sequences (ENCODE Project Consortium 2012; Parker et al. 2013; Tak and Farnham 2015). To understand the evolution of complex multicellular organisms and to untangle the mechanisms of common diseases, it is critical to develop a detailed understanding of how regulatory sequences integrate multiple cellular inputs to generate the correct transcriptional output. Below, I outline both our current understanding of transcriptional regulation and how we can utilize the relationship between sequence and function to better predict *cis*-regulatory impact. My thesis work aims to understand the combinatorial control of expression levels through the rearrangement of binding sites for key regulatory factors in a given cell type (Chapter 2), as well as how single nucleotide substitutions may impact the functional consequence of putative regulatory sequences from the human genome (Chapter 3). In Chapter 2, I hypothesized that the activity of sequences is controlled by a grammar of interactions between regulatory proteins. In Chapter 3, I hypothesized that computational predictions of fitness consequences will help prioritize sequences in the human genome with regulatory activity and help predict the regulatory consequence of variation. Together, my thesis works aims to connect our understanding of how changes to DNA sequences, either on the context of TFBSs arrangements or single nucleotide substitutions change *cis*-regulatory output.

## **Mechanisms of eukaryotic transcription**

Eukaryotic transcription requires the coordination of several layers of regulatory processes. Modifications to histones or DNA itself act to modulate the accessibility of regions of the genome, while factors that directly or indirectly bind DNA sequences stabilize or destabilize the transcriptional machinery at specific loci to regulate mRNA production (Maston, Evans, and Green 2006; Robert G. Roeder 2003; R. G. Roeder 1996; Sandelin et al. 2007). Transcriptional regulation can be defined more specifically as the *trans*-action of proteins known as transcription factors (TFs) bound to *cis*-regulatory sequences (CRSs) that modulate the expression of one or more gene targets (Mathelier, Shi, and Wasserman 2015; Lelli, Slattery, and Mann 2012; Y. Pan et al. 2010; Badis et al. 2009; Maston, Evans, and Green 2006). This definition highlights two critical components of transcriptional regulation: DNA sequences and the TFs that bind them.

In the case of promoters, regulatory sequences function to modulate the assembly of pre-initiation complexes at transcription start sites to fine-tune the rate of expression and therefore transcript abundance. Enhancers are sequences that impact the likelihood of expression but unlike promoters, act through long range interactions. Critical for defining the expression patterns of distal genes across cell-types, enhancers are pointed to as drivers of many developmental processes (Whyte et al. 2013; Parker et al. 2013; Evans, Swanson, and Barolo 2012; Lelli, Slattery, and Mann 2012; Frankel 2012; Nolis et al. 2009). Identified enhancers for most genes are large, from approximately 400 basepairs (bp) to over 10 kilobases in length, with only a handful studied in sufficient detail to determine what particular regions of the full length sequence are necessary for regulatory activity (H. Y. Zhou et al. 2014; Segal et al. 2008; Shen et al. 2012; Piens et al. 2010; Maston, Evans, and Green 2006; Kulkarni and Arnosti 2003). The limited number of validated long range enhancers in both humans and mice makes it difficult to connect differences in DNA features or biochemical features measured genome-wide, such as TF binding or histone marks, to the expected expression output. Furthermore, the few positive examples make it impossible to predict the expected regulatory impact of mutations or sequence rearrangements for enhancers regulated by the same factors. These challenges mean that a

reductive approach, measuring the activity of potential CRSs and making manipulations to determine the impacts on regulatory activity, can make important contributions to our understanding of gene regulation. Reductive approaches has been used successfully to understand key developmental enhancers, by comparing regulatory activity and specificity across evolutionary time, and as well as used to dissect the contribution of regulatory regions to disease (Pennacchio et al. 2006; Evans, Swanson, and Barolo 2012). The ultimate goal of these efforts is to be able to identify functional regulatory regions of the genome and predict the level of activation or repression that will be driven by a given CRS.

### **Transcription factors bind specific but degenerate DNA sequences**

To predict regulatory activity from sequence, we need to first identify DNA sequences that are likely to be bound by a given TF. The strength of TF binding to embedded DNA signals, or transcription factor binding sites (TFBSs), is dependent on the biochemical interaction between protein residues and the DNA structure, which is in turn dependent on the concentration of the factors and the affinity of a factor for a sequence of nucleotide residues (Segal et al. 2008). TFBSs are usually short (5-15 bp) sequences where TFs preferentially bind. A motif represents that preferential binding, representing a collection of positions with one or more possible nucleotides that fall within all TFBS experimentally discovered for a particular TF (D'haeseleer 2006; Bulyk 2003). While restriction enzymes tend to have very precise motifs, ie: EcoRI only binds to a single 6-bp sequence, even the most conserved TF motifs like the TATA box are degenerate (D'haeseleer 2006). Quantification of the binding preferences of a given TF can be measured by in vitro experimental methods, primarily gel-shift assays, systematic evolution of ligands by exponential evolution (SELEX), and protein binding microarrays (PBMs) (Levy and Hannenhalli 2002; Ponomarenko et al. 2002; Kadonaga and Tjian 1986). Quantitative representations of the binding preferences of TFs is an important tool for understanding predicting *cis*-regulation.



One representation of TFBSs, a positional weight matrix (PWM), is generated by aligning sets of sequences produced by experimental methods and comparing the sequence preferences of a TF to the background nucleotide frequency in the genome of interest to generate a statistical representation of binding preference. The PWM for a TF relates the information content of the motif, or how different is the motif compared to the nucleotide frequency in the genome of interest, to the binding energy of the protein-DNA interaction (Robasky and Bulyk 2011). A PWM provides a convenient method for scanning genomes for potential transcription factor binding sites (TFBSs) without relying on a single consensus sequence that fails to capture the full range of binding preferences for a given TF (Stormo and Fields 1998; Robasky and Bulyk 2011). A PWM or the complementary energy weight matrix, can be used to determine the affinity, or probability that a DNA sequence will be bound given a measured or assumed concentration of a TF of interest (Zhao, Granas, and Stormo 2009; White et al. 2013). Although PWMs are powerful in representing large amount of information, a limitation is in the difficulty in representing gaps or accounting for TFs with multiple DNA-binding interfaces due to the assumption that all positions are independent (Ghandi et al. 2014; Weirauch et al. 2013). An alternative method is to examine all possible  $k$ -mers for DNA sequences, which is possible due to the development of high throughput methods such as PBMs that can explicitly test in vitro all possible sequences of length  $k$ . These sequences can then be scored to characterize the full DNA binding specificity of a TF of interest (Weirauch et al. 2013; Alleyne et al. 2008). Either approach results in the ability to determine the likelihood and/or strength of TF binding to a potential CRS, with the goal that an understanding of where TFs are expected to bind will help predict *cis*-regulatory activity.

However, predicting the regulatory activity of a CRS is still difficult even with possible binding sites for relevant TFs for a given cell type in hand. A microarray-based assay of a hundred mouse TFs, selected from various structural classes, found that while each TF bound distinct sets of sequences, only half bound distinguishable DNA motifs (Badis et al. 2009). Interestingly, the authors also found that closely related factors could have similar high-affinity site binding

preferences but different low-affinity site preferences. This indicates that not only do TFs frequently recognize degenerate sequences, but that high-affinity sites could be bound by multiple factors within a family, a finding consistent with the high degree of DNA binding domain similarity between TF family members even over evolutionary time (Badis et al. 2009; Weirauch et al. 2013; Jauch et al. 2012). Since many proteins bind similar sequences, but not all sequences that match consensus TFBSs are bound, a more dynamic picture of how TFs compete and synergize is necessary for a detailed understanding of transcriptional regulation.

Additionally, it is unclear how relevant *in vitro* derived binding motifs are when tested *in vivo*. The presence of different cofactors could possibly skew binding site preferences and would be detectable by using more biologically relevant assays. Data from ChIP-seq experiments suggest that only approximately 10% of the possible binding sites for a TF of interest are occupied in the genome (Cusanovich et al. 2013; Kheradpour et al. 2013; ENCODE Project Consortium 2012; Levy and Hannenhalli 2002; Kadonaga and Tjian 1986). Our lab has also shown, in mouse retina as well as two human cell lines, not only that many genomic regions with TFBS underlying ChIP-seq peaks are unable to drive expression, but also that many active regulatory elements lack a motif for the TF of interest, even if considered occupied in the genome (White et al. 2013; Maricque, Dougherty, and Cohen 2017; Chaudhari and Cohen 2018; White et al. 2016). These results are corroborated by efforts by the ENCODE consortium, with only 55% of reported ChIP-seq peaks enriched for a detectable motif matching the target factor (ENCODE Project Consortium 2012). However, despite the lack of a recognizable motif for the factor of interest, sequence based models perform well to classify active and inactive sequences (Maricque, Dougherty, and Cohen 2017; Chaudhari and Cohen 2018; White et al. 2013). Together these results suggest that while TF binding is an important step, the use of binding information is insufficient to explain patterns of expression and additional sequence features, such as the spacing, arrangement, and quality of TFBSs, likely contribute to the activity of putative regulatory sequences.

## **The problem of specificity**

An explanation for how the specificity of TF binding is determined lies in the physical architecture of the genome. In vitro binding affinity assays for TFs ignore the highly compacted, histone-bound structure of the eukaryotic genome as well as modifications to the local chromatin that could influence binding. Histone positioning must have an effect on the ability of a TF to bind to a particular regulatory site, as bulky histones could easily sequester regulatory sequences or hinder motif recognition and binding (Thurman et al. 2012; Zentner and Scacheri 2012; Natarajan et al. 2012; Y. Pan et al. 2010). Additionally, well-positioned nucleosomes might facilitate some TF binding, with factors such as FOXA1 showing reduced DNA binding in the absence of proximal methylated histones (Lupien et al. 2008). Several studies have examined how chromatin states correlate with the activity of cell-specific regulatory elements as well as how chromatin states change between cell types and how the dynamics of histone modifications influence gene expression (Parker et al. 2013; Ernst et al. 2011; Xi et al. 2007). However, it is difficult to determine in most instances if chromatin changes are instructive or merely consequences of TF binding, with support of the latter growing in recent years (Guertin and Lis 2010, 2013; Dorigi et al. 2017; Wijchers et al. 2016). Causal changes in nucleosome identities or modifications are difficult to demonstrate with current techniques, especially for particular histone marks (Viñuelas et al. 2012; Kouzarides 2007), and histone marks alone have poor predictive power for identifying genomic sequences with regulatory potential (Kwasnieski et al. 2014). Therefore, taking putative CRSs out of their genomic context is a means to focus on how DNA sequence contributes to regulatory activity in a tractable and reproducible manner.

As the size of the average CRS is in the 1-2 kb range, eight to twenty-four TFs could directly be bound to a given sequence, or more if protein-protein interactions are taken into account (Y. Pan et al. 2010). Since a nucleosome core particle covers 147 bp, these regions could also be fully or partially occluded by a well placed nucleosome, reducing the number of TFs able to access a regulatory sequence. The number of possible TF and nucleosome configurations of any given enhancer is therefore very large, making it difficult to determine what sequence features are

necessary to drive expression. We are far from understanding how the presence or absence of a TFBS relates to the activity expected from an enhancer. Fortunately, shorter sequences can represent the possible regulatory potential of genomic sequences, as work from our lab demonstrates that control over expression can be localized within very short genomic regions, 84 bp in length, when isolated placed into an high throughput assay for regulatory activity (White et al. 2013). Even though well characterized enhancers usually require large genomic fragments to recapitulate proper transgene expression, short sequences are a better starting point for understanding patterns of TFBS and other sequence features like nucleosome positioning signals or dinucleotide content (Parker et al. 2013; Evans, Swanson, and Barolo 2012; Frankel 2012; Johnson et al. 2008; Bonifer 2000). Understanding how smaller functional units of regulatory regions control expression could lead to a better understanding of how information is encoded into the DNA sequence of full length enhancers.

### **Three models for the requirements for TFBSs in regulatory elements**

A central question for predicting function from DNA sequence is understanding how grammar, the number, order, spacing, and orientation of TFBSs, relate to *cis*-regulatory activity driven by key TFs (Weingarten-Gabbay and Segal 2014). A possible explanation for the discrepancy between bound and active genomic sequences is that particular combinations of TF binding trigger a particular level of transcriptional activity, a model best described as “enhanceosomes” (Panne 2008; Yie, Senger, and Thanos 1999). Enhanceosomes require combinatorial binding, where specific TFs bind independently in a particular order, or cooperative binding, where physical interactions with other TFs stabilize TF-DNA interactions and are necessary for better binding and modulation of expression, and are considered to follow a strict grammar (Lelli, Slattery, and Mann 2012; Evans, Swanson, and Barolo 2012; Maston, Evans, and Green 2006). The enhanceosome model predicts that any change in the arrangement of TFBS in a CRS will disrupt the activity of the sequence. Two examples of a well-defined enhanceosome are the 55 bp human interferon-B enhancer, where loss of a single TF binding site will abolish IFN-B activity,

and the tumor necrosis factor- $\alpha$  enhancers (Lelli, Slattery, and Mann 2012; Kulkarni and Arnosti 2003).

Alternatively, the “billboard” model, describes a CRS with flexible grammar acting as a general landing pad for TFs and specificity is determined due to a threshold for the number of bound TFs before activation (Lelli, Slattery, and Mann 2012; Evans, Swanson, and Barolo 2012; Kulkarni and Arnosti 2003). The billboard model predicts that most configurations of TFBSs will similarly drive expression. The flexibility of the billboard model is used to explain observations from comparative genetics, where a specific expression pattern can be achieved with many different TFBS configurations, such as in the case of binding site turnover in the *sparkling* and *even-skipped* enhancers from different *Drosophila* species and conservation in enhancer activity between humans and mice despite sequence divergence (Lelli, Slattery, and Mann 2012; Evans, Swanson, and Barolo 2012; Ludwig et al. 2000; Hare et al. 2008; Visel et al. 2009).

A third alternative hypothesis is the TF “collective” model. In the collective model, specific TFs must be recruited to enhancers but can be recruited either by direct contact with DNA or indirectly through other TFs (Spitz and Furlong 2012; Junion et al. 2012; Uhl, Zandvakili, and Gebelein 2016). In the collective model no specific TFBS is required for activity even though individual TFs are required. So the TF collective model predicts that while different TFBS configurations or even identities would be tolerated, all the factors required must be recruited to the locus for activity to be observed. An example of regulatory elements that follow the collective model is in *Drosophila* where the assembly of five transcription factors cooperatively regulates heart-specific enhancer activity (Junion et al. 2012).

These three *cis*-regulatory models differ in the importance that they give to TF-TF interactions and TF-DNA interactions in setting the activity of enhancers and predict different means of generating specificity for transcriptional regulation through DNA sequence. The billboard model emphasizes the additive contribution of each TF bound to a CRS, the enhanceosome model

postulates geometrically constrained interactions between TFs, and the TF collective model highlights the joint action of TFs acting on a CRS without requiring any particular TF-TF or TF-DNA interactions. Measuring the activity of different arrangements of binding sites would be a powerful tool for distinguish between the different models. Determining the possible combinations of TFBSs in functional CRSs and the grammar could help point to possible mechanisms of regulation by a particular suite of TFs and possibly help make predictions for the expected impact of mutations in noncoding sequences more broadly.

### **Interpreting noncoding variation in context of human disease and evolution**

There are instances in which is it impractical to identify the specific TFs or binding sites that act on noncoding sequences of interest. Following the sequencing of the human genome, decades of studies have described and analyzed the genetic basis of human phenotypic variation, ranging from whole-organism diseases or disorders to molecular-level phenotypes, with technological advances making sequencing based genome-wide association studies of human variation more affordable to study complex phenotypes (Stranger, Stahl, and Raj 2011; Mansur et al. 2018; F. Zhang and Lupski 2015; Mardis 2011; 1000 Genomes Project Consortium et al. 2015). However, number of loci and variants associated with different phenotypes without a clear understanding of which variants are causal means that relying solely a reductive approach looking at specific TFBSs or other sequence based features is intractable.

One approach to identify functional variants, including cis-regulatory variants, is to use evolutionary signals. Whole genome comparisons between species show that protein-coding sequences generally vary little between related organisms, with some notable exceptions such as immune function (Mouse Genome Sequencing Consortium et al. 2002; Chimpanzee Sequencing and Analysis Consortium 2005). Because noncoding cis-regulatory regions encode critical spatio-temporal and quantitative information, these sequences are thought to be under strong purifying selection and indeed mutate at a slower rate than flanking neutrally evolving regions (Rubinstein and de Souza 2013). Although *cis*-regulatory regions can be evolutionarily

conserved, they generally evolve faster than coding regions, suggesting that changes in regulation may contribute to fitness and evolution (S. L. Clarke et al. 2012; Konopka et al. 2009; John Wiley & Sons, Ltd 2001). Additionally, Pennacchio and colleagues showed that 45% of tested conserved human noncoding sequences showed tissue specific enhancer activity in mouse embryos (Pennacchio et al. 2006). Therefore measures of conservation, while neither exhaustive or perfect predictors for regulatory activity, are often used to prioritize variants that may be contributing to differences in gene regulation or other traits (Kheradpour et al. 2013; 1000 Genomes Project Consortium et al. 2012; Asthana et al. 2007).

More detailed work has identified likely causal variants that impact regulatory functions for genes related to the relevant phenotype. At a locus associated with type two diabetes, a single nucleotide substitution was shown to modify the TF binding of two key regulators, PAX6 and PAX4 and increase the expression of ARAP1 in pancreatic beta cells, linking binding and expression to a possible risk marker for disease (Kulzer et al. 2014). For pulmonary diseases, association studies combined with reporter assays identified three variants that impact the transcriptional response to oxygen deprivation, with one variant shown to causally impact EGFR expression through genome editing, linking a possible molecular mechanism to variability in disease susceptibility and severity (Roche et al. 2016). In addition, over 600 common variants associated with pulmonary disease were screened with high-throughput reporter assays, resulting in the identification of several variants that impact regulatory activity, including one that was shown using genome editing to affect cell proliferation (Castaldi et al. 2018). These efforts have provided a window into how *cis*-regulatory changes may lead to disease phenotypes. However, these studies have been single loci due to the large number of variants that are associated with a locus due to spurious factors like linkage disequilibrium or the lack of sufficient locus coverage in phenotyping arrays (Majumder and Ghosh 2005; Roukos 2009; Mansur et al. 2018). Broadly identifying possible regulatory variants for a given cell-type or parsing down the relevant regions of the genome, through either conservation or functional annotations will aid in the discovery process necessary to untangle the genetics basis of phenotypic variation.

## **Focus of Dissertation**

My thesis work utilizes technological advances to examine how TFBS grammar contributes to regulatory activity in mESCs and how single substitutions may impact the functional consequence of CRSs in the human genome. Our lab has pioneered the use of massively parallel reporter assays (MPRAs), a high-throughput assay that links short putative CRSs directly to the expression of a minimal promoter and reporter gene in a tractable manner by linking each CRS with unique barcode sequences in the 3'UTR (Mogno, Kwasnieski, and Cohen 2013; Kwasnieski et al. 2012). This technology allows for the construction of large plasmid based libraries on the order of thousands of unique CRSs, which can then be measured by next-generation sequencing to compare the number of RNAs generated by each construct compared to its representation in the DNA plasmid pool, resulting in highly reproducible, quantitative measures of regulatory activity (Fiore and Cohen 2016; Kwasnieski et al. 2014; Chaudhari and Cohen 2018). The use of MPRAs provides the statistical power and sensitivity necessary to determine what DNA sequences features or other annotations are common to active versus inactive regulatory elements (White et al. 2013; Kheradpour et al. 2013). While a plasmid-based assay is unlikely to recapitulate the possible looping mechanism of an enhancer, placing a regulatory element adjacent to a promoter should capture most of the complexity of TF- TF and TF-DNA interactions that control changes in expression attributable to a particular element since looping appears to be a means of bringing regulatory factors in close proximity (Noonan and McCallion 2010; Nolis et al. 2009). Additionally, testing variants of the same sequence of interest allows us to directly measure the possible impact of a given substitution, instead of being limited to correlations with the expression of nearby genes (Tewhey et al. 2016; Ulirsch et al. 2016; Mohammadi et al. 2017).

### ***Determining the grammar of pluripotency factors in mouse embryonic stem cells***

In Chapter 2, I examine the *cis*-regulatory grammar of pluripotency TFBSs in mouse embryonic stem cells (mESCs) by testing hundreds of synthetic and genome putative CRS. The guiding



hypothesis of this project is that most of the information necessary to guide individual TFs to the correct targets and activate regulatory elements through combinations of TF binding is contained in the DNA sequence and that a quantitative analysis of expression driven by regulatory elements will help uncover how cell-type expression is encoded into the genome. Previous work in the lab used synthetic combinations of TFBSs for the core pluripotency factors POU5F1 (OCT4), SOX2, ESRRB, and KLF4 in mESCs (Fiore and Cohen 2016). These pluripotency factors contribute to the maintenance of pluripotency in mESC and are sufficient to induce pluripotency in terminally differentiated cells (Masui et al. 2005; Feng et al. 2009; X. Zhang et al. 2008; Liu et al. 2008).

Based on known physical and genetic interactions, multiple interacting TFs specify target gene expression in mESCs (Reményi et al. 2003; Williams, Cai, and Clore 2004; Reményi, Schöler, and Wilmanns 2004; Huang et al. 2009), so only specific patterns of TFBSs that facilitate the necessary interactions would be expected to drive regulatory activity. However, the MPRA library used in Fiore and Cohen was underpowered to distinguish between the three primary models of regulatory grammar summarized above and did not test sequences from the genome for regulatory activity. Therefore, it is unclear what role, if any, is played by TFBS grammar in determining target specificity in the genome or driving specific patterns of gene expression in mESCs. Understanding the grammar model followed by these factors is central to identifying sequences that are part of the regulatory network in mESCs to better understand the establishment and maintenance of the pluripotent state.

### ***Prioritizing non-coding variants in the human genome that may have functional impact***

In Chapter 3, I evaluate the utility of two computational predictions of fitness for prioritizing *cis*-regulatory variants by testing almost two thousand putative CRSs from the human genome and an additional several hundred reference-alternative variants pairs that span the range of predictions for non-coding regions. The guiding hypothesis for this project is that selection acts on functional noncoding regions, so predictions of fitness will help identify functional CRS in

the human genome as well as help predict the consequence of variation. Although benchmarks exist for evaluating algorithms for coding sequences, such as predicting protein truncations, comparing to known pathogenic examples, and evaluating sequences for *in vivo* function or phenotype (Ng and Henikoff 2003; Miosge et al. 2015; Walters-Sen et al. 2015), similar benchmarks for functional roles of non-coding sequences have not been standardized.

One way of parsing down the genome into the regions that are necessary for at least some function is to focus on the areas of the genome that are under selection and therefore contribute to some function that impacts fitness (Pollard et al. 2006; Ludwig et al. 2000; Castillo-Davis 2005). Several groups have published computational approaches to predict the expected impact of human genetic variation (Lindblad-Toh et al. 2011; Spielman and Kosakovsky Pond 2018). Two such predictions, fitCons and CADD scores, integrate diverse annotations into a single measure (Gulko et al. 2015; Kircher et al. 2014), characterizing the fitness consequence and deleteriousness of variants, respectively, with a shared goal of characterizing the functional consequences of substitutions genome-wide. The authors of fitCons used functional data from the ENCODE project in several cell types to group sequences and assigned scores to grouped regions based on divergence and polymorphism frequencies (Gulko et al. 2015). The authors of CADD contrasted observed variants that are fixed or nearly fixed in human populations to simulated substitutions used as proxies for rare mutations that are more likely to impact organismal fitness to assign scores of likely deleteriousness (Kircher et al. 2014). Understanding how fitCons and CADD perform for identifying regions with regulatory activity and/or variants that impact *cis*-regulatory activity is key to determining if predictions of fitness are viable options for the prioritization of possible causal disease variants for follow-up study.

This overall body of work contributes to the field of transcriptional regulation by examining fundamental questions regarding *cis*-regulatory grammar and *cis*-regulatory variants. First, this study provides the first direct comparison between synthetic and genomic putative regulatory sequences in mESCs. This comparison informs future design of synthetic MPRA libraries as one

of the goals of the field is to identify sequences in the genome that are likely to drive *cis*-regulatory activity and adds to our knowledge of how patterns of TFBSs regulate expression. Second, the detailed testing of fitness predictions adds to the efforts of many for interpreting variation in the human genome. Specifically, it is important to determine the value of computational approaches before broader applications such as for identifying possible diagnostic markers or for applications like patient stratification for the growing field of precision medicine.

## **Chapter 2: Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in Mouse Embryonic Stem Cells**

In embryonic stem cells (ESCs), a core transcription factor (TFs) network establishes the gene expression program necessary for pluripotency. To understand 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 sites configurations revealed regulatory grammar requirements. While binding site quality is important for activity in both contexts, the expression of synthetic elements is driven by both binding site number and a grammar that includes position effect. This grammar plays a small role for genomic sequences, as their relative activity is best explained by the predicted affinity of binding sites, regardless of identity, and spacing between sites. Our findings highlight the need for detailed examinations of complex sequence space to understand *cis*-regulatory grammar in the genome.

This chapter was written as a paper, *Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in Mouse Embryonic Stem Cells*, with Brett Maricque and Barak Cohen that is currently under review at *eLife* and is available as a preprint on the *bioRxiv* server (doi: <https://doi.org/10.1101/398107>). I am the first author and contributed to the writing of the paper along with Barak Cohen. Brett Maricque and I performed the experiments and analyzed the initial sequencing data. I performed all additional analyses.

## Introduction

Combinations of transcription factors (TFs) act at enhancers to specify cell states. Three models describe how TFs collaborate at enhancers, the billboard model, the enhanceosome model, and the TF collective model (Kulkarni and Arnosti 2003; Spitz and Furlong 2012). These models differ in the importance they ascribe to *cis*-regulatory *grammar*, defined as the extent to which the order, orientation, and affinity of transcription factor binding sites (TFBSs) impact the activity of enhancers. The enhanceosome model posits a strict grammar in which only precise arrangements of TFBSs activate target genes. The enhanceosome model is supported by structural studies of the IFN- $\beta$  enhancer, where a specific order and spacing of TFBSs is required to activate expression (Yie, Senger, and Thanos 1999; Panne 2008). In contrast, the billboard model posits a more flexible grammar. In the billboard model enhancers can tolerate changes to the order, spacing, or orientations of TFBS with little change to target gene expression (Kulkarni and Arnosti 2003; Giorgetti et al. 2010) This model was proposed to explain binding site turnover in developmental enhancers and conservation in enhancer activity between species despite sequence divergence (Ludwig et al. 2000; Hare et al. 2008; Hare, Peterson, and Eisen 2008; Visel et al. 2009). In the TF collective model, specific TFs must be recruited to enhancers but can be recruited either by direct contact with DNA or indirectly through other TFs (Spitz and Furlong 2012; Junion et al. 2012; Uhl, Zandvakili, and Gebelein 2016). In the collective model no specific TFBS is required for activity even though individual TFs might be. These three *cis*-regulatory models differ in the importance that they give to TF-TF interactions and TF-DNA interactions in setting the activity of enhancers. The billboard model emphasizes the additive contribution of each TF bound to an enhancer, the enhanceosome model postulates geometrically constrained interactions between specific TFs, and the TF collective model highlights the joint action of TFs without requiring any particular TF-TF or TF-DNA interactions.

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 (Fiore and Cohen 2016; Dunn et al. 2014; Williams, Cai, and Clore 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 (Niwa 2014; Feng et al. 2009; Liu et al. 2008; X. Zhang et al. 2008; Takahashi and Yamanaka 2006). 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 (Niwa 2014; Huang et al. 2009; Reményi, Schöler, and Wilmanns 2004; Williams, Cai, and Clore 2004; Reményi et al. 2003). 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 target specificity in the genome (Chambers and Tomlinson 2009; X. Chen, Vega, and Ng 2008). 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 TFBSs by assaying an exhaustive set of synthetic *cis*-regulatory elements, composed of TFBS for SOX2, OCT4, KLF4 and ESRRB. The pattern of expression of synthetic regulatory elements is well predicted by a grammar that incorporates aspects of both the billboard and the enhanceosome models. However, this grammar seems to only play a small role in setting the activity of genomic regulatory elements with comparable configurations of binding sites. Genomic sequences appear to conform to a model where predicted TFBS affinity and favorable spacing between TFBSs contribute to activity levels, along with signals that recruit additional TFs, either directly through TF-DNA interactions or possibly indirectly through TF-TF interactions.

## 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), is an exhaustive set of synthetic combinations of consensus TFBSs for OCT4 (O), SOX2 (S), KLF4 (K), and ESRRB (E). We did not include sites for NANOG in our libraries as its PWM has low information content and is not amenable to a synthetic binding site approach, in addition to being dispensable for reprogramming terminal cells to a pluripotent state (Jie Wang et al. 2013, 2012; Jauch et al. 2008; G. Pan and Thomson 2007; Takahashi and Yamanaka 2006). We also did not incorporate MYC binding sites in our libraries because MYC often acts independently of the core pluripotency TFs (C.-Y. Chen, Morris, and Mitchell 2012; Xi Chen et al. 2008; Liu et al. 2008). For each TF we used a binding site based on its Position Weight Matrix (PWM) in the JASPAR database (Sandelin et al. 2004; Fiore and Cohen 2016). We embedded each TFBS in a constant 20 bp sequence to ensure all the sites sit on the same side of the DNA helix (Fiore and Cohen 2016). We designed the SYN library to include all possible strings of two, three, and four TFBS building blocks (2-mers, 3-mers, and 4-mers, respectively), with each TFBS in either the forward or reverse direction, and each TFBS occurring no more than once per sequence, totaling 624 unique synthetic elements (Supplemental Table 1). The highly controlled nature of the SYN library provides maximum power to detect interactions mediated by the arrangement of binding sites.

The second library includes sequences from the mouse genome picked to 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 TFBSs for O, S, K, & 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; Xi Chen et al. 2008). 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 TFBSs that could be compared to the exhaustive set of synthetic 3-mer elements. The resulting genomic wild-type library (gWT) is composed of the 407 unique genomic sequences with combinations of any three of the four TFBSs, with each site represented no more than once per sequence (Methods, Supplemental Table 3). Although these sequences differ from SYN elements in the individual site affinities, distances between TFBSs, as well as intervening sequence composition, our expectation was that the gWT sequences would directly 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 TFBSs, we generated matched genomic mutant sequences (gMUT) in which all three of the identified pluripotency TFBSs were mutated by changing two positions in each TFBS from the highest information content base to the lowest information base according to the PWM (Supplemental Figure S7). The final gMUT sequences lack detectable TFBSs 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) (Kheradpour et al. 2013; Mogno, Kwasnieski, and Cohen 2013; White et al. 2013; Patwardhan et al. 2012; 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; Mogno, Kwasnieski, and Cohen 2013; Kwasnieski et al. 2012). To determine the relative activation of each sequence compared to the minimal promoter included in each reporter construct, we included copies of plasmids with only the minimal promoter paired with over a hundred unique BCs in each library



(See Methods) (White 2015; Fiore and Cohen 2016; Mogno, Kwasnieski, and Cohen 2013; White et al. 2013; Kwasnieski et al. 2012). Our measurements were highly reproducible between biological replicates, with  $R^2$  between 0.98-0.99 for replicates of the SYN library and 0.96-0.98 for the gWT/gMUT library (Figure S1A-B). 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**

Synthetic regulatory elements have some characteristics of the billboard model. Most synthetic elements drive expression over basal activity regardless of the number, order, or orientation of sites within the element (Figure 1A). 77% of all SYN elements (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 Supplemental File 3). In most cases, three or four consensus binding sites at fixed spacing are sufficient to increase expression above basal levels, which suggests strong independent contributions by binding sites in synthetic elements, as reported previously (Fiore and Cohen 2016). Elements with more binding sites generally drive higher expression than elements with fewer binding sites, which supports a billboard model of regulation, as the TFBSs can contribute to expression in an independent, additive manner. However, the wide range of expression observed for 4-mer elements (2.5-12.5, normalized expression) rules out a pure billboard model as differences between these elements must be due to the arrangement of the TFBSs, as site number and identity are fixed. The strong positive effect of adding sites supports a billboard model, while the diversity of expression of elements with the same number of sites reveals that grammar can quantitatively modulate activity.

In contrast to the synthetic elements, most genomic sequences do 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 Supplemental File 4). The low fraction of active gWT sequences demonstrates that three binding sites for the pluripotency TFs are often insufficient to increase expression above basal levels, a result that differs from the additive, billboard-like behavior observed for the SYN elements, but is consistent with observations from functional testing of genomic sequences bound by key TFs in other cell types (Fisher et al. 2012; Grossman et al. 2017; White et al. 2013). 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 Supplemental File 4), indicating that the activity of these sequences depends on one or more of the pluripotency TFBSs. Our observation that the presence of high-quality pluripotency TFBSs is generally insufficient to drive expression above basal levels does not support a strict billboard model for genomic sequences, at least when tested in a functional assay.

### **Synthetic elements support a positional grammar**

While the overall pattern of expression of SYN elements supports the billboard model, direct comparisons of different TFBS configurations also support a role for interactions between factors. Pairwise comparisons between a particular 3-mer and 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 S2A and S2B). 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 position specific interactions between TFBSs. The ten highest expressing elements in the SYN library have a strong bias for S and O sites to be next to each other and in the first two positions (Figure 2B), while the ten lowest expressing 4-mers all have O and S in the last two positions (Figure 2C). The ten 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. We also found that the fourth position can have an especially large effect on expression. In the highest 25% of 4-mers (n=96) 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, E 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 (Supplemental Figures S2C-F). These patterns indicate a grammar that includes a positional bias to have adjacent SOX2 and OCT4 sites positioned upstream of KLF4 and ESRRB sites, which 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. The data suggest that synthetic elements have some properties consistent with the enhanceosome model, such as positional bias, with possible contributions by specific interactions between factors.

### **Modeling supports 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 likely has an effect on the relative activities of the SYN elements, the grammar of these sites does not appear to contribute to the activity genomic sequences. We compared the SYN and gWT libraries for elements with configurations of OKE, OSE, OSK, and SKE TFBSs. Unlike SYN 3-mer elements, all four classes of gWT sequences span the range of expression observed for the entire library (0.56 to 33.2, normalized expression), with only OSK sequences having a higher average expression (Figure S3A). 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 TFBSs. We observed no correlation in regulatory activity between matched site configurations, ( $R^2 = 0.001$ ; Figure S3B). 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 TFBSs 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 prediction (Basu et al. 2018).

We first trained a regression-case iRF model initialized with four features (Supplemental Table 4), representing only the presence or absence of each of the four pluripotency TFBSs to predict the expression of SYN elements. This “billboard” iRF model performed similarly to billboard-like thermodynamic based models trained on synthetic elements in mESCs (Fiore and Cohen 2016), with a  $R^2$  of 0.56 on a held-out test set for the final iRF iteration (2-fold cross-validation; Supplemental Figure S4). However, the billboard iRF model cannot account for the differences in activities between 4-mers, because all 4-mers have identical TFBS present (4-mers  $R^2 = 0.00$  (blue); Figure S4). To identify features that might distinguish between the activities of 4-mers, we trained an additional regression-case iRF model initialized with 20

features, representing both the presence and position of the four TFBS in each SYN element (Supplemental Table 4). 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 (4-mers  $R^2 = 0.52$  (blue); Figure 3A). The positional iRF model highly weighs the presence/absence of the sites, as expected from the performance of the billboard iRF model, but also has contributions from the presence of ESRRB in the 4th position and SOX2 in the 1st and 2nd positions (Figure 3B). These results reinforce the conclusion that the activity of synthetic sequences depends both on the composition and positioning of TFBS.

Models trained on the SYN library failed to predict or classify the expression of genomic sequences. While synthetic elements drove continuous expression across 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 (dark green; Figure S3A). This presents a challenge for predicting gWT sequence activity with the features used to predict SYN elements. Retraining iRF regression models to predict gWT expression fails during the training step and has no correlation with the observed expression data (Billboard:  $R^2 = 0.03$ ; Billboard + Position:  $R^2 = 0.001$ ). However, training a classification model to distinguish between active and inactive gWT sequences (top 25%,  $n = 102$ ; bottom 75%,  $n = 305$ ) using either only billboard or billboard plus positional features also fails to perform better than chance (Billboard: AUROC = 0.52, AUPRC = 0.22; Positional: AUROC = 0.47, AUPRC = 0.25; Supplemental Table 5). When we directly compare the expression of gWT sequences to SYN elements with matching patterns of sites, we observe that not only do genomic sequences have a very different expression distribution from SYN elements, gWT sequences with the same pattern of TFBSs, as determined by motif matching, can drive drastically different expression levels (Supplemental Figure S3B). Other sequence features present in the genomic sequences and absent from the synthetic elements must therefore play a larger role in setting activity levels than the identity and position of the individual pluripotency TFBSs.

### Site affinity contributes to the activity of genomic sequences

We sought to identify features that differentiate active and inactive gWT sequences.

Sequence-based support vector machines (*k*mer-SVMs) are powerful tools to predict the activity of putative regulatory elements independent of motif calling for specific factors that might be acting on the sequences (Fletez-Brant et al. 2013; Chaudhari and Cohen 2018). To identify sequence features that explain the differences between genomic elements, we trained a gapped *k*mer SVM (gkm-SVM) (Ghandi et al. 2016, 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 TFBSs for the four pluripotency factors, many of the discriminative 8-mers (29/50) have possible motif matches that include at least one pluripotency family member (Fletez-Brant et al. 2013; Bailey et al. 2009) (Supplemental File 6). This suggests that the differences between high and low activity genomic sites could be due to properties of either the primary pluripotency sites, defined here as TFBSs originally identified in the gWT library design, or due to secondary pluripotency sites, defined here as low PWM matches that scored below the scanning threshold, but could be present in the intervening sequences.

Sequences with higher predicted affinity pluripotency TFBSs 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, Bailey, and Noble 2011). For SOX2, we found no difference in scores between high and low sequences (Figure 4B, panel 2;  $p = 0.07$ , 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 4B; KLF4:  $p = 0.001$ , ESRRB:  $p = 0.006$ , Welch's t-test). The 'OSKE\_TotalAffinity', the summed PWM scores for all four TFBS in each sequence, further separates high and low sequences ( $p = 4e-8$ , Welch's t-test). 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 (See Methods). Predicted occupancy is a metric that includes contributions from any primary, well-scoring TFBSs plus contributions from weaker sites that might be missed with traditional motif scanning (White et al. 2016, Guertin, Michael J., and John T. Lis 2013; Evans, Swanson, and Barolo 2012; Segal et al. 2008; Zhao, Granas, and Stormo 2009). We found evidence for additional low predicted affinity sites for SOX2 and OCT4 in both high and low sequences, making it unlikely that low affinity sites strongly contribute to expression differences (Figure S5). 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.

### **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 (Supplemental File 6). We then used PWMs for these additional TFs to identify instances of sites for other factors in the genomic sequences (See Methods) (Grant, Bailey, and Noble 2011; Sandelin et al. 2004). We found significant enrichment for FOXA1 sites (Figure 4C; 1-sided Fisher's exact test; 29 vs. 17 sites,  $p = 0.03$ , OR = 1.97). We also found that FOXA1 and NANOG had higher total PWM scores in the high activity sequences (Figure S6; 1-sided Welch's t-test; FOXA1:  $p = 0.04$ ; NANOG:  $p = 0.03$ ). 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 (G. Pan and Thomson 2007; Mulas et al. 2018; Gabut et al. 2011), and therefore could be acting on the gWT sequences through these binding sites.

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 differences in accessibility (Supplemental Table 4). We found that accessibility was generally high for both high and low activity gWT sequences, and therefore not enriched. High activity sequences were significantly enriched for overlap with NANOG peaks (Supplemental Figure S7; 1-sided Fisher's exact test, NANOG:  $p = 0.03$ , OR = 2.17). 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).

To understand the relative contributions of the features that had some predictive power on their own, we trained iRF models with subsets of these features and compared the performance of these models on a held-out test set (Supplemental Table 4). Including parameters only related to spacing between TFBSs, which is held constant in SYN elements but variable in gWT sequences, resulted in a model with an AUROC of 0.52 (Figure 6A) and AUPRC of 0.31 (Figure 6B) (model 'Spacing'). A model that includes parameters describing the attributes of the primary pluripotency sites, such as the scores for the sites, yields an AUROC of 0.64 and AUPRC of 0.34 (model 'PrimarySites'). We then limited our model to only parameters that



relate occupancy of these regions by TFs in the genome, as measured by ChIP-seq, which resulted in a model with an AUROC of 0.59 and AUPRC of 0.31 (model ‘ChIPSignals’). We trained a final iRF model initialized with 58 features that capture key differences between gWT sequences and SYN elements. These features include predicted affinity and spacing between the pluripotency TFBSs, predicted occupancy for the pluripotency TFs, 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 (as in Figure 4B) for each sequence (See Supplemental Table 6 for full list of features). This gWT iRF model performed fairly well on a held out test set (AUROC = 0.67, AUPRC = 0.46; model ‘All’). The features that best separate active 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 suggests that genomic sequences are able to drive higher expression when they contain stronger predicted affinity binding sites for pluripotency TFs with optimal 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 TFBSs for OCT4, SOX2, KLF4, AND ESRRB and comparable genomic sequences. Our experimental design allowed for direct comparisons between the regulatory potential and regulatory grammar of synthetic elements and genomic sequences. Using a massively parallel reporter assay (MPRA), we found that the regulatory potential synthetic elements and genomic sequences are both impacted by the quality of the TFBS. The majority of synthetic elements, with consensus binding sites, drive expression above basal promoter levels, while genomic sequences with high scoring sites, ie: closer to the consensus sequence, are more likely to be active. The relationship between activity and TFBS scores is consistent with recent observations, specifically for the

recruitment of ESRRB to areas enriched for OCT4, SOX2, and NANOG motifs (Adachi et al. 2018).

As for the grammar of combinations of TFBSs for these four pluripotency factors, in the controlled context of synthetic elements, we found clear evidence for a dependency on the position and number of the binding sites. While each site has a strong additive contribution to expression, consistent with the billboard model, we also uncovered evidence for some position-specific effects, best aligned with the enhanceosome model. The spacing and affinities of sites in synthetic elements were fixed, and the constant spacer sequences between sites reduced the likelihood of the presence of additional high quality TFBSs for other factors, constraints that are not representative of combinations of these sites in the genome. Indeed, when variable predicted affinity and distances between TFBSs is allowed, even with similar patterns of binding sites and evidence for activity dependent on one or more of those sites, the grammar learned from synthetic elements appears to have less of an impact on driving relative activity for genomic sequences. We found that many different arrangements of pluripotency sites can result in a high activity for genomic sequences, a prediction of the billboard model. However, because most genomic sequences were inactive, additional features, excluding identity and position, are also required to produce a sequence that can drive high activity. Surprisingly, active genomic sequences do not appear to have TFBSs for one or few additional factors that might point to a simple billboard model explanation for the patterns of expression observed. Instead, our modeling suggests that outside of the quality of the binding sites present, favorable distances between TFBSs and the ability to recruit TFs in their genomic context might contribute to the levels of expression observed for genomic sequences.

The literature demonstrates the utility of using synthetic elements to probe *cis*-regulatory grammar, understand interactions between factors, and predict true targets in the genome (Gertz, Siggia, and Cohen 2009; Cox, Surette, and Elowitz 2007). This work highlights the need for designing more nuanced synthetic elements, even when examining well-studied mammalian TFs,

if we hope to predict the expected activity of patterns of TFBSs we observe in the genome. Additionally, we speculate that repression by some members of the pluripotency network might account for some of the seeming disparate behavior observed for similar combinations of OCT4, SOX2, KLF4 and ESRRB TFBSs from the genome. In particular, OCT4 has evidence of interactions with Polycomb and NuRD complex subunits in addition to physical interactions with Nanog (Jianlong Wang et al. 2006; Liang et al. 2008). The recruitment of repressive complexes and subsequent active repression would be difficult to distinguish from inactivity in current MPRA approaches. In the future, synthetic libraries of equally complex combinations of TFBS but where the predicted affinity and spacing are varied may better capture trends that allow us to understand the quantitative impact of these features in the genome and innovations in MPRA design to specifically detect active repression may allow for a more comprehensive understanding of the *cis*-regulatory landscape.

## Methods

### 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:

```
CTTCTACTACTAGGGCCCA[SEQ]AAGCTT[FILL]GAATTCTCTAGAC[BC]TGAGCTCTA  
CATGCTAGTTCATG
```

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. Synthetic elements were generated using a custom python script, generating 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 (Supplemental Table 1). The sequence of each of the element is listed in Supplemental File 1. 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:

```
GACTTACATTAGGGCCCGT[SEQ]AAGCTT[FILL]GAATTCTCTAGAC[BC]TGAGCTCG  
GACTACGATACTG
```

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. 2008). After excluding poorly sequenced and repetitive regions (ENCODE Project Consortium 2012; Mouse Genome Sequencing Consortium 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, Bailey, and Noble 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 20bp 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 TFBSs (Supplemental Table 2).

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 (Supplemental Figure S8). 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 TFBSs 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 of the element is listed in Supplemental File 2. 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 any upstream element, each with a unique BC (Supplemental Table 3). The rest of the array contained sequences not used in this study.

### **Cloning of plasmid libraries:**

The synthesized oligos were prepared as previously described (Kwasnieski 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 (Supplemental Table 6). 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 a SYN element (SYN library) or either a gWT or gMUT sequence (gWT/gMUT library) cloned into the *Apal* 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 *SpeI* 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.

**Cell culture and transfection:**

RW4 mESCs were cultured as described previously (Xian, Werth, and Gottlieb 2005; C. T. L. Chen, Gottlieb, and Cohen 2008) 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 uM B-mercaptoethanol). Approximately 1 millions cells at 100% estimated viability were seeded into 6-well plates 24 hours prior to transfection. The SYN library and combined gWT/gMUT were transfected in parallel using 10 uL Lipofectamine 2000 (Life Technologies, Carlsbad, CA), 3 ug of plasmid library, and 0.3 ug 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 hours post-transfection. For three replicates of each library pool, RNA was extracted 24 hours 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 CRE as described previously (Fiore and Cohen 2016; Mogno, Kwasnieski, and Cohen 2013). Briefly, we used Illumina NextSeq™ (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 (Supplemental Table 6), for 13 cycles. The PCR amplification products were digested using XbaI and XhoI (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' XhoI overhang, each of which is comprised of annealed forward (F) and reverse (R) strands (Supplemental Table 6). An enrichment PCR with primers CF52 and CF53 was then used (Supplemental Table 6), 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 3 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 ( $R^2 = 0.982-0.987$ ; Figure S1A) and the combined gWT/gMUT library ( $R^2 = 0.96-0.98$ ; Figure S1B). 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 ( $R^2 = 0.02-0.02$ ; Figure S1C) and the combined gWT/gMUT ( $R^2 = 0.0005-0.008$ ; Figure S1D). 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 Supplemental File 3 for the SYN library and Supplemental File 4 for the gWT/gMUT library.

### **Predicted Occupancy:**

Custom code, based on Zhao & Stormo's BEEML algorithm (Zhao, Granas, and Stormo 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,  $(1 / (1 + e^{(\Delta G - \mu)}))$ . 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

$RT * \ln(\text{Freq}(\hat{\text{Base consensus}})/\text{Freq}(\hat{\text{Base}} i))$  to convert the frequency of each base at each position  $i$  to a pseudo  $\Delta\Delta G$  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  $P(\text{Occ}) \cong 3$  for each SYN element, i.e.: approaching 1 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, Esrrb, 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 Supplemental File 8 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) from (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 ( $w^k$ ), 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 Supplemental Tables 5 & 6.

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 Supplemental File 7 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 Supplemental File 8 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 (Supplemental File 11). 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), & 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 (See Figure

S3). The best average performance on training data as evaluated by AUCs was the model trained with parameters of  $L=8$  and  $gap=2$  (See Supplemental File 12 for output scores). The final *gkmer-SVM* model includes approximately 1 million unique *k*-mers (See Supplemental File 5 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, Haussler, and Kent 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 Supplemental File 9 for the SYN library and Supplemental File 10 for the *gWT/gMUT* library.

### **Disclosure Declaration**

The authors declare that they have no competing interests to disclose.

### **Acknowledgements**

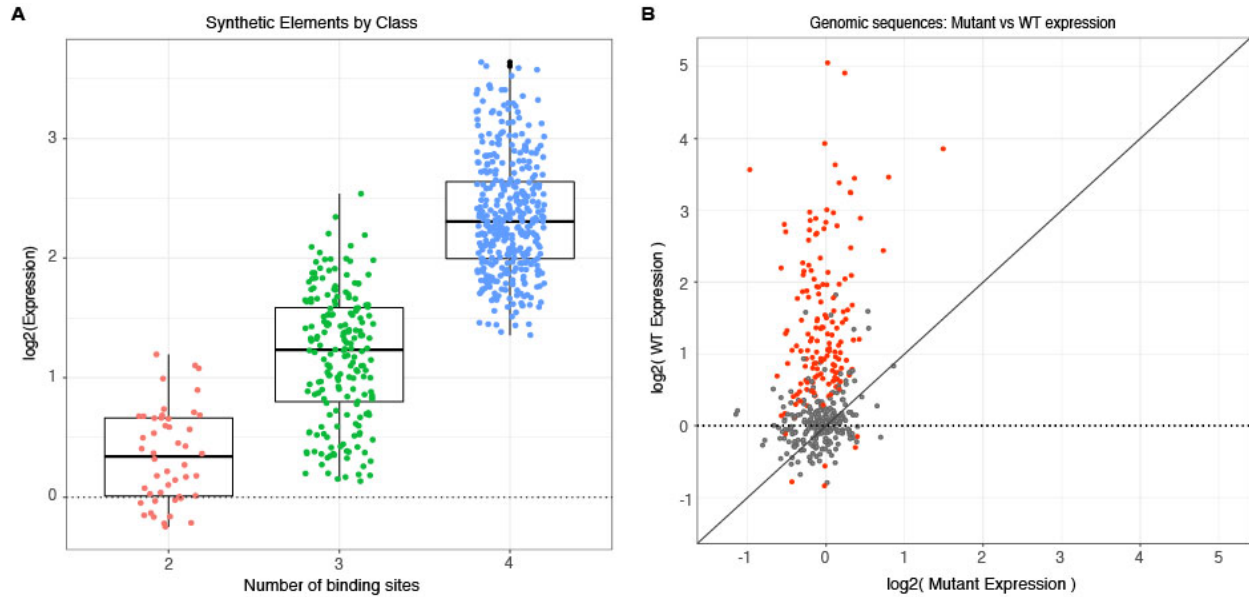
We thank members of the Cohen Lab for critical reading and feedback, particularly Michael White, Max Staller and Hemangi Chaudhari for helpful discussion over the course of the project,

and Jessica Hoisington-Lopez from the DNA Sequencing Innovation Lab for assistance with high-throughput sequencing. This work is supported by a grant from the National Institutes of Health, R01 GM092910 to B.A.C.

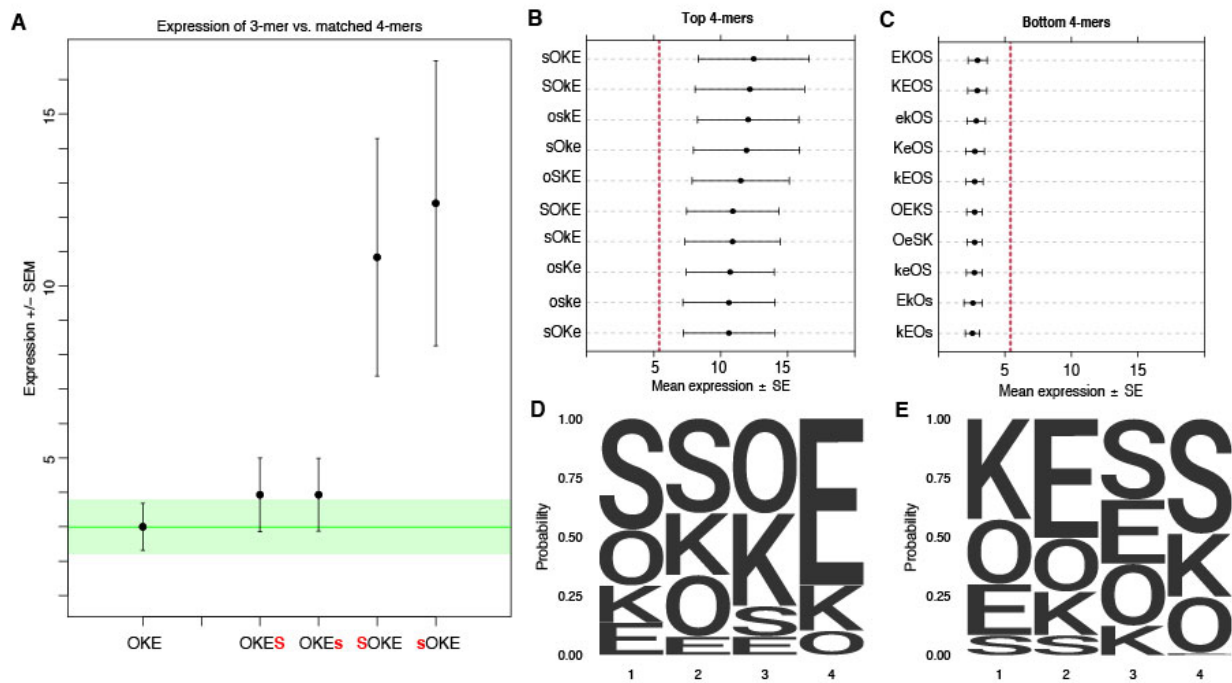
### **Author Contributions**

D.M.K and B.A.C. designed the experiments. D.M.K and B.M. collected the data. D.M.K analyzed the data. D.M.K and B.A.C wrote the manuscript.

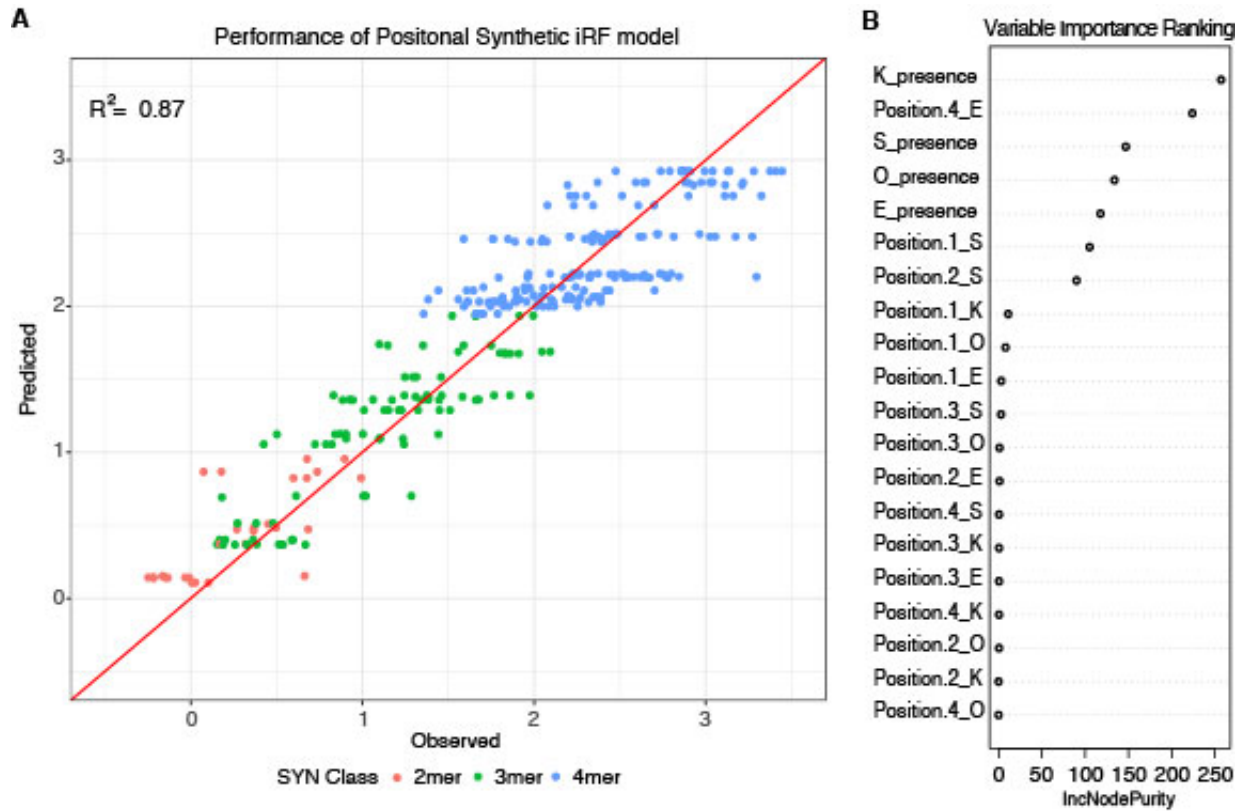
## Figures



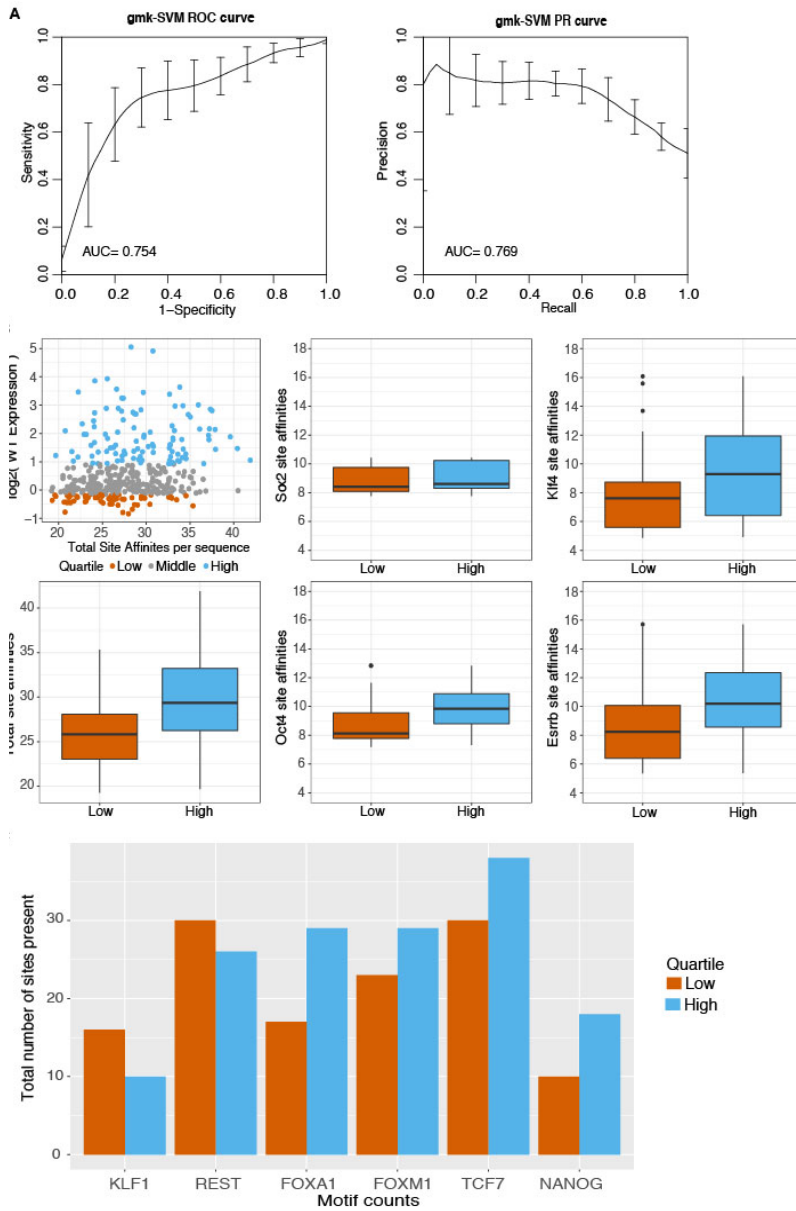
**Figure 2.1. Activity of synthetic elements and genomic sequences.** (A) The activity of synthetic elements with different numbers of binding sites. Expression is the average log of the ratio of cDNA barcode counts/DNA barcode counts for each synthetic element normalized to basal expression (dotted 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 TFBSs mutated (gMUT sequences). Red indicates sequences with significantly different expression between matched gWT and gMUT sequences. Diagonal solid line is expectation if mutation of TFBSs 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).



**Figure 2.2. 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  $\pm$  SEM (black whiskers). Green line for comparison to expression of 3-mer; Green transparency highlights SEM of 3-mer shown. Activity of the ten highest (B) and ten lowest (C) expressing 4-mers. Red line represents average expression of all synthetic 4-mer elements. Capital letter represents binding site in forward orientation and lower-case letter represents binding site in reverse orientation. Mean expression of each element across barcodes (black dot)  $\pm$  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.



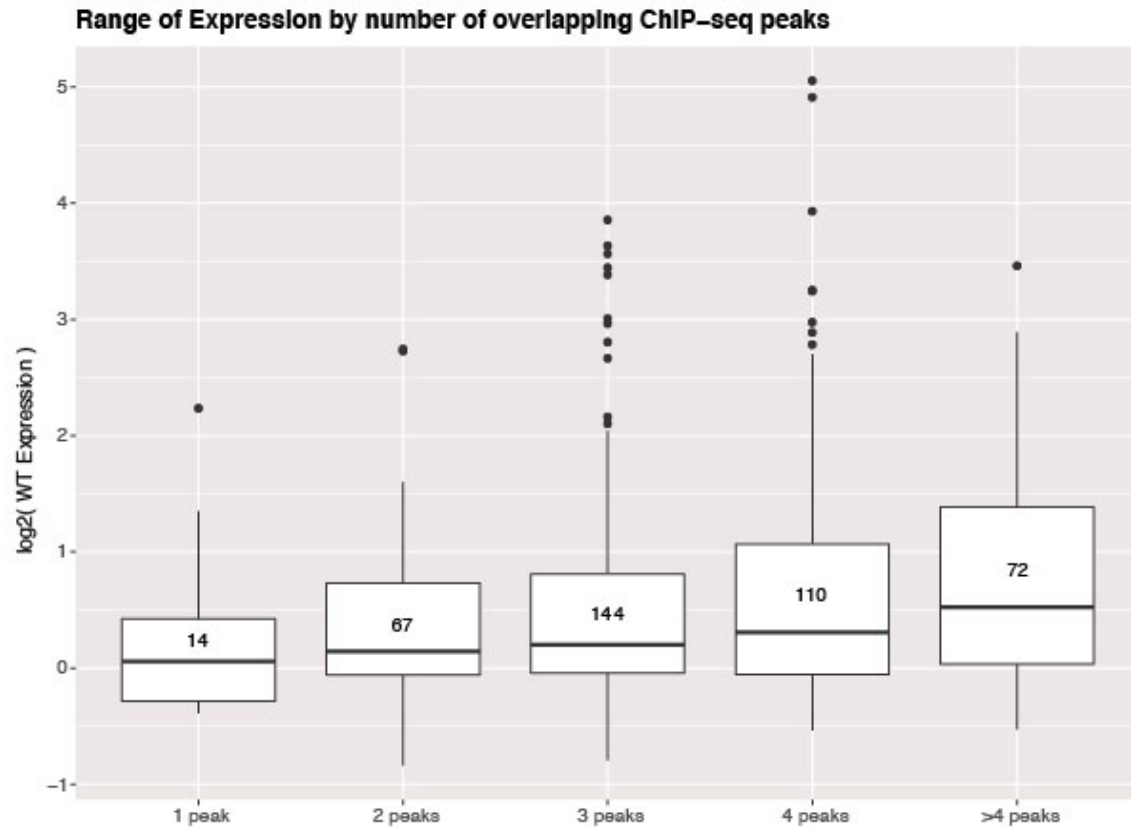
**Figure 2.3. Positional grammar in synthetic elements.** (A) Iterative random forest (iRF) regression model that includes features for presence and position of pluripotency TFBSs predicts relative expression of synthetic elements. Number of binding site per element is indicated in pink (2-mers), green (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.



**Figure 2.4. Sequence features separate active and inactive genomic sequences. (A)**

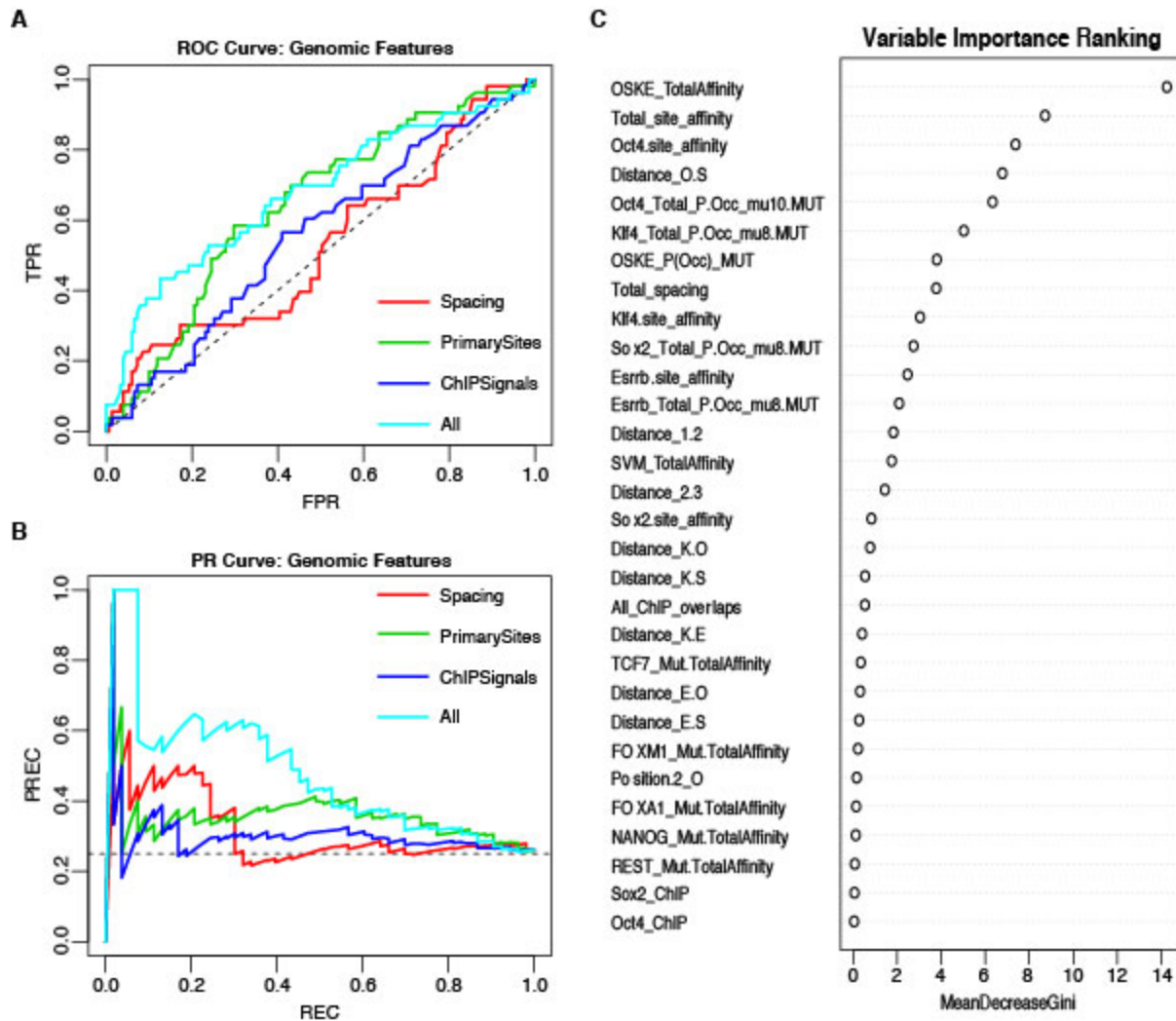
Performance of gkmer-SVM for genomic sequence supports contribution of sequence based features to activity. Word length of 8 bp with gap size of 2 bp was used for training with 3-fold cross validation. ROC curve (left panel) and PR curve (right panel) is plotted for the average across 3-fold cross-validation sets +/- standard deviation. (B) 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, n = 101) and low genomic sequences (bottom 25% as ranked by

expression, n =101). Total site affinities (bottom left panel) is calculated per sequences by summing predicted affinity of the three primary sites present in each sequence. (C) Total number of occurrences of TFBSs for additional TFs in high and low sequences (stratified as in B), as determined by motif scanning, excluding primary (O,S,K,E) sites.



**Figure 2.5. 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.





**Figure 2.6. Performance of iRF classification models that include features specific to genomic sequences.** (A) ROC Curve and (B) Precision-Recall (PR) Curve comparing genomic iRF models. Color indicates set of features used to train model. (C) Variable importance as evaluated for the feature by the average reduction in the Gini index, which is based on node impurity during training (Xi Chen and Ishwaran 2012).

## **Supplemental Materials**

Supplemental Figures S2.1-S2.7

Supplemental Tables 2.1-2.6

*Supplemental Files:*

Supplemental File 1: SyntheticLibraryDesign.txt

Supplemental File 2: GenomicLibraryDesign.txt

Supplemental File 3: SYN\_ExpressionSummary.txt

Supplemental File 4: GEN\_ExpressionSummary.txt

Supplemental File 5: gkmSVM\_8merScoreWeights.txt

Supplemental File 6: gkmSVM\_8merTop50TOMTOMe27.txt

Supplemental File 7: SYN\_FeaturesiRF.txt

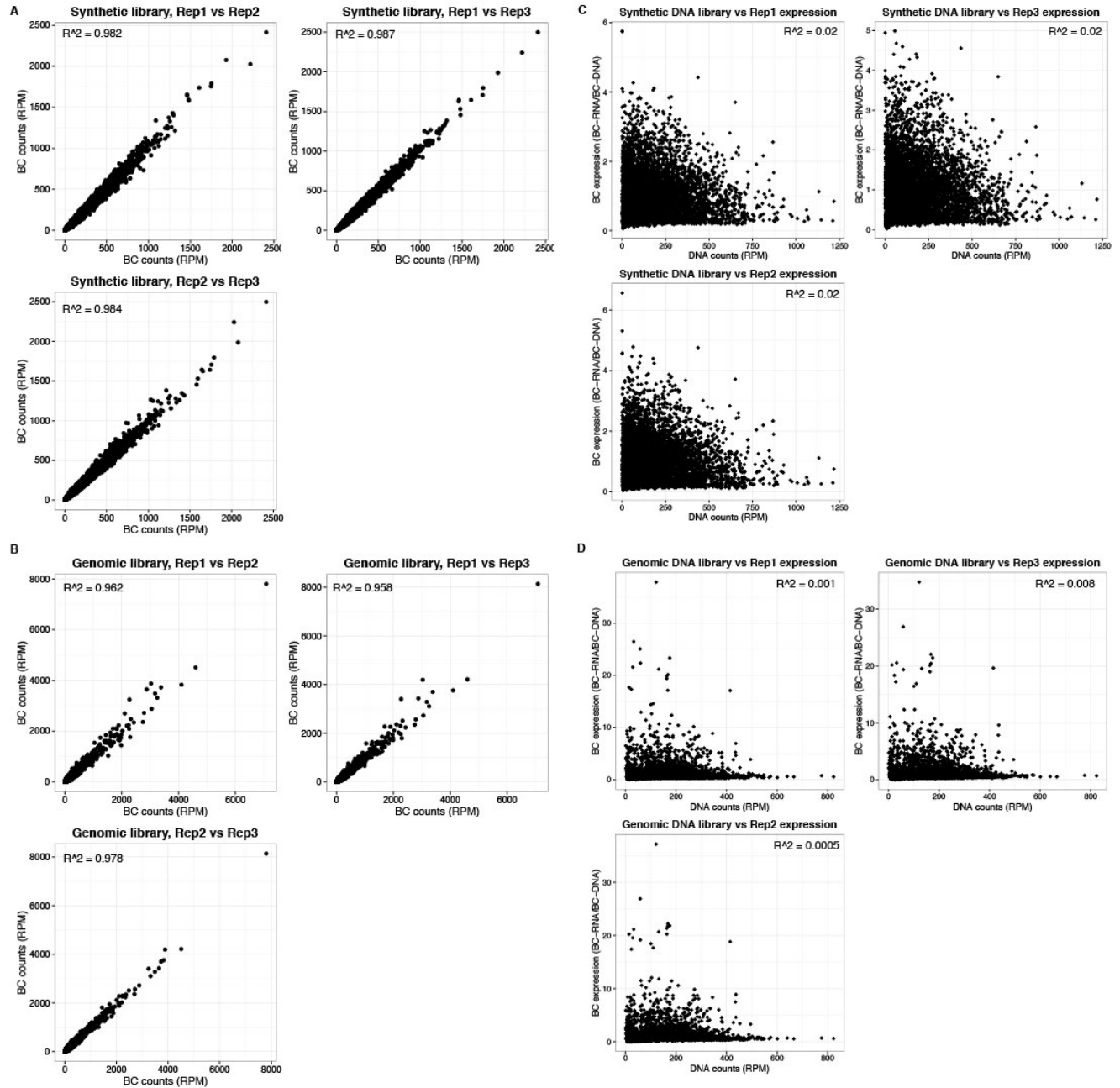
Supplemental File 8: gWT\_FeaturesiRF.txt

Supplemental File 9: SYN\_all\_RPM.txt

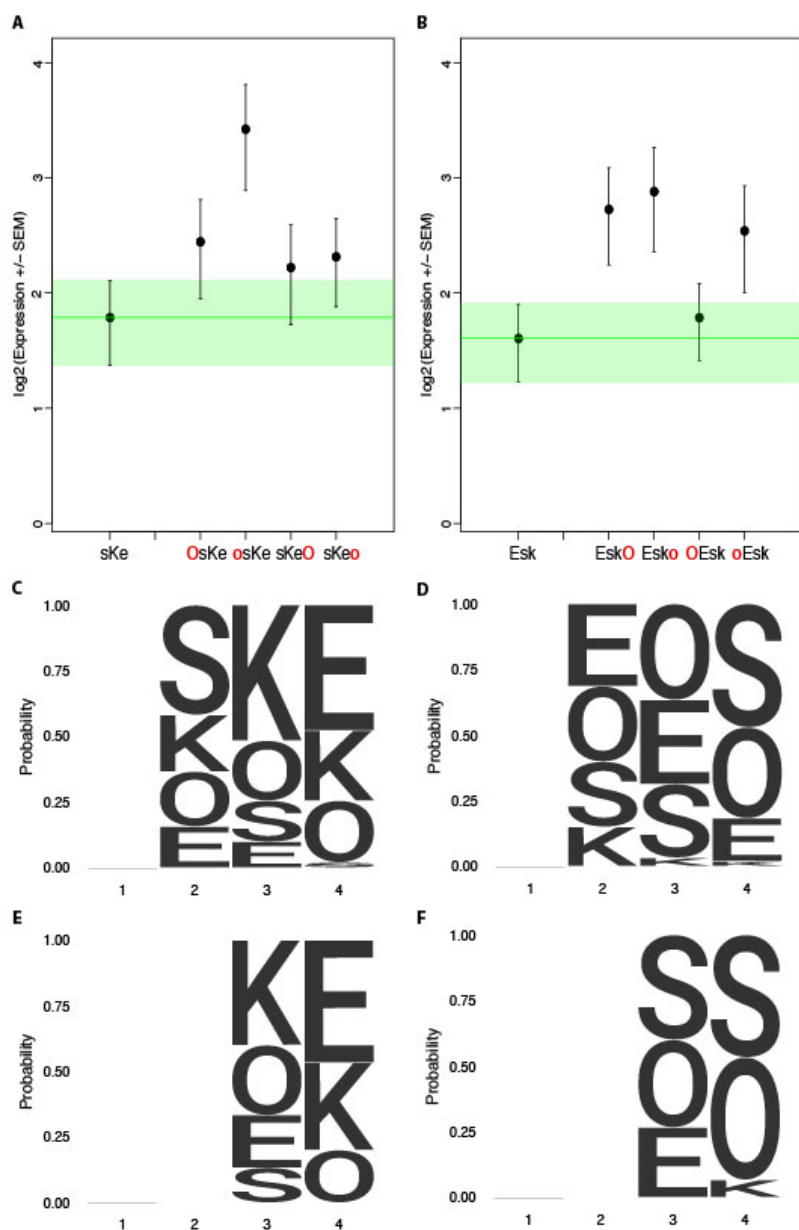
Supplemental File 10: GEN\_all\_RPM.txt

Supplemental File 11: gkmSVM\_8merInputs.fasta

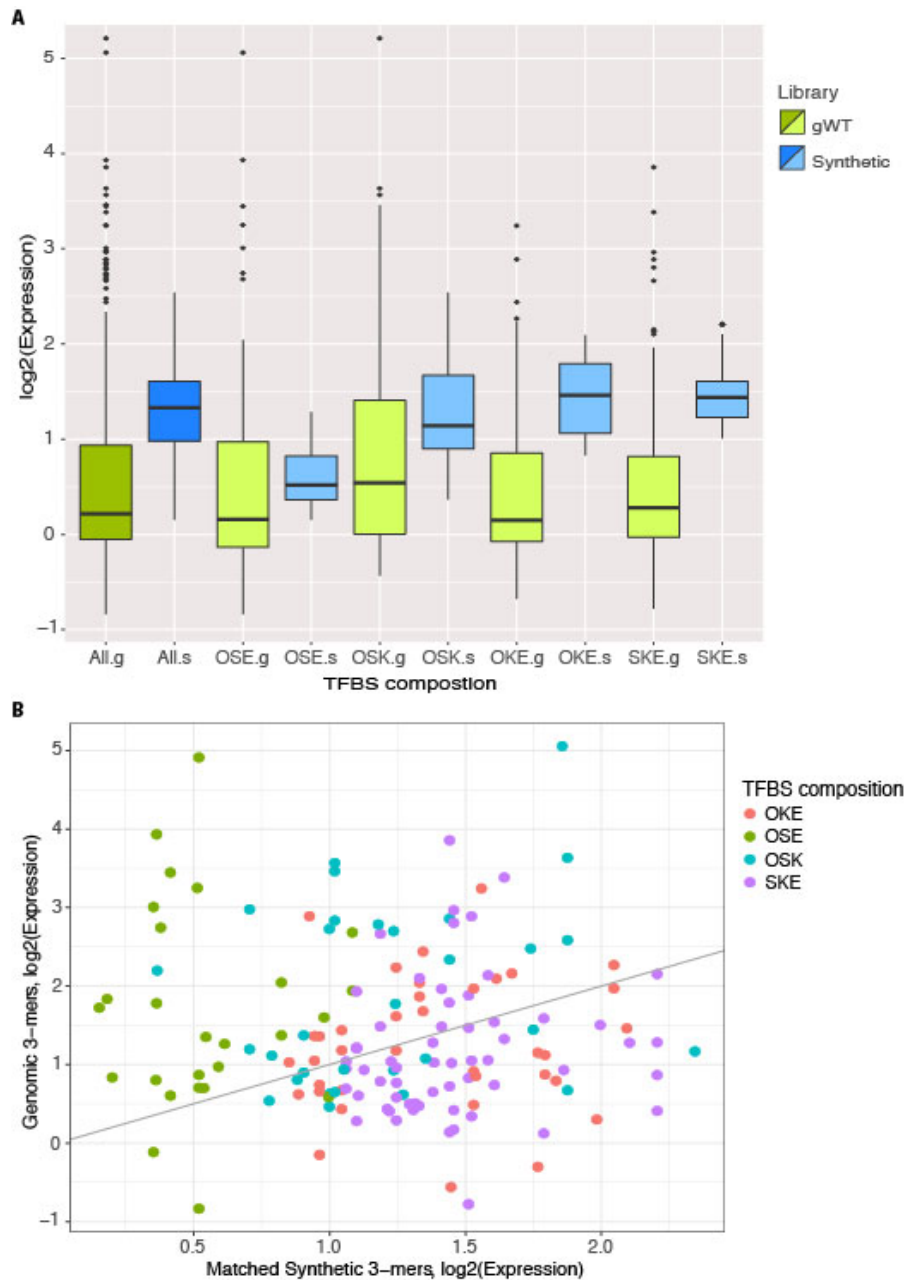
Supplemental File 12: gkmSVM\_8merOutputScores.txt



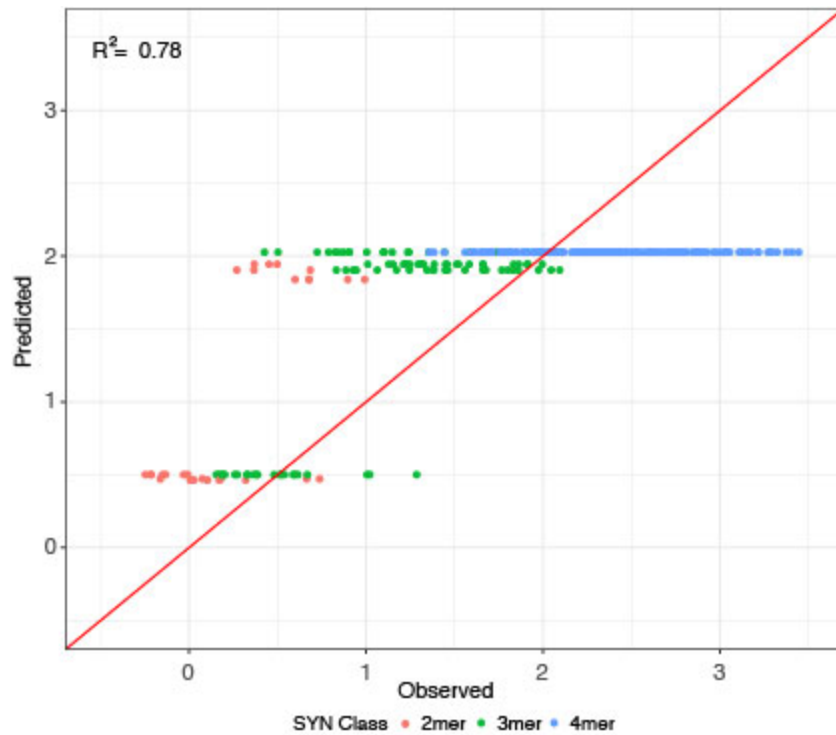
**Figure 2.S1. MPRA data quality.** Reproducibility of barcode (BC) counts between biological replicates, normalized as reads per million per RNA replicate for (A) Synthetic library and (B) Genomic, gWT and gMUT, library. Comparison of normalized BC expression ( $BC_{RNA}/BC_{DNA}$ ) versus DNA counts for (C) Synthetic library and (D) Genomic, gWT and gMUT, library.



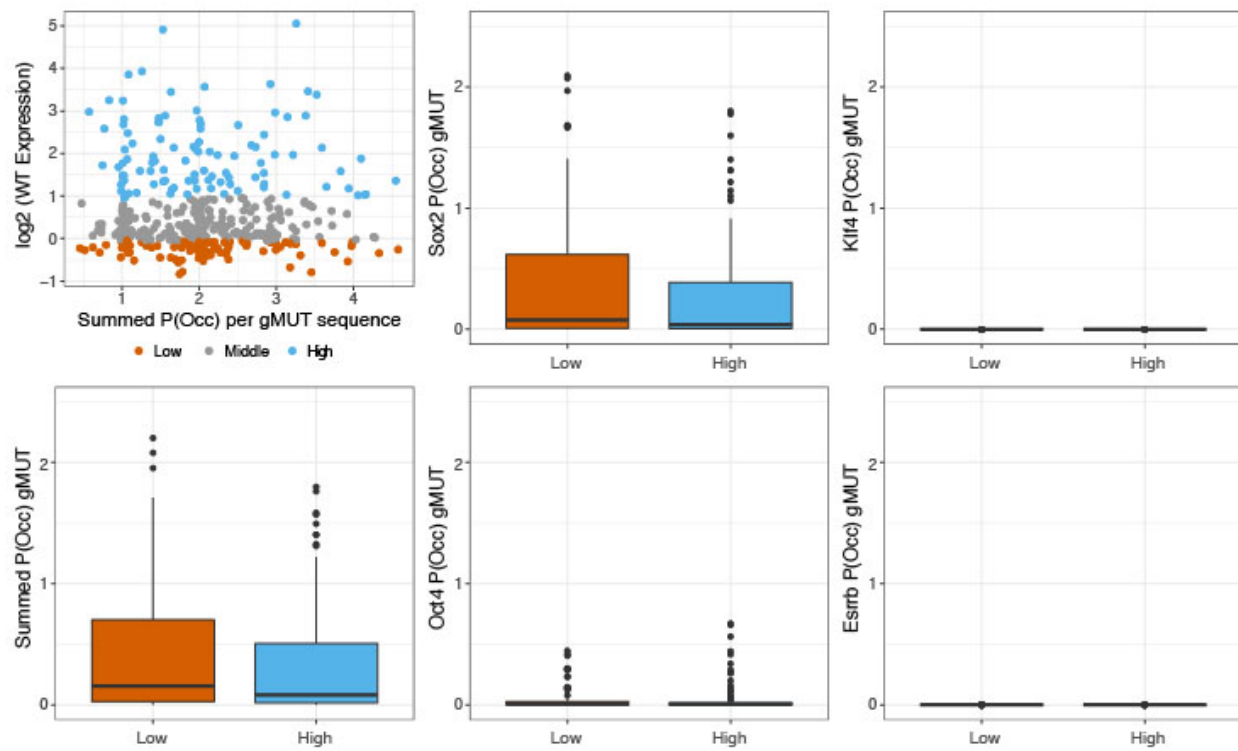
**Figure 2.S2. Additional examples of non-additivity in synthetic elements.** Comparisons of synthetic 3-mer elements with matched 4-mer elements containing one additional site in the first or fourth position with (A) three of four matched 4-mers with overlapping expression despite an additional binding site and (B) one of four matched 4-mers with overlapping expression. Activity logos for the top 25% (C), bottom 25% (D) of 3-mer synthetic elements (n= 48 each), and top 25% (E), bottom 25% of 2-mer synthetic elements (n= 12 each). Height of letter is proportional to frequency of site in indicated position. Positions organized as in **Figure 2**.



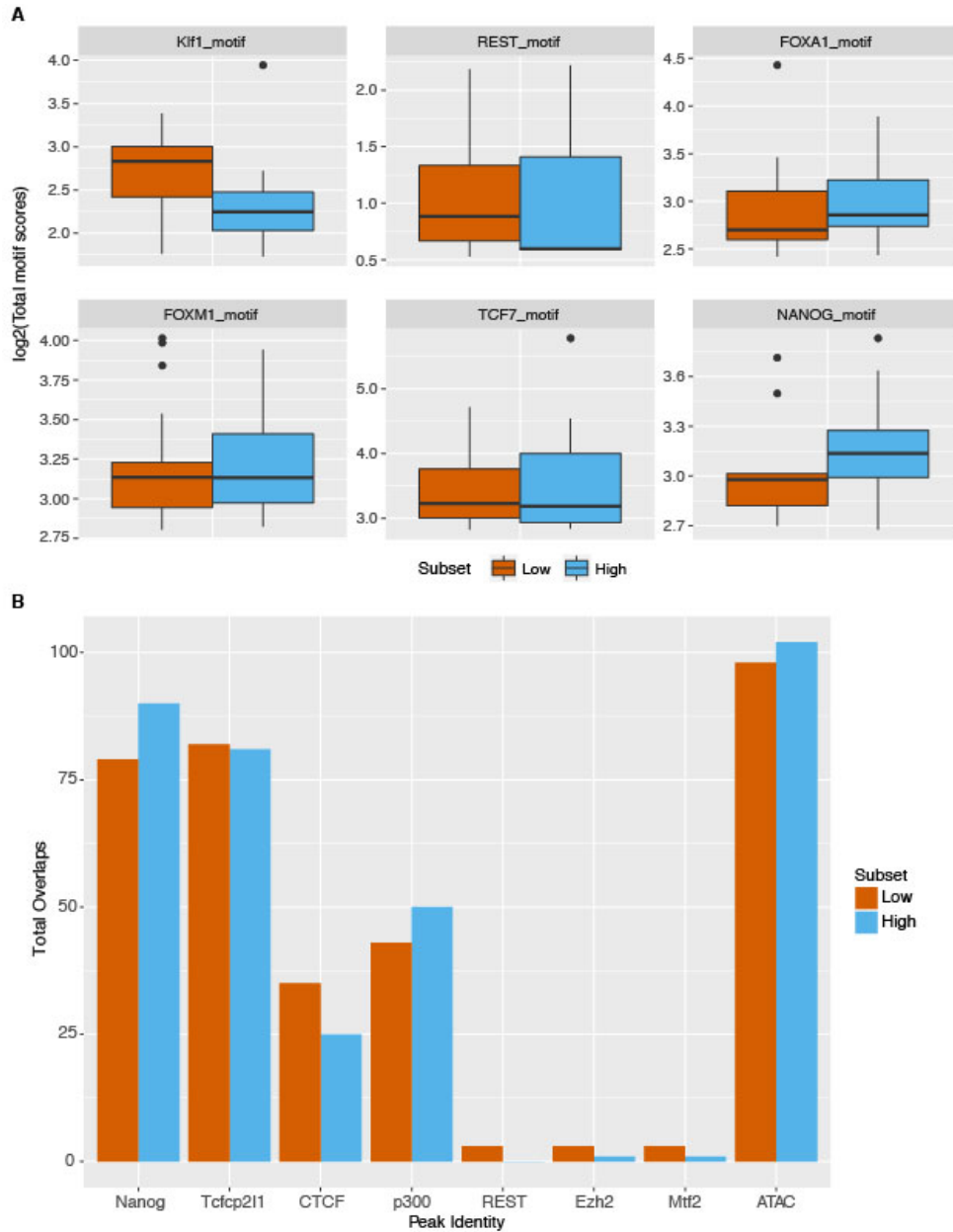
**Figure 2.S3. Comparison of synthetic and genomic patterns of transcription factor binding sites (TFBSs).** (A) Expression ( $\log_2$ ) of all synthetic (dark blue) and gWT (dark green) library members subset by TFBS composition (light blue and light green, respectively). (B) Expression ( $\log_2$ ) of synthetic (x-axis) and gWT (y-axis) library members, matched by composition and order of binding sites for OCT4 (O), SOX2 (S), KLF4 (K), and ESRRB (E). Subsets of TFBS composition indicated by color. Grey line indicates x-y diagonal as axis scales differ.



**Figure 2.S4. Additive effects in synthetic elements.** Iterative random forest (iRF) regression model that includes features for only presence of pluripotency TFBSs to predict the relative expression of synthetic elements. Number of binding site per element indicated as in **Figure 2.3**. Observed and predicted expression are both plotted in  $\log_2$  space.

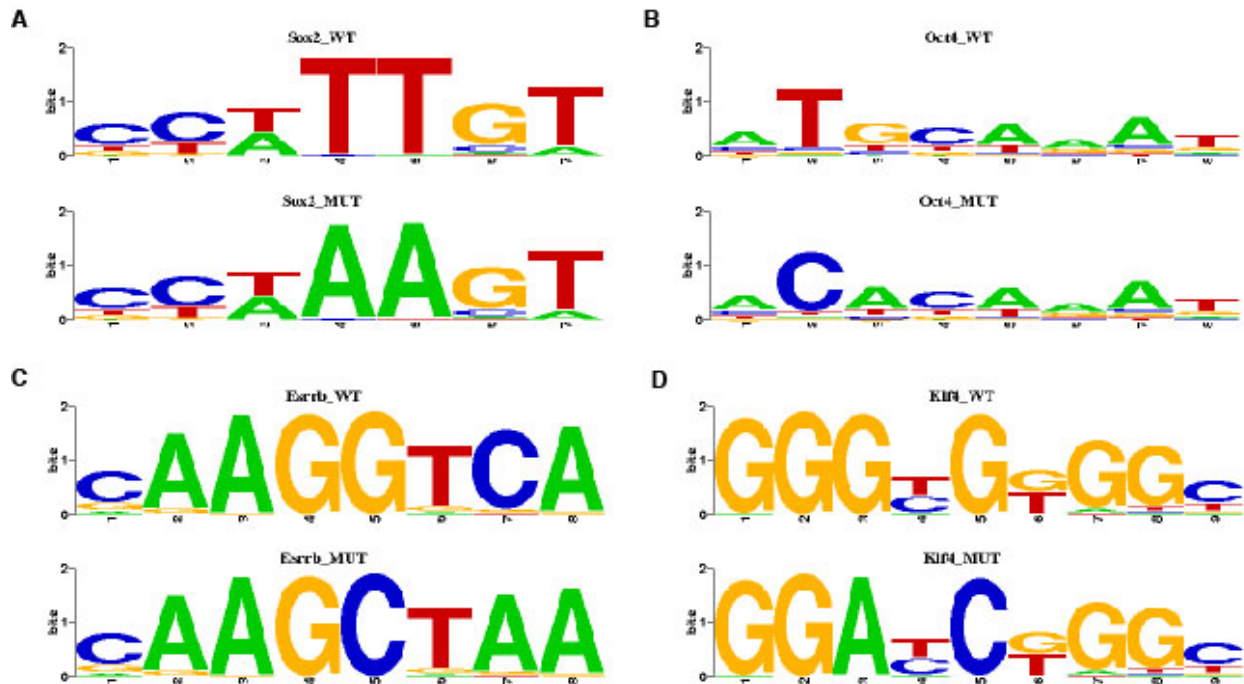


**Figure 2.S5. Predicted occupancy of genomic sequences.** Predicted occupancy (P(Occ)) for genomic sequences in the absence of the primary pluripotency sites (gMUT sequences) for high assumed protein concentration ( $\mu$ ) for SOX2 ( $\mu = 8$ ), OCT4 ( $\mu = 10$ ), KLF4 ( $\mu = 8$ ), and ESRRB ( $\mu = 8$ ) shown in middle and right panels. Summed P(Occ) of all factors per gMUT sequence, compared to expression (top left panel) or binned as low or high library members (bottom 25% and top 25% of sequences, ranked by gWT expression,  $n = 101$ ).



**Figure 2.S6. Genomic sequences show signatures for other factors. (A)** Summed motif scores for indicated motif across genomic sequences, excluding primary pluripotency sites. Site scores output during motif scanning of high (top 25% as ranked by gWT expression,  $n = 101$ ) and low (bottom 25% as ranked by gWT expression,  $n = 101$ ) gMUT sequences to prevent scoring of O, S, K, or E TFBS sequences. **(B)** Overlapping TF occupancy, as measured by CHIP-seq, or accessibility, as measured by ATAC-seq, for high (top 25% as ranked by gWT expression,  $n = 101$ ) and low (bottom 25% as ranked by gWT expression,  $n = 101$ ) genomic sequence intervals.





**Figure 2.S7. Pluripotency motif substitutions for gMUT sequences.** Highest information content positions in each motif were substituted with least frequent nucleotide for that position. (A) For mutating Sox2 motifs, the reference nucleotides were substituted for ‘A’ in position 4 and 5. (B) For mutating Oct4 motifs, the reference nucleotide was substituted for ‘C’ in position 2 and for ‘A’ in position 3. (C) For mutating Esrrb motifs, the reference nucleotide was substituted for ‘C’ in position 5 and ‘A’ for position 7. (D) For mutating Klf4 motifs, the reference nucleotide was substituted for ‘A’ in position 3 and ‘C’ in position 5.

**Table 2.1: SYN library composition**

<b>Element class</b>	<b>Unique Elements</b>	<b>Unique Element-Barcode Pairs</b>
2-mers	48	384
3-mers	192	1,536
4-mers	384	3,072
Basal	1	112
<i>Total library size</i>	<i>625</i>	<i>5,104</i>

**Table 2.2: gWT site composition**

<b>Sequence composition (Primary sites)</b>	<b>Unique Sequences</b>
OKE	117
OSE	65
OSK	68
SKE	157

**Table 2.3: gWT/gMUT library composition**

<b>Sequence class</b>	<b>Unique Sequences</b>	<b>Unique Sequence-Barcode Pairs</b>
gWT	407	3,256
gMUT	407	3,256
Basal	1	112
<i>Total library size</i>	<i>815</i>	<i>6,624</i>

**Table 2.4: Primer sequences**

Name	Sequence	Demultiplexing BC
Synthetic_FW-1	CTTCTACTACTAGGGCCCA	-
Synthetic_Rev-2	CATGAACTAGCATGTAGAGCTC	-
Genomic_FW-1	GACTTACATTAGGGCCCGT	-
Genomic_Rev-1	CAGTATCGTAGTCCGAGCTC	-
CF121	TAGCGTCGAGGACATCAAGA	-
CF122	TGGTTTGTCCAAACTCATCAA	-
CF150	TACACCGTGGTGGAGCAGTA	-
CF151b	AGCGTACTCGAGTTGTTAACTTGTTTATTGCAGCTT	-
CF52	AATGATACGGCGACCACCGAG	-
CF53	CAAGCAGAAGACGGCATA CGA	-
P1_XbaI_1_F	AATGATACGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGCTCTTCCGATCTAACCTCA	AACCTCA
P1_XbaI_1_R	/5Phos/C*TAGTGAGGTTAGATCGGAAGAGCGTCGT GTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGT ATCATT	AACCTCA
P1_XbaI_2_F	AATGATACGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGCTCTTCCGATCTTCTAAGC	TCTAAGC
P1_XbaI_2_R	/5Phos/C*TAGGCTTAGAAGATCGGAAGAGCGTCGT GTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGT ATCATT	TCTAAGC
P1_XbaI_3_F	AATGATACGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGCTCTTCCGATCTCTGTCAT	CTGTCAT
P1_XbaI_3_R	/5Phos/C*TAGATGACAGAGATCGGAAGAGCGTCGT GTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGT ATCATT	CTGTCAT
P1_XbaI_4_F	AATGATACGGCGACCACCGAGATCTACACTCTTTC	GGAGGTG

	CCTACACGACGCTCTTCCGATCTGGAGGTG	
P1_XbaI_4_R	/5Phos/C*TAGCACCTCCAGATCGGAAGAGCGTCGT GTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGT ATCATT	GGAGGTG
P1_XbaI_5_F	AATGATACGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGCTCTTCCGATCTGCTCGAT	GCTCGAT
P1_XbaI_5_R	/5Phos/C*TAGATCGAGCAGATCGGAAGAGCGTCGT GTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGT ATCATT	GCTCGAT
P1_XbaI_6_F	AATGATACGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGCTCTTCCGATCTTAGAGTA	TAGAGTA
P1_XbaI_6_R	/5Phos/C*TAGTACTCTAAGATCGGAAGAGCGTCGT GTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGT ATCATT	TAGAGTA
P1_XbaI_7_F	AATGATACGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGCTCTTCCGATCTTCAGTCT	TCAGTCT
P1_XbaI_7_R	/5Phos/C*TAGAGACTGAAGATCGGAAGAGCGTCGT GTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGT ATCATT	TCAGTCT
P1_XbaI_8_F	AATGATACGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGCTCTTCCGATCTTTCCAAG	TTCCAAG
P1_XbaI_8_R	/5Phos/C*TAGCTTGGAAGATCGGAAGAGCGTCGT GTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGT ATCATT	TTCCAAG
PE2_SIC69_SalI_F	/5Phos/T*CGAAGATCGGAAGAGCACACGTCTGAAC TCCAGTCACAGCGTGCCCATCTCGTATGCCGTCTT CTGCTTG	-
PE2_SIC69_SalI_R	CAAGCAGAAGACGGCATAACGAGATGGGCACGCTG TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	-

**Table 2.5: iRF SYN feature matrix**

<i>Model comparison</i>	<b>Billboard model</b>	<b>Positional model</b>	
<b>Test Set, R<sup>2</sup> (overall)</b>	0.56	0.87	
<b>Test Set, R<sup>2</sup> (4mers)</b>	0.00	0.52	
<b>Features included</b>			<b>Source</b>
O_presence	Yes	Yes	Same terms used for SYN iRF Billboard Model. Identity of TFBSs present in the sequence are determined via FIMO.
S_presence	Yes	Yes	
K_presence	Yes	Yes	
E_presence	Yes	Yes	
Position.4_O	(-)	Yes	Same terms used for SYN Billboard + Position iRF Model. Relative position of pluripotency TFBSs as determined via FIMO; as sequences contain only three primary sites Position.1 is FALSE for all factors and omitted from table.
Position.4_S	(-)	Yes	
Position.4_K	(-)	Yes	
Position.4_E	(-)	Yes	
Position.3_O	(-)	Yes	
Position.3_S,	(-)	Yes	
Position.3_K	(-)	Yes	
Position.3_E	(-)	Yes	
Position.2_O	(-)	Yes	
Position.2_S	(-)	Yes	
Position.2_K	(-)	Yes	
Position.2_E	(-)	Yes	
Position.1_O	(-)	Yes	
Position.1_S	(-)	Yes	
Position.1_K	(-)	Yes	
Position.1_E	(-)	Yes	

**Table 2.6: iRF gWT Feature Matrix**

<i>Model Comparison</i>	<b>Billboard</b>	<b>Positional</b>	<b>Spacing</b>	<b>Primary sites</b>	<b>ChIP</b>	<b>All</b>	
<b>AUROC</b>	0.52	0.47	0.52	0.64	0.56	0.67	
<b>AUPRC</b>	0.22	0.25	0.31	0.34	0.29	0.46	
<b>Features included</b>							<b>Source</b>
O_presence	Yes	Yes	(-)	(-)	(-)	(-)	Same terms used for SYN iRF Billboard Model. Identity of TFBSs present in the sequence are determined via FIMO.
S_presence	Yes	Yes	(-)	(-)	(-)	(-)	
K_presence	Yes	Yes	(-)	(-)	(-)	(-)	
E_presence	Yes	Yes	(-)	(-)	(-)	(-)	
Position.4_O	(-)	Yes	(-)	(-)	(-)	(-)	Same terms used for SYN Positional iRF Model. Relative position of pluripotency TFBSs as determined via FIMO; as sequences contain only three primary sites Position.1 is FALSE for all factors and omitted from table.
Position.4_S	(-)	Yes	(-)	(-)	(-)	(-)	
Position.4_K	(-)	Yes	(-)	(-)	(-)	(-)	
Position.4_E	(-)	Yes	(-)	(-)	(-)	(-)	
Position.3_O	(-)	Yes	(-)	(-)	(-)	(-)	
Position.3_S	(-)	Yes	(-)	(-)	(-)	(-)	
Position.3_K	(-)	Yes	(-)	(-)	(-)	(-)	
Position.3_E	(-)	Yes	(-)	(-)	(-)	(-)	
Position.2_O	(-)	Yes	(-)	(-)	(-)	(-)	
Position.2_S	(-)	Yes	(-)	(-)	(-)	(-)	
Position.2_K	(-)	Yes	(-)	(-)	(-)	(-)	
Position.2_E	(-)	Yes	(-)	(-)	(-)	(-)	
Distance_O.S	(-)	(-)	Yes	(-)	(-)	Yes	Distance between FIMO identified sites (OCT4, SOX2, KLF4, & ESRRB); if a site is absent from the sequence distance between the two factors is sent to
Distance_K.E	(-)	(-)	Yes	(-)	(-)	Yes	
Distance_K.O	(-)	(-)	Yes	(-)	(-)	Yes	
Distance_K.S	(-)	(-)	Yes	(-)	(-)	Yes	
Distance_E.S	(-)	(-)	Yes	(-)	(-)	Yes	
Distance_E.O	(-)	(-)	Yes	(-)	(-)	Yes	

							total length of sequence (81 or 82 bps).
Distance_1.2	(-)	(-)	Yes	(-)	(-)	Yes	Distance between sites, regardless of identity, present in sequences (1st to 2nd, 2nd to 3rd site).
Distance_2.3	(-)	(-)	Yes	(-)	(-)	Yes	
Total_spacing	(-)	(-)	(-)	(-)	(-)	Yes	Sum of 'Distance_1.2' and 'Distance_2.3'
Oct4.site_affinity	(-)	(-)	(-)	Yes	(-)	Yes	Scores assigned by FIMO, a log-likelihood score ratio based on the PWM provided (Grant et al. 2011).
Sox2.site_affinity	(-)	(-)	(-)	Yes	(-)	Yes	
Klf4.site_affinity	(-)	(-)	(-)	Yes	(-)	Yes	
Esrrb.site_affinity	(-)	(-)	(-)	Yes	(-)	Yes	
OSKE_TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	Sum of OCT4, SOX2, KLF4, & ESRRB site affinities for each sequence (above terms)
Oct4.Occ_10	(-)	(-)	(-)	(-)	(-)	Yes	Total predicted occupancy across the sequence for each pluripotency factor, annotated with custom code (See Methods)
Sox2.Occ_8	(-)	(-)	(-)	(-)	(-)	Yes	
Klf4.Occ_8	(-)	(-)	(-)	(-)	(-)	Yes	
Essrb.Occ_8	(-)	(-)	(-)	(-)	(-)	Yes	
OSKE_P(Occ)	(-)	(-)	(-)	(-)	(-)	Yes	Sum of predicted occupancies for pluripotency factors (above terms)
Klf1_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	Number of identified sites for gMUT sequences scanned using FIMO with PWMs of SVM supported
REST_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	

FOXA1_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	factors. gMUT sequences were scored to prevent assigning a score for another factor to any of the primary pluripotency sites.
FOXM1_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	
TCF7_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	
NANOG_Mut.count	(-)	(-)	(-)	(-)	(-)	Yes	
KLF1_Mut.TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	Sum of scores assigned for FIMO scanning with PWMs of SVM supported factors for gMUT sequences. gMUT sequences were scored to prevent assigning a score for another factor to primary pluripotency sites.
REST_Mut.TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	
FOXA1_Mut.TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	
FOXM1_Mut.TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	
TCF7_Mut.TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	
NANOG_Mut.TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	
SVM_TotalAffinity	(-)	(-)	(-)	(-)	(-)	Yes	Sum of KLF1, REST, FOXA1, FOXM1, TCF7, & Nanog site affinities (above).
Total_site_affinity	(-)	(-)	(-)	(-)	(-)	Yes	Sum of 'SVM_TotalAffinity' and 'OSKE_TotalAffinity'.
Oct4_ChIP	(-)	(-)	(-)	(-)	Yes	Yes	Chen et al. ChIP-seq overlaps. GEO dataset: GSE11431; GEO IDs: GSM288346 (O), GSM288347 (S), GSM288354 (K),
Sox2_ChIP	(-)	(-)	(-)	(-)	Yes	Yes	
Klf4_ChIP	(-)	(-)	(-)	(-)	Yes	Yes	
Essrb_ChIP	(-)	(-)	(-)	(-)	Yes	Yes	



							GSM288355 (E); E14 mESCs.
Nanog	(-)	(-)	(-)	(-)	Yes	Yes	Additional ChIP-seq peaks from Chen et al. GEO dataset: GSE11431; GEO IDs: GSM288345 (Nanog), GSM288350 (Tcfcp211), GSM288351 (CTCF), GSM288359 (p300); E14 mESCs.
Tcfcp211	(-)	(-)	(-)	(-)	Yes	Yes	
CTCF	(-)	(-)	(-)	(-)	Yes	Yes	
p300	(-)	(-)	(-)	(-)	Yes	Yes	
Peak_count	(-)	(-)	(-)	(-)	(-)	Yes	Total overlapping pluripotency ChIP seq signals for all peaks from Chen et al. (above), including O,S,K,E, Nanog, Tcfcp211, CTCF, & p300.
REST	(-)	(-)	(-)	(-)	Yes	Yes	ChIP-seq from Yu et al. PMID: 21632747; GEO dataset: GSE28233; GEO ID: GSM698696; E14 mESCs.
Mtf2	(-)	(-)	(-)	(-)	Yes	Yes	ChIP-seq from Perino et al. PMID: 29808031; GEO dataset: GSE94300; MAnorm bed files from dataset used for respective factors; E14 mESCs.
Ezh2	(-)	(-)	(-)	(-)	Yes	Yes	
H3K27me3	(-)	(-)	(-)	(-)	Yes	Yes	
H3K4me3	(-)	(-)	(-)	(-)	Yes	Yes	
ATAC	(-)	(-)	(-)	(-)	Yes	Yes	ATAC-seq from Wu et al. PMID: 27309802; GEO ID: GSM2156965; 50k cell stage of mESCs.
H3K27ac	(-)	(-)	(-)	(-)	Yes	Yes	All files downloaded from <a href="http://www.encodeproject.org">www.encodeproject.org</a> . Note: all data sets from Yue et al. (PMID: 25409824)
H3K36me3	(-)	(-)	(-)	(-)	Yes	Yes	
H3K4me1Ren	(-)	(-)	(-)	(-)	Yes	Yes	

H3K4me1Snyder	(-)	(-)	(-)	(-)	Yes	Yes	are under review due to library complexity and/or read depth issues. E14 mESCs.
H3K4me3Ren	(-)	(-)	(-)	(-)	Yes	Yes	
H3K4me3Snyder	(-)	(-)	(-)	(-)	Yes	Yes	
H3K9ac	(-)	(-)	(-)	(-)	Yes	Yes	
H3K9me3	(-)	(-)	(-)	(-)	Yes	Yes	

## **Chapter 3: Evaluating fitness predictions for identification of cis-regulatory variants using massively parallel reporter assays**

Identifying regions of the human genome that drive regulatory activity is still a major challenge and prioritization of variants that may impact that key processes the thousands catalogued is even more difficult. Massively Parallel Reporter Assays (MPRAs) can be used to assess the regulatory potential of genomic sequences by providing the means to test hundreds or thousands of sequences at once and allows for direct comparisons between the activity of reference and alternate alleles without possible confounding variables such as chromatin state or the presence of other variants in the loci. Efforts have been made to integrate functional annotations and evolutionary and/or polymorphism metrics to predict regions or variants that contribute to fitness, or the quantitative representation of natural selection. To assess the predictive power of the genome classifications and any additional information that may be provided by fitness predictions such as those generated by programs like fitCons or CADD (Kircher et al. 2014; Gulko et al. 2015), we tested reference sequences from genome-wide regulatory segmentations and found that sequences with higher fitCons scores have higher regulatory activity. To assess the utility of fitness predictions for prioritizing variants that may impact *cis*-regulation, we tested reference and alternate alleles from the 1000 genome database and found that neither high fitCons or high CADD predictions were enriched for variants that drove significant allelic skew. Our findings highlight the need for critical evaluation of computational tools for and the challenge of predicting the regulatory impact of variants in noncoding regions of the human genome.

This chapter is a manuscript currently being prepared for submission, *Evaluating fitness predictions for identification of cis-regulatory variants using massively parallel reporter assays*,

with Avinash Ramu and Barak Cohen. I am the first author and contributed to the writing of the manuscript along with Barak Cohen. I performed the experiments and analyzed the initial sequencing data. Avinash Ramu performed CADD annotations and analyses. I performed all additional analyses.

## Introduction

A major challenge for interpreting growing collections of noncoding human variation, particularly variants associated with disease, is the prioritization of the sometimes thousands of associated variants for follow-up. Estimates for the proportion of functional DNA in the human genome have ranged from as little as 2.5-12%, based on sequence constraints and comparative genomics, to over 80% based on experimentally measured biochemical signals (Castillo-Davis 2005; Ponting and Hardison 2011; ENCODE Project Consortium 2012; Eddy 2013). With the majority of genome-wide association study (GWAS) hits mapping to noncoding regions, determining which regions of the genome are functional in relevant cells types is critical for identifying the causal variants underlying these associations (Hrdlickova et al. 2014; Stranger, Stahl, and Raj 2011). However, identifying functional regions of the genome and predicting the impact of individual noncoding variants remains difficult. In this study, we sought to determine the value of fitness predictions for identifying genomic sequences with regulatory potential in the lymphoblastoid cell line GM12878 and prioritizing variants that might impact *cis*-regulatory activity.

Evidence suggests that a large fraction of noncoding variants under selection may contribute to *cis*-regulatory activity but it remains difficult to identify variants that contribute to disease or organismal level traits in humans (Asthana et al. 2007). For example, lactase persistence into adulthood is a trait strongly associated with two noncoding variants, C/T-13910 and G/A-22018, that are at high allele frequencies in European populations and are thus likely under positive selection (Poulter et al. 2003; Ridefelt and Håkansson 2005). One possible causal noncoding variant likely modulates the expression of the enzyme lactase-phlorizin hydrolase, which breaks down lactose into glucose and galactose (Ranciaro et al. 2014; Tishkoff et al. 2007). Additional distal putative regulatory variants for the same enzyme have been found in pastoral African populations that consumed milk, supporting selective pressure on the expression of this gene (Ranciaro et al. 2014; Tishkoff et al. 2007; Swallow 2003). In contrast, a common noncoding

variant in a locus associated with levels of plasma low-density lipoprotein cholesterol (LDL-C), heart attack, and coronary artery disease, impacts the expression of a key liver enzyme, SORT1, when tested in hepatic cell lines (Musunuru et al. 2010; Linsel-Nitschke et al. 2010). Functional assays indicate that the modulation of SORT1 transcript abundance alters the associated phenotype of plasma LDL-C, potentially explaining the observed disease risks (Musunuru et al. 2010). Additional regulatory variants have been found with links to phenotypes such as diabetes (Kulzer et al. 2014), hepatitis C clearance (Lu et al. 2015), transcriptional response to hypoxia (Roche et al. 2016) and hypertension (Rana et al. 2017). These examples suggest a complex link between the impact of a genetic variant on *cis*-regulatory activity and expected fitness contributions, i.e. the quantitative representation of natural selection, but highlight the need for higher throughput approaches to prioritize variants beyond single loci for detailed, high effort follow up experiments.

Higher throughput approaches to identify variants of interest from human populations have traditionally included identifying variants associating expression quantitative trait loci (eQTLs) with quantitative expression differences between individuals. Studies that aim to identify eQTLs, including large scale studies in GM12878 cell lines, assume that susceptibility to common disease, a possible contributor to fitness, is at least in part mediated by variation in gene expression (Cookson et al. 2009; Lappalainen et al. 2013). This assumption is supported by evidence that variants associated with complex human phenotypes are more likely to be associated with differences in transcript abundance than variants with matched allele frequencies (Nicolae et al. 2010). However, a eQTL for a single gene can include hundreds of significantly associated variants (Lappalainen et al. 2013; Mohammadi et al. 2017). Even intersecting significantly associated variants with additional data, such as biochemical signals or functional annotations the ENCODE project can leave several dozen variants with equal likelihood of being causal, limiting follow up experiments from testing multiple loci (Oldoni et al. 2016).

Functional annotations, evolutionary evidence, and even associations with expression changes alone have been insufficient to prioritize variants on a genome-wide scale. Therefore, integrative computational tools that predict what regions of the genome are sensitive to substitutions or which noncoding variants could impact key genomic functions, including *cis*-regulation, are an attractive approach for the prioritization of variants necessary for follow-up experiments. Several groups have published computational approaches to make predictions of the expected impact of human genetic variation. Two such programs, fitCons and CADD, integrate diverse annotations into a single metric (Gulko et al. 2015; Kircher et al. 2014), to characterize the fitness consequence and deleteriousness of variants, respectively, with a shared goal of characterizing the functional consequences of substitutions genome-wide. While fitCons uses functional data from the ENCODE project in several cell types to group sequences and assign scores based on divergence and polymorphism frequencies (Gulko et al. 2015), CADD contrasts observed variants that are fixed or nearly fixed in human populations to simulated substitutions as proxies for rare mutations, which are hypothesized to be more likely to impact organismal fitness, to assign scores of likely deleteriousness (Kircher et al. 2014).

Although there are accepted benchmarks for evaluating algorithms that predict the impact of variants on coding sequences, such as predicting protein truncations and evaluating *in vivo* function or phenotypes, similar benchmarks have not been standardized for evaluating predictions in noncoding sequences (Ng and Henikoff 2003; Miosge et al. 2015; Walters-Sen et al. 2015). For fitCons, performance was assessed by relying on correlations with measurements such as ChIP-seq (Gulko et al. 2015), although only a fraction of bound regions are likely to have biochemical activity (White et al. 2013, 2016; Chaudhari and Cohen 2018). For CADD, performance was assessed by generalizing correlations of activity from saturation mutagenesis of three validated *cis*-regulatory sequences to the genome-wide predictions (Kircher et al. 2014). Because many propose using *cis*-regulatory activity as a possible read out of fitness or deleteriousness, including the authors that developed fitCons and CADD, a more rigorous approach is to directly test these predictions. Although it remains unclear if the ability of a

sequence to modulate expression in the context of a reporter assay corresponds to *in vivo* regulatory activity, transient plasmid-based assays such as MPRA have the advantage of determining the *cis*-regulatory potential of a sequence in isolation and allow us to measure the exact impact of variants on expression (Castillo-Davis 2005; Ulirsch et al. 2016; Chaudhari and Cohen 2018). Tewhey and colleagues used MPRA to identify significant allelic skew, the ratio of alternative allele and reference allele activity, for 2.6% of tested eQTL-associated and control variants in two lymphoblastoid cell lines and identify a causal variant for one locus using genome engineering, showing that MPRA have value in identifying putative regulatory variants (Tewhey et al. 2016).

Here we use MPRA to evaluate the predictive power of two conservation & population genetics-based algorithms for the *cis*-regulatory activity of sequences from the human genome in the lymphoblastoid cell line GM12878, an ENCODE tier one cell line that was generated as part of the International HapMap project and has extensive annotation data available as well as being used as a model cell line for B-cell related diseases (International HapMap Consortium 2005; Arvey et al. 2012; Sanyal et al. 2012; H. Zhou et al. 2015; Oakes et al. 2016). We tested putative regulatory sequences and variants to understand how predictions of fitness relate to both level and allelic differences for *cis*-regulatory activity. We find that 60% of regions annotated as regulatory by chromatin-based models are active in our assay, with a surprising fraction acting as silencers or repressors. We also find that approximately half of tested variants have a detectable impact on expression but that neither CADD or fitCons scores help prioritize variants with allelic effects. These results suggest that caution should be exercised if CADD and fitCons are used to prioritize variants that may have *cis*-regulatory impacts and that alternative computational and experimental approaches for prioritizing noncoding variants may need to be developed.

## Results



## **Rationale and description of *cis*-regulatory libraries**

If natural selection acts on the *cis*-regulatory activity of genomic sequences, we would expect that noncoding sequences predicted to contribute to fitness to be enriched for *cis*-regulatory activity, either as activating or repressing sequences. To test this prediction, we designed a library of reporter genes (REF) composed of noncoding sequences from the reference human genome (genome build hg19). We chose sequences from four annotation classes for the GM12878 cell line, Weak Enhancers (WE), Enhancers (E), Predicated Promoter Flank (PF), and Repressed (R), based on chromatin data from the ENCODE project (Flicek et al. 2013; ENCODE Project Consortium 2012). For each annotation class, we chose sequences that span the range of possible scores from fitCons. For every nucleotide in the genome, fitCons assigns a *rho* score, which represents the probability that the nucleotide contributes to fitness. The resulting library of ~1,400 sequences allow us to assess the predictive power of the genome classifications compared to any additional information fitCons may provide for selecting genomic regions with *cis*-regulatory activity.

If natural selection acts on regulatory elements, then we might expect that noncoding variants to impact *cis*-regulatory activity. As fitCons predicts the probability of a sequence being under selection, a high probability of selection would suggest that the sequence may also be sensitive to substitutions (Gulko et al. 2015). Therefore, we designed a variant (VAR) library to test this prediction. We selected reference sequences from the WE annotation, as we previously found this annotation to be enriched for sequences with regulatory activity (Kwasnieski et al. 2014) and to control for confounding effects that could result from comparing sequences with different annotations. As in the REF library, we chose sequences across the range of fitCons scores for noncoding sequences. The coordinates for these sequences were then used to identify alternative alleles reported in the 1000 Genome Project database (Lappalainen et al. 2013). The resulting VAR library is composed of 403 reference sequences and their matching alternative allele sequences. By assaying the activities of these pairs of sequences, we tested the relationship

between predicted fitness and the impact of naturally occurring variation on cis-regulatory activity.

Together these libraries test two complementary hypotheses: 1) that regions with higher fitCons scores will be enriched for active regulatory DNA elements and 2) that variants in regions with higher fitCons or CADD scores will be enriched for having an impact on regulatory DNA activity.

### **MPRA of reporter gene libraries**

We assayed the *cis*-regulatory activity of the REF and VAR libraries in immortalized human primary B-lymphoblastoid cells (cell line GM12878; Coriell) using a plasmid-based MPRA (Patwardhan et al. 2012; Tewhey et al. 2016; Kwasnieski et al. 2014). Each unique library member described above was designed to be present in the library pools six times, paired with a unique barcode (BC) sequence (Kwasnieski et al. 2014; Chaudhari and Cohen 2018; Fiore and Cohen 2016). To determine the relative activity of each sequence, we included plasmids designed to include only the minimal promoter paired (See Methods) (White 2015; Mogno, Kwasnieski, and Cohen 2013; White et al. 2013; Kwasnieski et al. 2012; Fiore and Cohen 2016). Our BC counts were highly reproducible between biological replicates, with  $R^2$  between 0.994-0.996 for replicates of the REF library and 0.972-0.988 for the VAR library (Figure S1A-B). After thresholding on read counts for RNA and DNA samples, we recovered reads for 99.7% (1759/1763) of REF library sequences and 99.4% (691/697) of VAR library sequences, demonstrating the quality of these data. Expression was then calculated by computing the average of the ratio of  $BC_{RNA}$  counts to  $BC_{DNA}$  counts for each for each unique sequence tested. These data are of extremely high quality, particularly when compared to previous efforts to use MPRA to identify allele-specific expression, which should, in principle, increase our power to detect potential predictive power of CADD and fitCons (Ulirsch et al. 2016).

## **Predicted Fitness identifies activity of genomic sequences better than functional annotations alone**

Activity of ENCODE segmentation predictions consistent with functional testing in other cell types. For the REF library, we detect significant activity for 60% (1051/1763) of tested sequences, of which 24% (251/1051) drive expression above basal controls and 76% (800/1051) drive expression below basal controls ( $p < 0.05$ , Wilcoxon-rank sum test with Bonferroni correction,  $n = 1764$ ). Activity effect sizes (Figure 1A) are consistent with previous MPRAs in GM12878, with a notable enrichment for sequences that drive expression below basal controls, which has been observed previously in MPRAs in other cell types (Kwasnieski et al. 2014; Tewhey et al. 2016; White et al. 2013). For the segmentation predictions, regions annotated as WE and E have the highest average activity compared to Repressed and Promoter Flanking annotated regions (Figure 1B; Tukey ANOVA analysis). This result is consistent with the higher activity previously observed for sequences annotated as Weak Enhancers tested in K562 cells (Kwasnieski et al. 2014). We observed no significant difference in the average expression levels between tested REF library sequences annotated as R and PF (Figure 1B; Tukey ANOVA analysis). However, within each annotation group there are sequences that drive expression across the whole range of observed activity, including sequences predicted to be repressive that drive expression more than 4-fold over and 4-fold under basal levels. Therefore, chromatin-based functional annotations alone only explains a small portion of the differences in regulatory activities of genomic sequences.

We next asked whether sequences predicted to contribute to fitness add value to the functional annotations. We did not observe a quantitative correlation between fitCons scores and the activities of genomic sequences measured in our REF library (Figure 2A;  $R^2 = 0.01$ ). However, the *rho* scores from fitCons represent the probability that a sequence contributes to fitness and are not quantitative measures of how much sequences might contribute to fitness. After binning sequence with low ( $rho < 0.162$ ) and high scores ( $rho > 0.162$ ), we did find a significant difference between average sequence activity (Figure 2B; Student's t-test,  $p = 7.68e-05$ ), but high

scoring sequences were not any more enriched for sequences with significant activity than low scoring sequences (low *rho*: 522 significant out of 879, high *rho*: 529 significant out of 880; Chi-square test,  $\chi^2(1, N = 1759) = 0.0639$ ;  $p = 0.8003$ ). Our results suggest that sequences predicted to be under selection are not necessarily more likely to have regulatory potential, but do show greater levels of *cis*-regulatory activity.

### **Fitness or deleteriousness predictions do not identify variants that drive allelic skew**

We next asked whether genomic sequences with signatures of selection are more likely to contain genetic variants that affect *cis*-regulation by testing sequences that were all annotated as WE and overlapped a variant in the 1000 genome project database. Our expectation for the VAR library was that regardless of the activity level of the sequences, genetic variants in sequences under selection would be more likely to cause a detectable difference in *cis*-regulatory activity compared to the reference allele. Overall, we detect significant activity for 65.6% (454/692) of the entire VAR library, regardless of allele status, of which 32.6% (148/454) drive expression above basal controls and 67.4% (306/454) drive expression below basal controls ( $p < 0.05$ , Wilcoxon-rank sum test with Bonferroni correction,  $n = 691$ ). We detected significant allelic skew, the log ratio of alternative variant to reference allele expression, for 45% of alternative-reference pairs tested in the VAR library (Figure 3A;  $p < 0.05$ , Student's t-test with FDR  $< 0.05$  correction;  $n = 346$ ). The majority of allelic differences are of modest effect sizes (Figure 3B). The large fraction of tested alternative-reference pairs with detectable allelic skew shows that variation in human populations can frequently impact the relative activity of a putative *cis*-regulatory sequence, but rarely cause large changes.

All regions tested in the VAR library are annotated as WE and although alternative alleles could be enhance or reduce the activity of a putative CRS, since fitCons and CADD claim to integrate regulatory annotations and variation metrics, we expected that the impact of tested variants might be similarly predicted by the two fitness predictions. For fitCons, we found that the magnitude of observed allelic skew does not correlate with the fitCons score for each parent sequence, again

unsurprising as fitCons scores are probabilities ( $R^2 = 0.001$ ; Figure 4A). We then binned sequences into ‘low’ and ‘high’ groups based on fitCons scores, with low ( $\rho < 0.158$ ) and high scores ( $\rho > 0.158$ ), and to determine if regions with higher predicted probabilities of contributing to fitness enriched for variants that show allelic skew. However, we found no significant difference between average allelic effect sizes for low and high scored sequences (Figure 4B; Wilcoxon rank-sum test,  $p = 0.1028$ ), or in the number of sequences with significant allelic skew (Pearson’s Chi-square test;  $\chi^2(1, N = 346) = 0.15179$ ;  $p = 0.6968$ ). The independently generated CADD scores, which according to the authors provide a score of deleteriousness, or negative fitness impacts were less evenly distributed for the tested substitutions. (Kircher et al. 2014). When we compared the precompiled CADD C-scores for each tested variant, we also found poor correlation with the observed magnitude of allelic skew ( $R^2 = 0.005$ ; Figure 4C). When we binned CADD scores into high (C-score  $> 4.6$ ) and low (C-scores  $< 4.6$ ) based on the average CADD score tested in the VAR library, we did not find that there was a significant difference between average allelic effect sizes (Figure 4D; Wilcoxon rank-sum test;  $p = 0.1028$ ). Binned CADD scored sequences are also not enriched for significant allelic skew (high C-score: 97 with significant allelic skew of 219, low C-score: 61 with significant allelic skew of 127; Pearson’s Chi-square test,  $\chi^2(1, N = 346) = 0.45298$ ;  $p = 0.5027$ ). Together, these data show that signatures of natural selection, as measured by two independent predictors, do not help predict the expected regulatory impact of variants observed in human populations.

Population genetics predicts that rare variants may have larger impacts on phenotypes than common variants. If there is a *cis*-regulatory contribution to the rare variant-common disease model (Saint Pierre and Génin 2014; Gibson 2012; Schork et al. 2009), then rare variants, even those observed in human populations, might be more likely to drive allelic skew. In our data, there is not a quantitative correlation between allelic effect size and minor allele frequency (MAF). Although we observe what appears to be a trend towards larger effect sizes for rarer variants (Figure 3D) this trend does not pass the threshold for significance testing. We found no

significant difference between rare alleles (MAF < 0.01) and common alleles (MAF > 0.05) for either the average allelic effect size (Wilcoxon rank-sum test;  $p = 0.4865$ ) or the number of sequences with significant allelic skew (Chi-square test;  $\chi^2(1, N = 288) = 0.81923$ ,  $p = 0.365$ ). This result suggests that allele frequency alone is not a robust predictor of *cis*-regulatory impact.

## Discussion

In this study, we sought to determine the value of fitness predictions for identifying genomic sequences with regulatory potential in the lymphoblastoid cell line GM12878 and prioritizing variants that might impact *cis*-regulatory activity. By testing sequences classified as potentially regulatory in GM12878, we found that sequences with high probabilities of contributing to fitness as predicted by fitCons had higher average activity in our assay. The predictive power of fitCons scores for genomic sequences indicates that short noncoding sequences might be selected for high or low *cis*-regulatory activity. When we tested variants in regions annotated as WE, our method was sensitive enough to detect differences in expression for matched reference-alternative alleles, although these differences were relatively modest. The percentage of tested regions with significant allelic skew is slightly higher than similar studies, likely due to differences in normalization processes, sequencing depth, and minimal promoter choice, but the modest effect sizes measured for single nucleotide substitutions are consistent (Ulirsch et al. 2016; Tewhey et al. 2016). However, we did not find a significant difference in allelic skews between sequences with low and high probabilities of contributing to fitness as predicted by fitCons nor did we find a significant difference for variants with high and low scores of deleteriousness as predicted by CADD. These results are in contrast to our prediction that sensitivity to substitutions may have a greater impact on organismal level fitness than activity level, but are consistent with recent work that concluded that CADD has poor diagnostic value for identifying pathogenic variants for clinical noncoding cancer variants (Mather et al. 2016).

A shared goal for developing fitCons and CADD was to predict the impact of substitutions genome-wide. While we observed a small but significant difference between high and low fitCons scores for general activity, neither CADD nor fitCons scores align with the frequency or magnitude of measured allelic expression differences. This suggests that these algorithms are predicting something besides fitness and deleteriousness, respectively. Alternatively this suggests that that potential regulatory sequences tolerate substitutions observed in human populations just as well as sequences that are unlikely to contribute to fitness. A major assumption in the field is that differences in regulatory activity contribute to phenotypic differences (Cookson et al. 2009; Raghavan et al. 2014; Mohammadi et al. 2017; Myint et al. 2018). If most cell types or most gene networks are robust to minor expression changes, then it is possible that only variants that impact a particularly sensitive gene or network will have any impact on fitness or disease risk. This prediction is consistent with recent work to determine if GWAS risk and eQTL association are driven by the same underlying variant, which found that although the noncoding loci overlap, only 21% of disease associations show a statistical shared effect with a local eQTL (Chun et al. 2017). As both the study by Chun and colleagues and our study used immune-related cell types, the smaller impact of regulatory variants could also be a feature of this lineage.

An unexpected result was the proportion of sequences that appear to act repressively. While repression should be equally well modeled in predictions of fitness and deleteriousness, MPRAs are strongly skewed for measuring activation if very weak minimal promoters, such as TATA-box only promoters are used (White et al. 2016; Tewhey et al. 2016). The fraction of sequences measured as repressive in our study, thanks to a stronger minimal promoter and deep sequencing of the libraries and replicates (See Methods), is consistent with the fraction of the genome expected to be repressive by segmentation predictions (Supplemental Figure S5), histone marks alone as supported by ENCODE data, and observations from development, including lymphoblastoid differentiation (ENCODE Project Consortium 2012; Thiel, Lietz, and Hohl 2004; Reynolds, O’Shaughnessy, and Hendrich 2013; van Keimpema et al. 2015). If *cis*-regulatory activity does contribute to fitness and disease risk, our ability to identify key

regions and variants of interest may be hampered without more sensitive approaches to enrich for and identify repressive elements from the genome, as many of the sequences that showed repressive activity in the VAR and REF libraries may have been missed without sufficiently deep sequencing.

These results suggest that the link between fitness and how we test for expression changes may need to be re-evaluated as two independent algorithms that aim to predict fitness consequences fail to predict the allelic skew as measured using plasmid reporters, despite MPRA being used as a widespread tool to evaluate the impact of variation from the human genome (Ulirsch et al. 2016; Tewhey et al. 2016; Y. Li et al., n.d.). Additionally, the large fraction of sequences that drive expression below basal activity suggests that our understanding of the function of the genome could be enhanced by developing methods with better sensitivity to detect sequences that repress expression. Moving forward, testing a sufficient number of *de novo* mutations or very rare mutations associated with a disease or phenotype would help determine if the pattern of small effect sizes we observe for 1000 Genomes variants generalize to most substitutions or not. In addition, more examples of causal noncoding alleles in enhancers and repressors will need to be identified if we are to reliably predict the impacts of noncoding variants on *cis*-regulatory activity.

## **Methods**

### **Library design**

First, genomic regions with uniform fitCons scores across the entire lengths (approximately 14 million regions) were filtered to exclude blacklisted and poorly mapped regions (approximately 13.5 million remaining regions) (See Supplemental Table 4) (Derrien et al. 2012). Regions were then intersected with putative regulatory regions (Combined ChromHMM/Segway models, Supplemental Table 4), such that the entire genome region has a consistent fitCons score and



single Combined annotation. Regions were then subsetted to only blocks of length 80-130 bp (approximately 8.8 million regions) due to limitations in library size oligo synthesis. The regions were subsetted further to be restricted to likely regulatory regions, Combined annotations of Weak Enhancers, Enhancers, Repressors, and Proximal Promoters (approximately 530,000 regions). These were then assessed to determine relative distributions of fitCons scores (Supplementary Figure S5). Each annotation group was then randomly sampled between *rho* scores of 0.00 and 0.35, with all regions with scores between 0.40 and 0.50, the maximum values observed for 80-130 bp sequences, were retained (n= 28), to better capture the entire distribution of fitCons scores for noncoding regions. After filtering for restriction sites, the remaining sequences were used to design oligos for the REF library.

Regions annotated as Weak Enhancers (WE) were then used as query regions for the Tabix tool (H. Li 2011) to retrieve variants from the The International Genome Sample Resource (IGSR) 1000 Genome Project database (L. Clarke et al. 2017; 1000 Genomes Project Consortium et al. 2015). As not all regions selected for the REF library overlapped variants, additional WE annotated regions were randomly selected with scores between *rho* scores of 0.00 and 0.30 and then used as query regions to retrieve additional variants from the database. After filtering for restriction sites, the remaining 346 variant and alternative allele pairs were used to design oligos for the VAR library.

### **Library cloning**

To generate libraries for our regions of interest, we ordered a custom pool of 15,000 unique 200 bp oligonucleotides (oligos) from Agilent Technologies (Santa Clara, CA) through a limited licensing agreement. Each oligo in the REF library pool was 200 bp in length with the following sequence:

```
GATACATGGTCAGCTAGCGT[SEQ]AAGCTT[FILLER]CTCGAGGCATGCC[BC]TGAGC  
TCTACATTCGCATACTG
```

Where [SEQ] is a 80-130 bp reference sequence (Reference assembly GRCh37 (hg19)) annotated as a putative regulatory region, [FILLER] is a random filler sequence of variable length to bring the total length of each sequence to 200 bp, and [BC] is a unique 9 bp barcode sequence. Each unique [SEQ] in the REF library was assigned six different [BC] sequences.

Each oligo in the VARS library pool was 200 bp in length with the following sequence:

```
GACTTGACATGTCTGCTAGCA[SEQ]AAGCTT[FILLER]CTCGAGGCATGCC[BC]TGAG  
CTCTGAACAGTACGATC
```

Where [SEQ] is either a 80-130 bp ‘Weak Enhancer’-annotated reference sequence (Reference assembly GRCh37 (hg19)) or a matched sequence with a single nucleotide substitution from the IGSR 1000 Genome Project database, [FILLER] is a random filler sequence of variable length to bring the total length of each sequence to 200 bp, and [BC] is a unique 9 bp barcode sequence. Each unique [SEQ] in the VARS library was assigned six different [BC] sequences.

The plasmid library was prepared as described previously (Kwasnieski et al. 2014, [a] 2012), except using primers RefLib\_FW and RefLib\_Rev for the REF library with an annealing temperature of 53°C and VarLib\_FW and VarLib\_Rev for the VAR library with an annealing temperature of 56°C (Supplemental Table 1). Amplified sequences were cloned into the NheI and SacI sites of pGL-hsp68 and inserted into pGL-hsp68 (Kwasnieski et al. 2014) digested with the same enzymes. These plasmid pools were then isolated and digested with HindIII and SphI (New England Biolabs, Ipswich, MA). A minimal *hsp68* promoter and *dsRed* reporter gene was then amplified from pGL-hsp68 using primers FW\_hsp68\_dsRed and Rev\_hsp68\_dsRed (Supplemental Table 1), and inserted into the plasmid library pools from the previous step at the HindIII and SphI sites. The final libraries then each have putative regulatory elements upstream

of the *hsp68* minimal promoter and *dsRed* reporter gene, with a 9 bp barcode sequence in the 3' UTR of the reporter gene.

### **Cell culture and transfection**

GM12878 cell line was maintained in standard conditions (Roswell Park Memorial Institute (RPMI) 1640 media plus 15% fetal bovine serum and 10 U/mL Penicillin-Streptomycin), as described by ENCODE protocols (ENCODE Project Consortium 2012). The plasmid library was purified by phenol-chloroform extraction followed by ethanol precipitation prior to transfections. The NEON transfection system (Life Technologies, Carlsbad, CA) was used to transfect approximately  $5 \times 10^6$  cells per replicate using two 30-msec pulses at 1100V with 18 ug of each library plasmid pool plus 2 ug of pMAX-GFP as a positive control. Transfected cells were seeded into T-25 flasks with 5 mLs of pre-warmed media lacking antibiotics, as directed by the manufacturer. Three replicates were transfected with the REF library and four replicates were transfected with the VAR library. Transfection efficiencies were estimated at 30% based on GFP signal from control plasmid in transfected cells (data not shown).

### **MPRA**

After 24 hours, approximately  $3-4 \times 10^6$  cells were harvested for RNA using the PureLink RNA mini kit (Life Technologies, Carlsbad, CA), 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 pools were amplified using primers CF150 and BCenrichment\_Rev with a 60°C annealing temperature for 12 cycles. Barcode enrichment PCR products were then digested with XbaI and SphI and then ligated with adaptors PE2\_Ind71\_XbaI and P1\_SphI\_X (where X is 1-4) for the REF library and adaptor PE2\_Ind72\_XbaI and P1\_SphI\_X (where X is 1-5) for the VAR library. An enrichment PCR with primers CF150 and BCenrichment\_Rev was then used (Supplemental Table 1), and the resulting products were mixed at equal concentration and sequenced on one Illumina (San Diego, CA) NextSeq lane.

Sequencing reads were filtered to ensure that the BC sequence perfectly matched the expected sequence. For the REF library, this resulted in 54.3 million total reads combined for the three demultiplexed RNA samples (PE\_Ind71\_XbaI/P1\_SphI\_1-3 adaptors; 16.6-19.2 million each), and 18.3 million reads for the DNA library sample (PE\_Ind71\_XbaI/P1\_SphI\_4 adaptors). For the VAR library, this resulted in approximately 65.9 million reads combined for the four demultiplexed RNA samples (PE\_Ind72\_XbaI/P1\_SphI\_1-4 adaptors; 15.6-16.8 million each), and 23.2 million reads for the DNA library sample (PE\_Ind72\_XbaI/P1\_SphI\_5 adaptors). For each library, BCs that had less than 10 raw counts in any RNA replicate or less than 10 raw counts in the DNA sample for the REF library and less than 30 raw counts in any RNA or less than 30 raw counts in the DNA sample for the VAR library were removed before proceeding with downstream analyses.

Expression normalization was performed by first calculating reads per million (RPM) per BC for each RNA replicate for both the REF library ( $R^2 = 0.994-0.996$ ; Figure S1A) and the VAR library ( $R^2 = 0.972-0.988$ ; Figure S1B). 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 removes 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 REF library ( $R^2 = 0.0003-0.0005$ ; Figure S2A) and the VAR library ( $R^2 = 0.0001-0.0010$ ; Figure S2B). Within each biological replicate the BCs corresponding to each unique putative regulatory sequences were averaged and then normalized by basal mean expression in that replicate. These normalized expression values were then averaged across biological replicates. Expression summaries per replicate are reported in Supplemental File 1 for the REF library and Supplemental File 2 for the VARS library.

### **Other analysis and data sources**

The position of each alternative allele in the VAR library were annotated with CADD scores (v1.3) using the CADD server (<https://cadd.gs.washington.edu/score>). Genome coordinate intersections and merges were performed with the bedtools suite version 2.27. All downstream analyses were performed in R version 3.3.3 and plotted with ggplot2 version 2.2.1.

### **Data Access**

Raw data will be uploaded to SRA and processed data will be available through the GEO database prior to submission for publication.

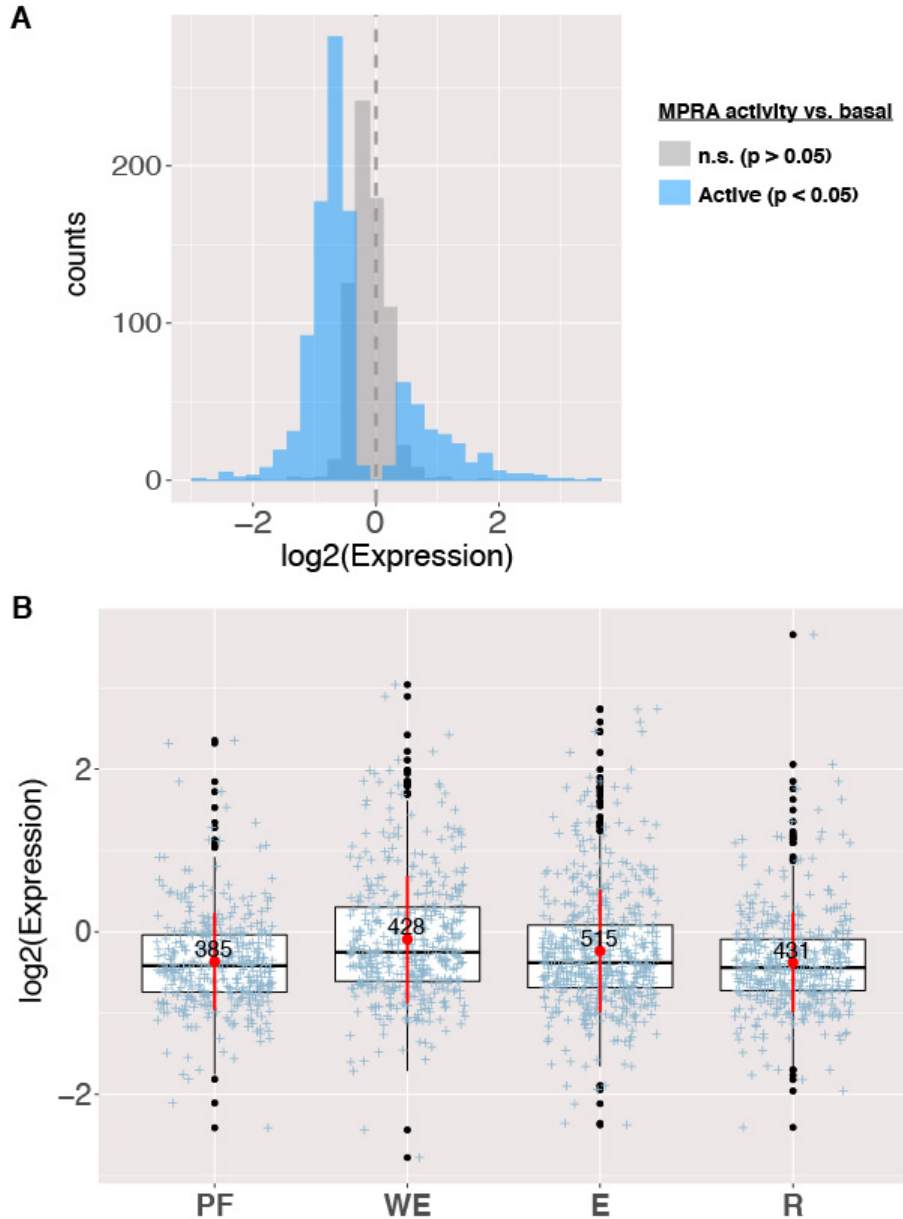
### **Disclosure Declaration**

The authors declare that they have no competing interests to disclose.

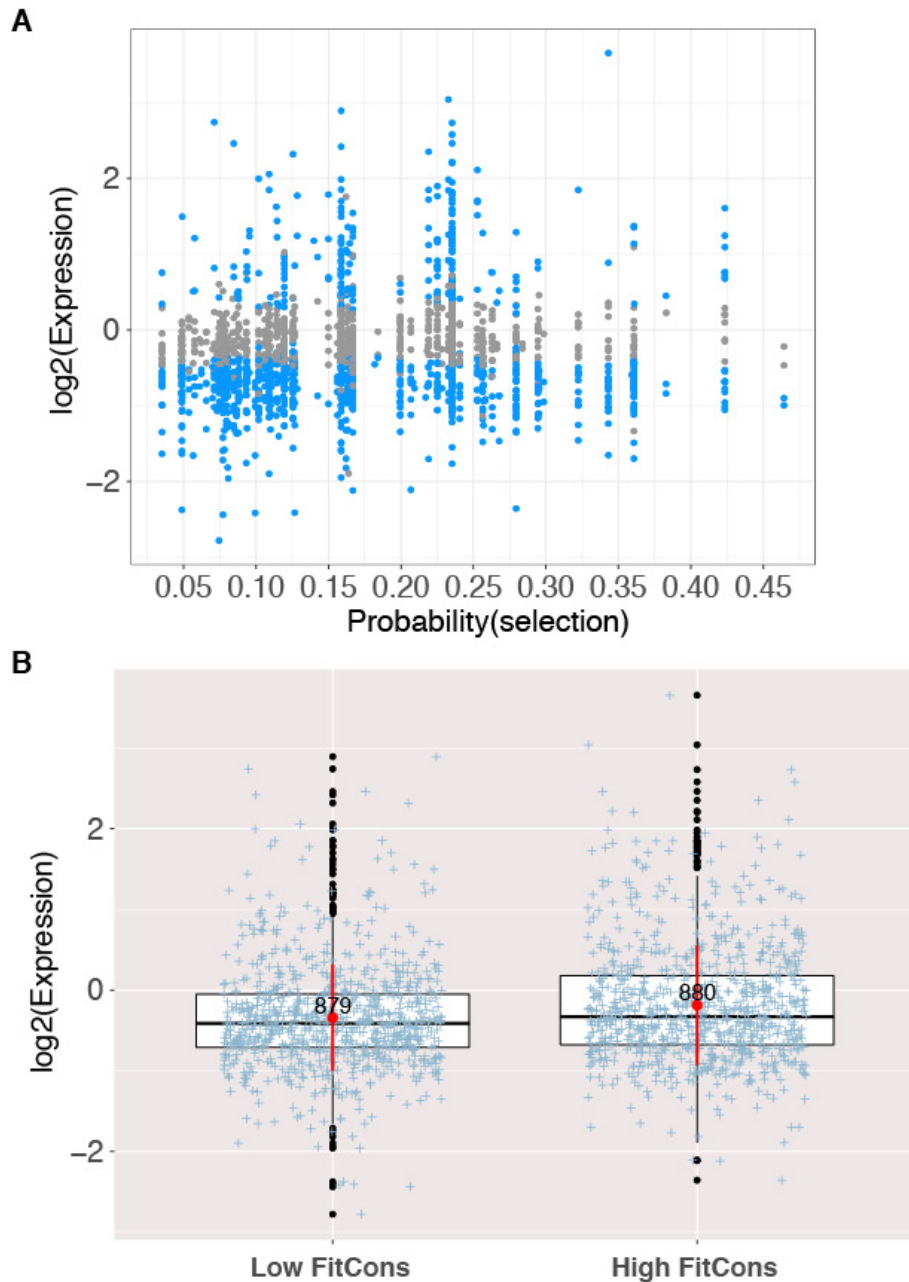
### **Acknowledgements**

We thank members of the Cohen Lab for critical reading and feedback, Jessica Hoisington-Lopez from the DNA Sequencing Innovation Lab for assistance with high-throughput sequencing, as well as Brad Gulko and Adam Siepel for their assistance with interpreting fitCons scores. This work is supported by a grant from the National Institutes of Health, R01 GM092910 to B.A.C.

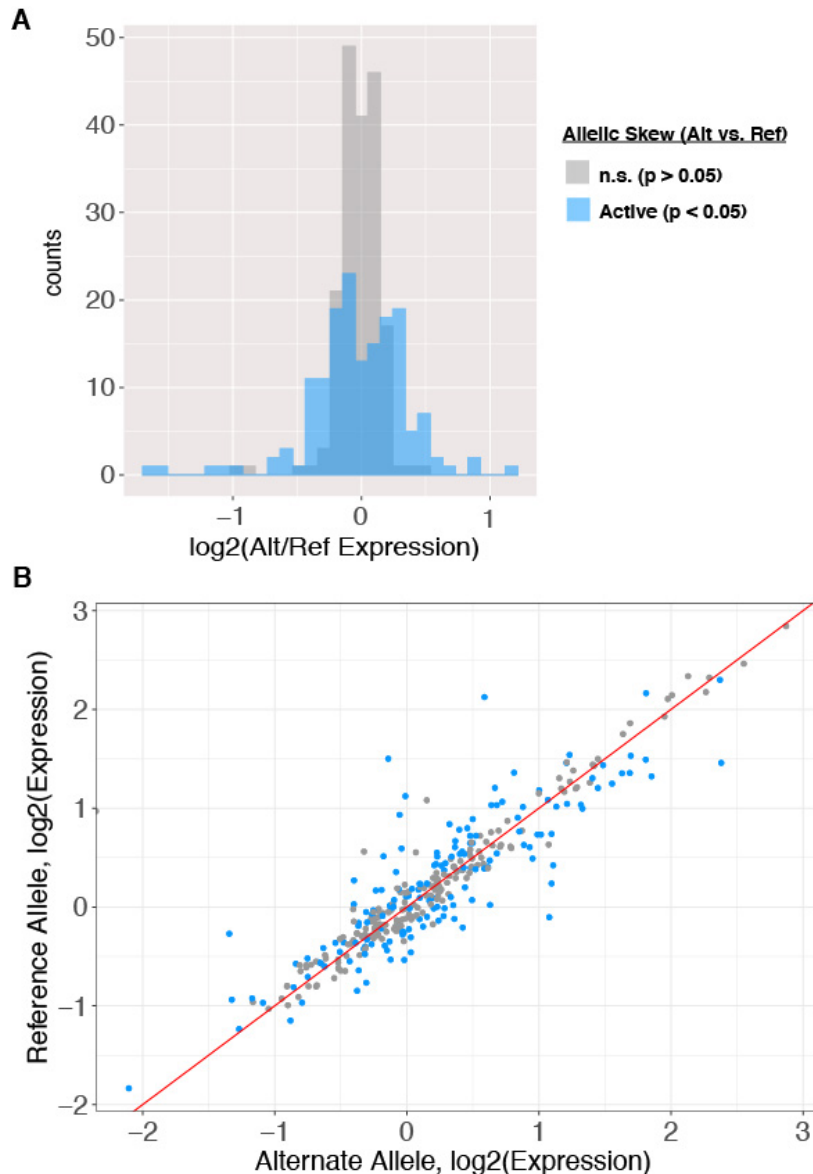
## Figures



**Figure 3.1. Activity of reference sequences in GM12878.** (A) Distribution of activity for REF library sequences that are significantly different from basal controls (blue) and that failed to reach significance (grey);  $p < 0.05$ , Wilcoxon-rank sum test with Bonferroni correction,  $n = 1764$ . (B) Distribution of normalized expression of REF library sequences for Predicted Flanking Promoter (PF), Weak Enhancer (WE), Enhancer (E), and Repressed (R) Combined segmentations. Boxplots overlaid with mean (red dot) and SEM (red lines) for tested sequences, with expression average across barcodes per sequence shown with blue (+) symbol.

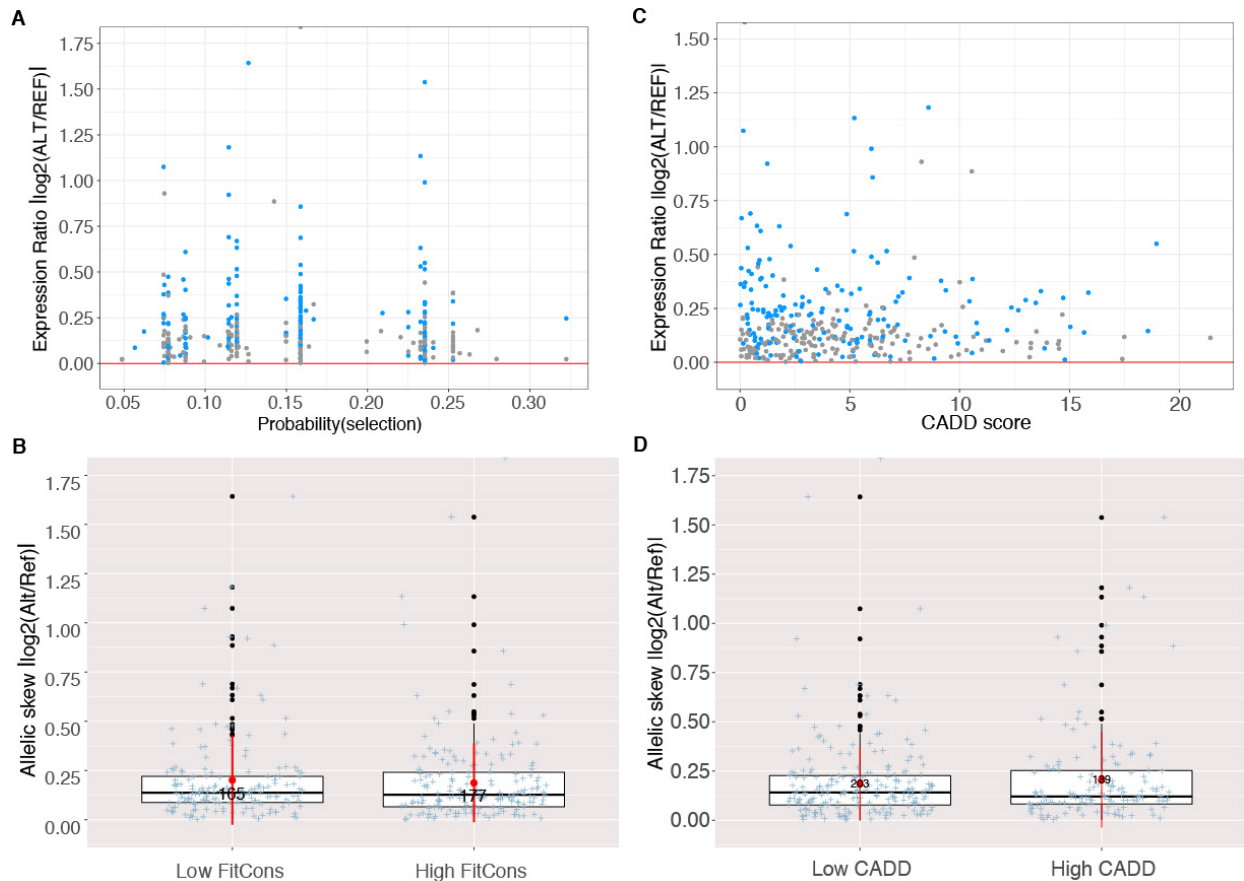


**Figure 3.2. Activity of reference sequences versus probability of selection.** (A) Normalized expression of REF library sequences versus fitCons score of probability of selection. Sequences with expression significantly different from basal controls shown in blue (Wilcoxon rank sum test,  $p < 0.05$ ; Bonferroni correction,  $n = 1763$ ). (B) Expression of REF sequences for low ( $\rho < 0.162$ ) or high ( $\rho > 0.162$ ) fitCons scores. Boxplots overlaid with mean and SEM (red) plus data points (blue) as in Figure 1.

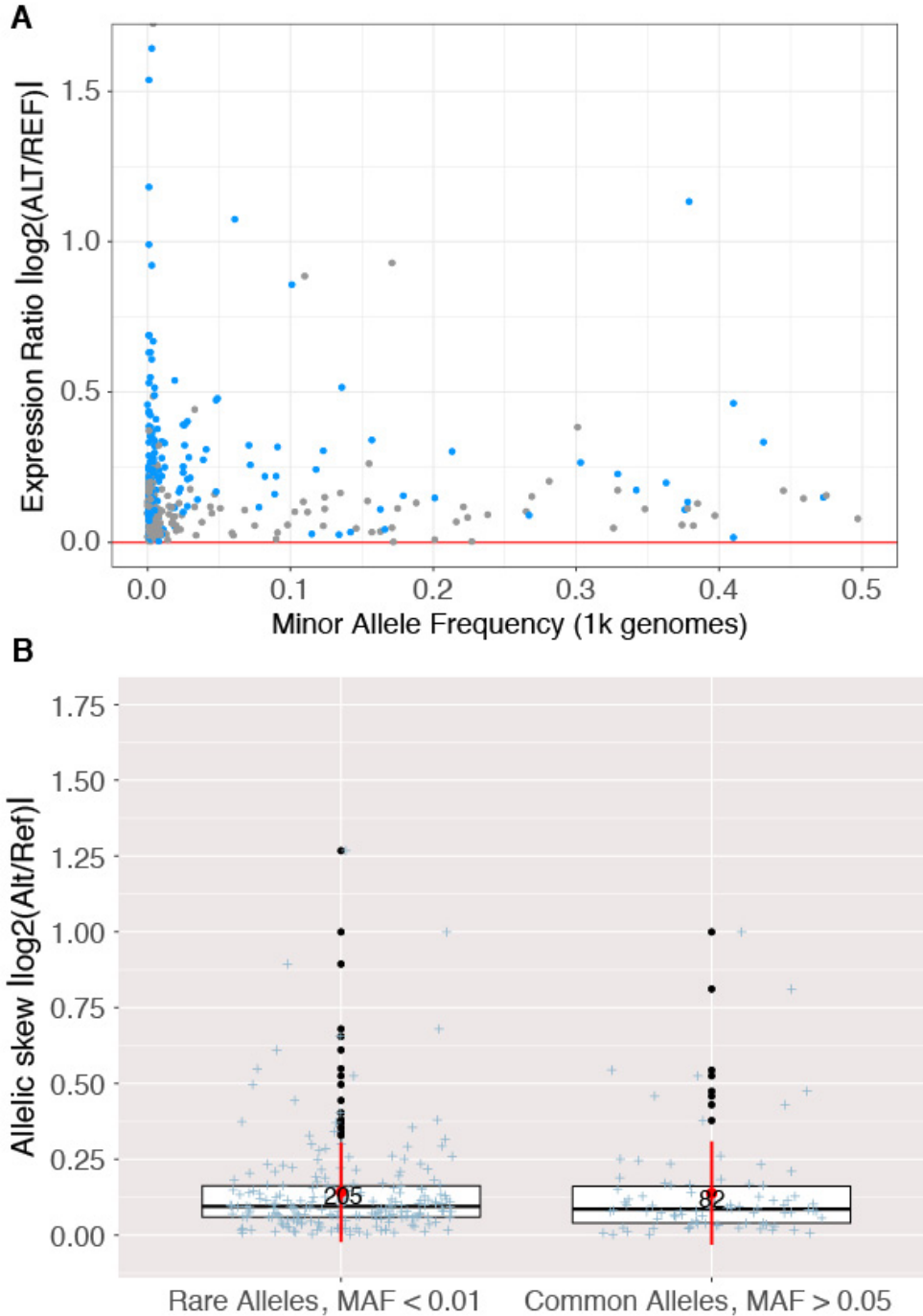


**Figure 3.3. Allelic skew of 1000 genome variants in GM12878.** (A) Normalized expression of reference and single substitution alternate alleles. Blue indicates sequences with significantly different expression between matched alleles (Student's t-test,  $p < 0.05$ ; B.H. correction, FDR < 5%). Each point is the average of multiple barcodes across replicates, so larger average changes that fail to meet significance can be observed if there is large variance in measurements. Diagonal red line is expectation if substitution had no impact on expression level. Both alleles have been normalized to basal controls (Basal expression = 1; not shown). (B) Distribution of allelic skew,  $\log_2(\text{Alternative/Reference expression})$ , for VAR library alleles. Pairs with significant skew are plotted in blue and sequences that failed to reach significance shown in grey;  $p < 0.05$ , Wilcoxon-rank sum test with B.H. correction, FDR < 5%,  $n = 346$ .





**Figure 3.4. Allelic skew versus probability of selection and deleteriousness score.** (A) Allelic skew, absolute value of  $\log_2(\text{Alternative}/\text{Reference expression})$  versus fitCons score of probability of selection. Blue indicates sequences with significantly different expression between matched alleles (Student's t-test,  $p < 0.05$ ; B.H. correction, FDR  $< 5\%$ ). (B) Allelic skew of VAR sequences for low ( $\rho < 0.158$ ) or high ( $\rho > 0.158$ ) fitCons scores. Boxplots overlaid with mean and SEM (red) plus data points (blue) as in Figure 1. (C) Allelic skew versus CADD score of deleteriousness. Blue indicates significance as in A. (D) Allelic skew of VAR sequences for low (C-score  $< 4.6$ ) or high (C-score  $> 4.6$ ) CADD scores. Boxplots overlaid with mean and SEM plus data points as in (B).



**Figure 3.5. Allelic skew versus allele frequency.** (A) Allelic skew of VAR sequences versus minor allele frequency of alternative allele as reported by the 1000 Genome database. Blue indicates sequences with significantly different expression between matched alleles (Student's t-test,  $p < 0.05$ ; B.H. correction,  $\text{FDR} < 5\%$ ). (B) Allelic skew of VAR sequences with rare (MAF < 1%) or common (MAF > 5%) alternative alleles. Boxplots overlaid with mean and SEM (red) plus data points (blue) as in Figure 1.

## **Chapter 4: Conclusion and Future Directions**

My thesis work has focused on understanding DNA sequence contributions to *cis*-regulatory activity in two species, human and mouse. The first part of my thesis, Chapter 2, was primarily focused at the resolution of transcription factor binding sites (TFBS), to understand the grammar of interactions between pluripotency TFs for synthetic and genomically occurring configurations of TFBS in mouse embryonic stem cells (mESCs). Using expression driven in massively parallel reporter assays (MPRAs), I found that the grammar learned from synthetic combinations of pluripotency TFBSs was not a strong driver of expression measured for genomic configurations of TFBSs for the same transcription factors (TFs). The second part of my thesis, Chapter 3, was primarily focused at the resolution of single nucleotide variation observed in human populations, to understand if predictions of fitness contributions can help prioritize noncoding variation that impacts *cis*-regulatory activity. Using MPRAs to test putative regulatory regions and variants from the human genome, I found that predicted fitness better distinguished the activity of putative *cis*-regulatory sequences (CRSs) than the impact of substitutions. These findings have broad implications for the study of *cis*-regulation in mammals, both in the context of synergistic regulation by TFs and for understanding the impacts of variation on regulatory activity.

### **Synthetic and Genomic grammar of pluripotency factors**

I have extended our usage of MPRAs to directly compare the expression of combinations of TFBSs in a controlled, synthetic context and more a complex genomic context. Previous work has been generally limited to either genomic or synthetic CRSs, resulting in either limited power to detect higher order interactions between transcription factors in the case of genomic only studies or uncertainty in how much higher order interactions learned from synthetic sequences contribute to expression when spacing, affinity, and other variables are no longer fixed

(Chaudhari and Cohen 2018; Kwasnieski et al. 2014; Fiore and Cohen 2016; Gertz, Siggia, and Cohen 2009; White et al. 2013). In Chapter 2, I designed synthetic elements to specifically test grammar that were complimented by genomic sequences selected to have comparable configurations of TFBSs and also mutated to demonstrate that activity is dependent on one or more of the pluripotency TFBSs. This design study plus quantitative modeling allowed me to definitively show that, at least in the context of mouse pluripotency factors, the features that are fixed in synthetic contexts, such as TFBS affinity and spacing, play a larger role in the activity of genomic sequences than the synthetic-learned grammar. The comparison between synthetic elements and genomic sequences in mESCs represents a significant contribution to identifying active CRSs in mESCs and informing the future design of synthetic CRSs.

The contributions of TFBS affinity and spacing to classifying active tested genomic sequences in Chapter 2 could help researchers determine what intergenic regions of mammalian genomes are functional, as my work suggests that clusters of high affinity TFBSs are more likely to drive regulatory activity than other sequences with ChIP-seq signals. This is particularly important as we and others have shown that occupancy by a given TF as measured by ChIP-seq does not mean that a given sequence will have regulatory activity, with only 20% to 30% of sequences with TF occupancy in the genome driving significant activity in MPRA (Maricque, Dougherty, and Cohen 2017; White et al. 2013; Chaudhari and Cohen 2018). A higher rate of validation for potential CRSs falling under ChIP-seq peaks could lead to a better understanding of how to distinguish between spurious binding events and true targets for TFs of interest. To build on the patterns of expression for the small sample of genomic sequences tested that best matched the combinations of no more than one of each TFBS, additional sequence space should be explored by testing additional genomic sequences that include no more of one of each pluripotency TFBS underlying Nanog and p300 ChIP-seq peaks to better resolve the sequence determinants of activity for regions regulated by pluripotency factors in mESCs (C.-Y. Chen, Morris, and Mitchell 2012; Bailey and Machanick 2012). For the sequences previous tested plus any additional genomic sequences should be more extensively mutated, with each pluripotency TFBS

mutated individually, as well as pairs, to compliment the complete mutation design used in Chapter 2. Additionally, more extensive manipulations of genomic sequences could test the grammar rules learned through modeling, including shuffling the sequences to determine if other sites could be contributing to activity, modifying the affinities of and spacing between identified TFBSs. This approach would extensively tease apart the sequence features necessary to drive expression in mESCs, adding to the body of work examining TFs networks that are necessary for the pluripotent state (Ferraris et al. 2011; Niwa 2014; Loh et al. 2006).

The lack of evidence for shared grammar between synthetic and genomic CRSs also suggests that future synthetic CRSs for multiple TFs should include variable spacing and affinity to better represent the complexity of genomic sequences, while maintaining statistical power by prioritize the sequence features that I found to contribute to genomic activity. Previous studies have used high affinity TFBSs with fixed spacing for multiple factors (Fiore and Cohen 2016; Mogno, Kwasnieski, and Cohen 2013) or two to three different affinity binding sites for no more than two factors (White et al. 2016). My work in Chapter 2 tested high affinity sites for four different factors, exhausting all possible combinations of no more than one of each TFBS per synthetic element, which was successfully modeled by a grammar that included position. However, the grammar model for the synthetic library failed to predict the activity of genomic sequences. To address this, improvements to synthetic library design could be made. For the pluripotency factors I tested, the possible sequence space could be further explored while maintaining the advantage that synthetic elements provide for exhausting design. Future synthetic libraries should at minimum modify site affinities and include multiple spacings between TFBSs to better represent the observed patterns of TFBSs in mammalian genomes.

### **Identifying regulatory variants using fitness predictions**

I have extended the use of MPRA to test genome-wide segmentation and fitness predictions in a lymphoblastoid cell line. By testing putative CRS from the human genome in Chapter 3, I show

that in GM12878, as in the K562 cell line, specific chromatin-based functional segmentations have variable levels of activity when tested in the same cell type as the segmentations were generated, with ‘weak enhancers’ driving the highest average activity in the MPRA (Kwasnieski et al. 2014). Segmentation predictions aim to classify the human genome into specific functional units based on histone modification measurements, but appear to fail to predict the likelihood of regulatory activity despite assignments of regulatory function (Kwasnieski et al. 2014; Ernst and Kellis 2012; Hoffman et al. 2011, 2013). This was especially surprising when I compared sequences labeled by ‘Repressed’ segmentations to ‘Flanking Promoters’ segmentations, as the average activity was not significantly different. The lack of evidence of regulatory potential for sequences with more ‘active’ histone modifications (Lelli, Slattery, and Mann 2012; Jenuwein and Allis 2001), suggests that these functional classifications would need to be refined to effectively identify regions of the human genome that are likely regulatory regions.

In Chapter 3, I show that there is a modest difference in expression between low and high fitCons scores that were specifically generated for the GM12878 cell line. This difference in regulatory activity could be due to actual selection on activating regulatory sequences in the human genome, or could be a result the fitCons scoring process which includes using functional information to bin sequences with similar biochemical markers prior to sequence conservation across species and diversity within populations (Gulko et al. 2015). To further develop the manuscript presented in Chapter 3, comparing the activity of the tested reference sequence CRSs to conservation scores alone and the individual biochemical markers used to classify sequences would help untangle if this signal is due to the end fitness predictions or the underlying data that is incorporated into those predictions. Further testing to control for possible confounding factors such as conservation and signals for regulatory activity such as ChIP-seq, would also help inform our expectations in regards to the relationship evolution and *cis*-regulatory activity.

By testing a relatively unbiased set of variants observed in human populations, I also showed that neither allelic expression frequency or effect sizes are correlated with predictions of regional

fitness consequence or individual variant deleteriousness. Recent work tested genome sequences with multiple variants using MPRA in GM12878, but was restricted only to variants associated with expression differences between individuals making these data difficult to extend to more general questions about the expected impact of observed variation on regulatory activity (Tewhey et al. 2016). Testing variants from the 1000 Genomes Project database for regulatory impacts without filtering on any potential expression or disease associations allows us to generate a baseline expectation for the expected frequency and effect size for variation observed in noncoding regions. Although our sample size of tested reference and alternative alleles ( $n = 346$ ) is too small to be called a null distribution, the observation that 45% of variants observed in human populations impact *cis*-regulatory activity suggests that a large proportion of noncoding variants could potentially have impacts but the relatively small effect sizes observed suggest that most variation would have small impacts on nearby gene expression. Larger allelic skews might be observed for more recent or *de novo* variation that we would expect to potentially have more extreme *cis*-regulatory impacts. Drawing from disease linked variants, *de novo* variants such as from cancer panels, as well as additional 1000 Genomes Project variants would provide a broader representation, especially if we can better prioritize regions of interest using the activity patterns and annotations of the larger reference only library tested in Chapter 3. Testing a larger and more diverse sample of variants could lead to a better understanding of the landscape of potential *cis*-regulatory impacts of variation in human genomes.

Although fitness predictions do not appear to help identify variants that will drive allelic skew, there could be several possible explanations for this observation. Minor expression changes likely propagate through gene networks, and thus, small changes may be dissipated. The discrepancy between fitness predictions and measured allelic impacts could perhaps be explained if fitness is only impacted when the affected genes are hubs within a gene regulatory network. This is consistent with network theory claims that complex networks demonstrate a surprising degree of error tolerance due to redundancy (Albert, Jeong, and Barabasi 2000). If this is true, fitness predictions for noncoding regions must account for these network sensitivities. Although

fitCons does attempt to account for differences between genomic region using annotation data to group similar regions prior to determining conservation over evolutionary time and diversity within human populations, and CADD uses ENCODE and UCSC genome browser data to annotate observed and simulated variants (Kircher et al. 2014; Gulko et al. 2015), it is unclear if network sensitivity should be better taken into account. Although network sensitivity is a difficult area of research, tools such as CRISPR allow genome manipulations that could address this question by swapping variants in the regulatory regions of central regulators for a cell type versus peripheral gene network member. Additionally, the discrepancy between fitness predictions and measured allelic impacts could be explained by the possibility that many variants that have an impact but low expected fitness impact could affect processes that would be consequential later in life, such as heart disease or type-2 diabetes, such that reproduction would likely not be impacted. In cases where variants impact processes later in life, GWAS studies and MPRA might be more useful than predictions of fitness. An alternative approach to addressing the question of fitness in relation to expected allelic impacts would be to test the libraries generated in Chapter 3 in cell lines of other species, such as one or more chimp lymphoblastoid cell line, as developed by Romero and colleagues (Gallego Romero et al. 2015), or test the libraries in mouse lines that share the B-cell progenitor lineage of GM12878, as can be generated by differentiations of mESCs down the hematopoietic lineage (Cho et al. 1999). Overall, more detailed exploration of the relationship is warranted.

### **Evidence for role of repression in mammalian *cis*-regulation**

Another tantalizing result that is shared between the parts of my thesis work is observations of repression for CRSs from the human and mouse genomes that were expected to be activating. In Chapter 2, over 60% of tested ChIP-seq peaks failed to drive expression overall minimal promoter controls, despite these sequences including three high quality pluripotency TFBSs. The high activity levels of some genomic sequences tested in mESCs made it difficult to sequence deeply enough to determine if some of the inactive sequences were actively repressing expression and from sequence analysis, I did not detect an enrichment of REST TFBS, a



canonical repressor of neuronal cell fates expressed in mESCs, for the lowest expressing sequences (McGann et al. 2014). However, OCT4 has been shown to interact with subunits of the repressive NuRD complex in mESCs, suggesting that some of the sequences that do not show activity in the assay may show evidence of repression if tested independently (Pietersen and van Lohuizen 2008; Rajasekhar and Begemann 2007; Liang et al. 2008). In Chapter 3, approximately 45% of all sequences tested in the GM12878 cell line and 76% of sequences with significant activity, drove expression levels below that of basal controls. Although the fraction of sequences that drove expression below controls was surprising considering only a quarter of regions selected were annotated as ‘Repressive’ in function by the Combined segmentations, but consistent with estimations that the majority of noncoding regions of the human genome is likely to be repressive (Thiel, Lietz, and Hohl 2004; Pennacchio et al. 2013; Reynolds, O’Shaughnessy, and Hendrich 2013). Repression is an important part of determining cell fate and proper cellular function, and both gain and loss of repression of key driver genes can contribute to cancer (Thiel, Lietz, and Hohl 2004; S. Zhou, Treloar, and Lupien 2016). Together, these results suggest that developing an MPRA that enriches for repressive sequences, perhaps by using a stronger minimal promoter and replacing mCherry with a reporter that can be weakly selected against to avoid the majority of sequencing reads going to highly activating sequences, would be a valuable tool for better understanding the *cis*-regulatory landscape of the human genome.

## **Conclusions**

Overall, my dissertation work has shown that it is possible to connect genotype to quantitative, *cis*-regulatory function but that a more detailed understanding of how changes in regulatory activity in the genome may impact an organism. Although positive results appear to be generally beyond the scope of this thesis, important hypotheses about *cis*-regulation have been ruled out and my work has highlighted the complex regulatory landscape of mammalian genomes. I have shown that modeling can uncover grammar from diverse genomic sequences and that testing the regulatory impact of human variants alone may not be sufficient to identify alleles that contribute

to disease. I hope that this work will help guide future efforts to avoid pitfalls in experimental design and assumptions and I am excited to see what the collect efforts of the field will uncover.

## **References**

- 1000 Genomes Project Consortium, Goncalo R. Abecasis, Adam Auton, Lisa D. Brooks, Mark A. DePristo, Richard M. Durbin, Robert E. Handsaker, Hyun Min Kang, Gabor T. Marth, and Gil A. McVean. 2012. “An Integrated Map of Genetic Variation from 1,092 Human Genomes.” *Nature* 491 (7422): 56–65.
- 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korb, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.
- Adachi, Kenjiro, Wolfgang Kopp, Guangming Wu, Sandra Heising, Boris Greber, Martin Stehling, Marcos J. Araúzo-Bravo, et al. 2018. “Esrrb Unlocks Silenced Enhancers for Reprogramming to Naive Pluripotency.” *Cell Stem Cell*, June.  
<https://doi.org/10.1016/j.stem.2018.05.020>.
- Albert, R., H. Jeong, and A. L. Barabasi. 2000. “Error and Attack Tolerance of Complex Networks.” *Nature* 406 (6794): 378–82.
- Alleyne, Trevis M., Lourdes Peña-castillo, Gwenael Badis, Shaheynoor Talukder, Michael F. Berger, Andrew R. Gehrke, Anthony A. Philippakis, Martha L. Bulyk, Quaid D. Morris, and Timothy R. Hughes. 2008. “Predicting the Binding Preference of Transcription Factors to In- Dividual DNA K-Mers,” 1–7.
- Arvey, Aaron, Phaedra Agius, William Stafford Noble, and Christina Leslie. 2012. “Sequence and Chromatin Determinants of Cell-Type-Specific Transcription Factor Binding.” *Genome Research* 22 (9): 1723–34.
- Asthana, Saurabh, William S. Noble, Gregory Kryukov, Charles E. Grant, Shamil Sunyaev, and John A. Stamatoyannopoulos. 2007. “Widely Distributed Noncoding Purifying Selection in the Human Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (30): 12410–15.
- Badis, Gwenael, Michael F. Berger, Anthony a. Philippakis, Shaheynoor Talukder, Andrew R. Gehrke, Savina a. Jaeger, Esther T. Chan, et al. 2009. “Diversity and Complexity in DNA Recognition by Transcription Factors.” *Science* 324 (5935): 1720–23.
- Bailey, Timothy L., Mikael Boden, Fabian a. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. 2009. “MEME SUITE: Tools for Motif Discovery and Searching.” *Nucleic Acids Research* 37 (Web Server issue): W202–8.
- Bailey, Timothy L., and Philip Machanick. 2012. “Inferring Direct DNA Binding from ChIP-Seq.” *Nucleic Acids Research* 40 (17): e128.
- Basu, Sumanta, Karl Kumbier, James B. Brown, and Bin Yu. 2018. “Iterative Random Forests to Discover Predictive and Stable High-Order Interactions.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (8): 1943–48.
- Bonifer, Constanze. 2000. “Developmental Regulation of Eukaryotic Gene Loci.” *Trends in Genetics: TIG* 16 (7): 310–14.
- Bulyk, Martha L. 2003. “Computational Prediction of Transcription-Factor Binding Site Locations.” *Genome Biology* 5 (1): 201.

- Castaldi, Peter J., Feng Guo, Dandi Qiao, Fei Du, Zun Zar Chi Naing, Yan Li, Betty Pham, et al. 2018. "Identification of Functional Variants in the FAM13A COPD GWAS Locus by Massively Parallel Reporter Assays." *American Journal of Respiratory and Critical Care Medicine*, August. <https://doi.org/10.1164/rccm.201802-0337OC>.
- Castillo-Davis, Cristian I. 2005. "The Evolution of Noncoding DNA: How Much Junk, How Much Func?" *Trends in Genetics: TIG* 21 (10): 533–36.
- Chambers, Ian, and Simon R. Tomlinson. 2009. "The Transcriptional Foundation of Pluripotency." *Development* 136: 2311–22.
- Chaudhari, Hemangi G., and Barak A. Cohen. 2018. "Local Sequence Features That Influence AP-1 Cis-Regulatory Activity." *Genome Research* 28 (2): 171–81.
- Chen, Chih-Yu, Quaid Morris, and Jennifer a. Mitchell. 2012. "Enhancer Identification in Mouse Embryonic Stem Cells Using Integrative Modeling of Chromatin and Genomic Features." *BMC Genomics* 13 (1): 152.
- Chen, Christina T. L., David I. Gottlieb, and Barak A. Cohen. 2008. "Ultraconserved Elements in the Olig2 Promoter." *PloS One* 3 (12): e3946.
- Chen, Xi, and Hemant Ishwaran. 2012. "Random Forests for Genomic Data Analysis." *Genomics* 99 (6): 323–29.
- Chen, Xi, Han Xu, Ping Yuan, Fang Fang, Mikael Huss, Vinsensius B. Vega, Eleanor Wong, et al. 2008. "Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells." *Cell* 133: 1106–17.
- Chen, X., V. B. Vega, and H-H Ng. 2008. "Transcriptional Regulatory Networks in Embryonic Stem Cells." *Cold Spring Harbor Symposia on Quantitative Biology* 73: 203–9.
- Chimpanzee Sequencing and Analysis Consortium. 2005. "Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome." *Nature* 437 (7055): 69–87.
- Cho, S. K., T. D. Webber, J. R. Carlyle, T. Nakano, S. M. Lewis, and J. C. Zúñiga-Pflücker. 1999. "Functional Characterization of B Lymphocytes Generated in Vitro from Embryonic Stem Cells." *Proceedings of the National Academy of Sciences of the United States of America* 96 (17): 9797–9802.
- Chun, Sung, Alexandra Casparino, Nikolaos A. Patsopoulos, Damien C. Croteau-Chonka, Benjamin A. Raby, Philip L. De Jager, Shamil R. Sunyaev, and Chris Cotsapas. 2017. "Limited Statistical Evidence for Shared Genetic Effects of eQTLs and Autoimmune-Disease-Associated Loci in Three Major Immune-Cell Types." *Nature Genetics* 49 (4): 600–605.
- Clarke, Laura, Susan Fairley, Xiangqun Zheng-Bradley, Ian Streeter, Emily Perry, Ernesto Lowy, Anne-Marie Tassé, and Paul Flicek. 2017. "The International Genome Sample Resource (IGSR): A Worldwide Collection of Genome Variation Incorporating the 1000 Genomes Project Data." *Nucleic Acids Research* 45 (D1): D854–59.
- Clarke, Shoa L., Julia E. VanderMeer, Aaron M. Wenger, Bruce T. Schaar, Nadav Ahituv, and Gill Bejerano. 2012. "Human Developmental Enhancers Conserved between Deuterostomes and Protostomes." *PLoS Genetics* 8 (8): e1002852.
- Cookson, William, Liming Liang, Gonçalo Abecasis, Miriam Moffatt, and Mark Lathrop. 2009. "Mapping Complex Disease Traits with Global Gene Expression." *Nature Reviews. Genetics* 10 (3): 184–94.

- Cox, Robert Sidney, 3rd, Michael G. Surette, and Michael B. Elowitz. 2007. “Programming Gene Expression with Combinatorial Promoters.” *Molecular Systems Biology* 3 (November): 145.
- Cusanovich, Darren A., Bryan Pavlovic, Jonathan K. Pritchard, and Yoav Gilad. 2013. “The Functional Consequences of Variation in Transcription Factor Binding.” *Author’s Manuscript* N/A (N/A): 1–30.
- Derrien, Thomas, Jordi Estellé, Santiago Marco Sola, David G. Knowles, Emanuele Raineri, Roderic Guigó, and Paolo Ribeca. 2012. “Fast Computation and Applications of Genome Mappability.” *PLoS One* 7 (1): e30377.
- D’haeseleer, Patrik. 2006. “What Are DNA Sequence Motifs?” *Nature Biotechnology* 24 (4): 423–25.
- Dorigi, Kristel M., Tomek Swigut, Telmo Henriques, Natarajan V. Bhanu, Benjamin S. Scruggs, Nataliya Nady, Christopher D. Still 2nd, Benjamin A. Garcia, Karen Adelman, and Joanna Wysocka. 2017. “Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation.” *Molecular Cell* 66 (4): 568–76.e4.
- Dunn, S-J, G. Martello, B. Yordanov, S. Emmott, and a. G. Smith. 2014. “Defining an Essential Transcription Factor Program for Naïve Pluripotency.” *Science* 344: 1156–60.
- Eddy, Sean R. 2013. “The ENCODE Project: Missteps Overshadowing a Success.” *Current Biology: CB* 23 (7): R259–61.
- ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.
- Ernst, Jason, and Manolis Kellis. 2012. “ChromHMM: Automating Chromatin-State Discovery and Characterization.” *Nature Methods* 9 (3): 215–16.
- Ernst, Jason, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shores, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, et al. 2011. “Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types.” *Nature* 473 (7345): 43–49.
- Evans, Nicole C., Christina I. Swanson, and Scott Barolo. 2012. *Sparkling Insights into Enhancer Structure, Function, and Evolution*. 1st ed. Vol. 98. Elsevier Inc.
- Feng, Bo, Jianming Jiang, Petra Kraus, Jia-Hui Ng, Jian-Chien Dominic Heng, Yun-Shen Chan, Lai-Ping Yaw, et al. 2009. “Reprogramming of Fibroblasts into Induced Pluripotent Stem Cells with Orphan Nuclear Receptor Esrrb.” *Nature Cell Biology* 11 (2): 197–203.
- Ferraris, Luciana, Allan P. Stewart, Jinsuk Kang, Alec M. DeSimone, Matthew Gemberling, Dean Tantin, and William G. Fairbrother. 2011. “Combinatorial Binding of Transcription Factors in the Pluripotency Control Regions of the Genome.” *Genome Research* 21: 1055–64.
- Fiore, Chris, and Barak A. Cohen. 2016. “Interactions between Pluripotency Factors Specify Cis-Regulation in Embryonic Stem Cells.” *Genome Research* 26 (6): 778–86.
- Fisher, William W., Jingyi Jessica Li, Ann S. Hammonds, James B. Brown, Barret D. Pfeiffer, Richard Weiszmman, Stewart MacArthur, et al. 2012. “DNA Regions Bound at Low Occupancy by Transcription Factors Do Not Drive Patterned Reporter Gene Expression in *Drosophila*.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (52): 21330–35.
- Fletez-Brant, Christopher, Dongwon Lee, Andrew S. McCallion, and Michael A. Beer. 2013.

- “Kmer-SVM: A Web Server for Identifying Predictive Regulatory Sequence Features in Genomic Data Sets.” *Nucleic Acids Research* 41 (Web Server issue): W544–56.
- Flicek, Paul, Ikhlak Ahmed, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, et al. 2013. “Ensembl 2013.” *Nucleic Acids Research* 41 (Database issue): D48–55.
- Frankel, Nicolás. 2012. “Multiple Layers of Complexity in Cis-Regulatory Regions of Developmental Genes.” *Developmental Dynamics: An Official Publication of the American Association of Anatomists* 241 (12): 1857–66.
- Gabut, Mathieu, Payman Samavarchi-Tehrani, Xinchun Wang, Valentina Slobodeniuc, Dave O’Hanlon, Hoon Ki Sung, Manuel Alvarez, et al. 2011. “An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming.” *Cell* 147 (1): 132–46.
- Gallego Romero, Irene, Bryan J. Pavlovic, Irene Hernando-Herraez, Xiang Zhou, Michelle C. Ward, Nicholas E. Banovich, Courtney L. Kagan, et al. 2015. “A Panel of Induced Pluripotent Stem Cells from Chimpanzees: A Resource for Comparative Functional Genomics.” *eLife* 4 (June): e07103.
- Gertz, Jason, Eric D. Siggia, and Barak A. Cohen. 2009. “Analysis of Combinatorial Cis-Regulation in Synthetic and Genomic Promoters.” *Nature* 457 (7226): 215–18.
- Ghandi, Mahmoud, Dongwon Lee, Morteza Mohammad-Noori, and Michael A. Beer. 2014. “Enhanced Regulatory Sequence Prediction Using Gapped K-Mer Features.” *PLoS Computational Biology* 10 (7): e1003711.
- Ghandi, Mahmoud, Morteza Mohammad-Noori, Narges Ghareghani, Dongwon Lee, Levi Garraway, and Michael A. Beer. 2016. “gkmSVM: An R Package for Gapped-Kmer SVM.” *Bioinformatics* 32 (14): 2205–7.
- Gibson, Greg. 2012. “Rare and Common Variants: Twenty Arguments.” *Nature Reviews. Genetics* 13 (2): 135–45.
- Giorgetti, Luca, Trevor Siggers, Guido Tian, Greta Caprara, Samuele Notarbartolo, Teresa Corona, Manolis Pasparakis, Paolo Milani, Martha L. Bulyk, and Gioacchino Natoli. 2010. “Noncooperative Interactions between Transcription Factors and Clustered DNA Binding Sites Enable Graded Transcriptional Responses to Environmental Inputs.” *Molecular Cell* 37 (3): 418–28.
- Grant, Charles E., Timothy L. Bailey, and William Stafford Noble. 2011. “FIMO: Scanning for Occurrences of a given Motif.” *Bioinformatics* 27 (7): 1017–18.
- Grossman, Sharon R., Xiaolan Zhang, Li Wang, Jesse Engreitz, Alexandre Melnikov, Peter Rogov, Ryan Tewhey, et al. 2017. “Systematic Dissection of Genomic Features Determining Transcription Factor Binding and Enhancer Function.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (7): E1291–1300.
- Guertin, Michael J., and John T. Lis. 2010. “Chromatin Landscape Dictates HSF Binding to Target DNA Elements.” *PLoS Genetics* 6 (9): e1001114.
- Guertin, Michael J., and John T. Lis. 2013. “Mechanisms by Which Transcription Factors Gain Access to Target Sequence Elements in Chromatin.” *Current Opinion in Genetics & Development* 23 (2): 116–23.
- Gulko, Brad, Melissa J. Hubisz, Ilan Gronau, and Adam Siepel. 2015. “A Method for Calculating Probabilities of Fitness Consequences for Point Mutations across the Human

- Genome.” *Nature Genetics* 47 (3): 276–83.
- Hare, Emily E., Brant K. Peterson, and Michael B. Eisen. 2008. “A Careful Look at Binding Site Reorganization in the Even-Skipped Enhancers of *Drosophila* and Sepsids.” *PLoS Genetics* 4 (11): 1–5.
- Hare, Emily E., Brant K. Peterson, Venky N. Iyer, Rudolf Meier, and Michael B. Eisen. 2008. “Sepsid Even-Skipped Enhancers Are Functionally Conserved in *Drosophila* despite Lack of Sequence Conservation.” *PLoS Genetics* 4 (6): e1000106.
- Hoffman, Michael M., Orion J. Buske, Jeff A. Bilmes, and William Stafford Noble. 2011. “Segway: Simultaneous Segmentation of Multiple Functional Genomics Data Sets with Heterogeneous Patterns of Missing Data.”  
<http://noble.gs.washington.edu/proj/segway/manuscript/temposegment.nips09.hoffman.pdf>.
- Hoffman, Michael M., Jason Ernst, Steven P. Wilder, Anshul Kundaje, Robert S. Harris, Max Libbrecht, Belinda Giardine, et al. 2013. “Integrative Annotation of Chromatin Elements from ENCODE Data.” *Nucleic Acids Research* 41 (2): 827–41.
- Hrdlickova, Barbara, Rodrigo Coutinho de Almeida, Zuzanna Borek, and Sebo Withoff. 2014. “Genetic Variation in the Non-Coding Genome: Involvement of Micro-RNAs and Long Non-Coding RNAs in Disease.” *Biochimica et Biophysica Acta* 1842 (10): 1910–22.
- Huang, Jinyan, Taotao Chen, Xiaosong Liu, Jing Jiang, Jinsong Li, Dangsheng Li, X. Shirley Liu, Wei Li, Jiuhong Kang, and Gang Pei. 2009. “More Synergetic Cooperation of Yamanaka Factors in Induced Pluripotent Stem Cells than in Embryonic Stem Cells.” *Cell Research* 19 (10): 1127–38.
- International HapMap Consortium. 2005. “A Haplotype Map of the Human Genome.” *Nature* 437 (7063): 1299–1320.
- Jauch, Ralf, Calista Keow Leng Ng, Kumar Singh Saikatendu, Raymond C. Stevens, and Prasanna R. Kolatkar. 2008. “Crystal Structure and DNA Binding of the Homeodomain of the Stem Cell Transcription Factor Nanog.” *Journal of Molecular Biology* 376 (3): 758–70.
- Jauch, Ralf, Calista K. L. Ng, Kamesh Narasimhan, and Prasanna R. Kolatkar. 2012. “The Crystal Structure of the Sox4 HMG Domain-DNA Complex Suggests a Mechanism for Positional Interdependence in DNA Recognition.” *Biochemical Journal* 443 (1): 39–47.
- Jenuwein, T., and C. D. Allis. 2001. “Translating the Histone Code.” *Science* 293 (5532): 1074–80.
- Johnson, Lisa a., Ying Zhao, Krista Golden, and Scott Barolo. 2008. “Reverse-Engineering a Transcriptional Enhancer: A Case Study in *Drosophila*.” *Tissue Engineering. Part A* 14 (9): 1549–59.
- John Wiley & Sons, Ltd, ed. 2001. “Transcriptional Regulation: Evolution.” In *Encyclopedia of Life Sciences*, 92:1684. Chichester, UK: John Wiley & Sons, Ltd.
- Junion, Guillaume, Mikhail Spivakov, Charles Girardot, Martina Braun, E. Hilary Gustafson, Ewan Birney, and Eileen E. M. Furlong. 2012. “A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History.” *Cell* 148 (3): 473–86.
- Kadonaga, J. T., and R. Tjian. 1986. “Affinity Purification of Sequence-Specific DNA Binding Proteins.” *Proceedings of the National Academy of Sciences of the United States of America* 83 (16): 5889–93.
- Keimpema, Martine van, Leonie J. Grüneberg, Michal Mokry, Ruben van Boxtel, Menno C. van

- Zelm, Paul Coffey, Steven T. Pals, and Marcel Spaargaren. 2015. "The Forkhead Transcription Factor FOXP1 Represses Human Plasma Cell Differentiation." *Blood* 126 (18): 2098–2109.
- Kheradpour, Pouya, Jason Ernst, Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Jessica Alston, Tarjei S. Mikkelsen, and Manolis Kellis. 2013. "Systematic Dissection of Regulatory Motifs in 2000 Predicted Human Enhancers Using a Massively Parallel Reporter Assay." *Genome Research* 23 (5): 800–811.
- Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O’Roak, Gregory M. Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." *Nature Genetics* 46 (3): 310–15.
- Konopka, Genevieve, Jamee M. Bomar, Kellen Winden, Giovanni Coppola, Zophonias O. Jonsson, Fuying Gao, Sophia Peng, Todd M. Preuss, James A. Wohlschlegel, and Daniel H. Geschwind. 2009. "Human-Specific Transcriptional Regulation of CNS Development Genes by FOXP2." *Nature* 462 (7270): 213–17.
- Kouzarides, Tony. 2007. "Chromatin Modifications and Their Function." *Cell* 128 (4): 693–705.
- Kuhn, Robert M., David Haussler, and W. James Kent. 2013. "The UCSC Genome Browser and Associated Tools." *Briefings in Bioinformatics* 14 (2): 144–61.
- Kulkarni, Meghana M., and David N. Arnosti. 2003. "Information Display by Transcriptional Enhancers." *Development* 130: 6569–75.
- Kulzer, Jennifer R., Michael L. Stitzel, Mario A. Morken, Jeroen R. Huyghe, Christian Fuchsberger, Johanna Kuusisto, Markku Laakso, Michael Boehnke, Francis S. Collins, and Karen L. Mohlke. 2014. "A Common Functional Regulatory Variant at a Type 2 Diabetes Locus Upregulates ARAP1 Expression in the Pancreatic Beta Cell." *American Journal of Human Genetics* 94 (2): 186–97.
- Kwasnieski, Jamie C., Christopher Fiore, Hemangi G. Chaudhari, and Barak A. Cohen. 2014. "High-Throughput Functional Testing of ENCODE Segmentation Predictions." *Genome Research* 24 (10): 1595–1602.
- Kwasnieski, Jamie C., Ilaria Mogno, Connie A. Myers, Joseph C. Corbo, and Barak A. Cohen. 2012. "Complex Effects of Nucleotide Variants in a Mammalian Cis -Regulatory Element." *Proceedings of the National Academy of Sciences of the United States of America* 109 (47): 19498–503.
- Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. ’t Hoen, Jean Monlong, Manuel A. Rivas, Mar González-Porta, et al. 2013. "Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans." *Nature* 501 (7468): 506–11.
- Lelli, Katherine M., Matthew Slattery, and Richard S. Mann. 2012. "Disentangling the Many Layers of Eukaryotic Transcriptional Regulation." *Annual Review of Genetics* 46 (January): 43–68.
- Levine, Michael, and Robert Tjian. 2003. "Transcription Regulation and Animal Diversity." *Nature* 424 (6945): 147–51.
- Levy, Samuel, and Sridhar Hannenhalli. 2002. "Identification of Transcription Factor Binding Sites in the Human Genome Sequence." *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 14: 510–14.
- Liang, Jiancong, Ma Wan, Yi Zhang, Peili Gu, Huawei Xin, Sung Yun Jung, Jun Qin, et al. 2008.



- “Nanog and Oct4 Associate with Unique Transcriptional Repression Complexes in Embryonic Stem Cells.” *Nature Cell Biology* 10 (6): 731–39.
- Li, Heng. 2011. “Tabix: Fast Retrieval of Sequence Features from Generic TAB-Delimited Files.” *Bioinformatics* 27 (5): 718–19.
- Lindblad-Toh, Kerstin, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, et al. 2011. “A High-Resolution Map of Human Evolutionary Constraint Using 29 Mammals.” *Nature* 478 (7370): 476–82.
- Linsel-Nitschke, Patrick, Jörg Heeren, Zouhair Aherrahrou, Petra Bruse, Christian Gieger, Thomas Illig, Holger Prokisch, et al. 2010. “Genetic Variation at Chromosome 1p13.3 Affects Sortilin mRNA Expression, Cellular LDL-Uptake and Serum LDL Levels Which Translates to the Risk of Coronary Artery Disease.” *Atherosclerosis* 208 (1): 183–89.
- Liu, Xiaosong, Jinyan Huang, Taotao Chen, Ying Wang, Shunmei Xin, Jian Li, Gang Pei, and Juhong Kang. 2008. “Yamanaka Factors Critically Regulate the Developmental Signaling Network in Mouse Embryonic Stem Cells.” *Cell Research* 18: 1177–89.
- Li, Yue, Alvin Houze Shi, Ryan Tewhey, Pardis C. Sabeti, Jason Ernst, and Manolis Kellis. n.d. “Genome-Wide Regulatory Model from MPRA Data Predicts Functional Regions, eQTLs, and GWAS Hits.” <https://doi.org/10.1101/110171>.
- Loh, Yui-Han, Qiang Wu, Joon-Lin Chew, Vinsensius B. Vega, Weiwei Zhang, Xi Chen, Guillaume Bourque, et al. 2006. “The Oct4 and Nanog Transcription Network Regulates Pluripotency in Mouse Embryonic Stem Cells.” *Nature Genetics* 38 (4): 431–40.
- Louppe, Gilles, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. 2013. “Understanding Variable Importances in Forests of Randomized Trees.” In *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, 431–39. Curran Associates, Inc.
- Ludwig, M. Z., C. Bergman, N. H. Patel, and M. Kreitman. 2000. “Evidence for Stabilizing Selection in a Eukaryotic Enhancer Element.” *Nature* 403 (6769): 564–67.
- Lupien, Mathieu, Jérôme Eeckhoute, Clifford A. Meyer, Qianben Wang, Yong Zhang, Wei Li, Jason S. Carroll, X. Shirley Liu, and Myles Brown. 2008. “FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription.” *Cell* 132: 958–70.
- Lu, Yi-Fan, David M. Mauger, David B. Goldstein, Thomas J. Urban, Kevin M. Weeks, and Shelton S. Bradrick. 2015. “IFNL3 mRNA Structure Is Remodeled by a Functional Non-Coding Polymorphism Associated with Hepatitis C Virus Clearance.” *Scientific Reports* 5 (November): 16037.
- Majumder, Partha P., and Saurabh Ghosh. 2005. “Mapping Quantitative Trait Loci in Humans: Achievements and Limitations.” *The Journal of Clinical Investigation* 115 (6): 1419–24.
- Mansur, Yasmina A., Elena Rojano, Juan A. G. Ranea, and James R. Perkins. 2018. “Chapter 7 - Analyzing the Effects of Genetic Variation in Noncoding Genomic Regions.” In *Precision Medicine*, edited by Hans-Peter Digner and Matthias Kohl, 119–44. Academic Press.
- Mardis, Elaine R. 2011. “A Decade’s Perspective on DNA Sequencing Technology.” *Nature* 470 (7333): 198–203.
- Maricque, Brett B., Joseph D. Dougherty, and Barak A. Cohen. 2017. “A Genome-Integrated Massively Parallel Reporter Assay Reveals DNA Sequence Determinants of Cis-Regulatory Activity in Neural Cells.” *Nucleic Acids Research* 45 (4): e16.

- Maston, Glenn a., Sara K. Evans, and Michael R. Green. 2006. "Transcriptional Regulatory Elements in the Human Genome." *Annual Review of Genomics and Human Genetics* 7 (January): 29–59.
- Masui, Shinji, Daisuke Shimosato, Yayoi Toyooka, Rika Yagi, Kazue Takahashi, and Hitoshi Niwa. 2005. "An Efficient System to Establish Multiple Embryonic Stem Cell Lines Carrying an Inducible Expression Unit." *Nucleic Acids Research* 33 (4): 1–8.
- Mathelier, Anthony, Wenqiang Shi, and Wyeth W. Wasserman. 2015. "Identification of Altered Cis-Regulatory Elements in Human Disease." *Trends in Genetics: TIG* 31 (2): 67–76.
- Mather, Cheryl A., Sean D. Mooney, Stephen J. Salipante, Sheena Scroggins, David Wu, Colin C. Pritchard, and Brian H. Shirts. 2016. "CADD Score Has Limited Clinical Validity for the Identification of Pathogenic Variants in Noncoding Regions in a Hereditary Cancer Panel." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 18 (12): 1269–75.
- McGann, James C., Jon A. Oyer, Saurabh Garg, Huilan Yao, Jun Liu, Xin Feng, Lujian Liao, John R. Yates 3rd, and Gail Mandel. 2014. "Polycomb- and REST-Associated Histone Deacetylases Are Independent Pathways toward a Mature Neuronal Phenotype." *eLife* 3 (September): e04235.
- Menze, Bjoern H., B. Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A. Hamprecht. 2009. "A Comparison of Random Forest and Its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data." *BMC Bioinformatics* 10 (July): 213.
- Miosge, Lisa A., Matthew A. Field, Yovina Sontani, Vicky Cho, Simon Johnson, Anna Palkova, Bhavani Balakishnan, et al. 2015. "Comparison of Predicted and Actual Consequences of Missense Mutations." *Proceedings of the National Academy of Sciences of the United States of America* 112 (37): E5189–98.
- Mogno, Ilaria, Jamie C. Kwasnieski, and Barak A. Cohen. 2013. "Massively Parallel Synthetic Promoter Assays Reveal the in Vivo Effects of Binding Site Variants." *Genome Research* 23 (11): 1908–15.
- Mohammadi, Pejman, Stephane E. Castel, Andrew A. Brown, and Tuuli Lappalainen. 2017. "Quantifying the Regulatory Effect Size of Cis-Acting Genetic Variation Using Allelic Fold Change." *Genome Research* 27 (11): 1872–84.
- Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, et al. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome." *Nature* 420 (6915): 520–62.
- Mulas, Carla, Gloryn Chia, Kenneth Alan Jones, Andrew Christopher Hodgson, Giuliano Giuseppe Stirparo, and Jennifer Nichols. 2018. "Oct4 Regulates the Embryonic Axis and Coordinates Exit from Pluripotency and Germ Layer Specification in the Mouse Embryo." *Development* 145 (12). <https://doi.org/10.1242/dev.159103>.
- Musunuru, Kiran, Alanna Strong, Maria Frank-Kamenetsky, Noemi E. Lee, Tim Ahfeldt, Katherine V. Sachs, Xiaoyu Li, et al. 2010. "From Noncoding Variant to Phenotype via SORT1 at the 1p13 Cholesterol Locus." *Nature* 466 (7307): 714–19.
- Myint, Leslie, Ruihua Wang, Leandros Boukas, Kasper D. Hansen, Loyal A. Goff, and Dimitrios Avramopoulos. 2018. "Testing the Regulatory Consequences of 1,049 Schizophrenia

- Associated Variants With a Massively Parallel Reporter Assay.” *bioRxiv*. bioRxiv. <https://doi.org/10.1101/447557>.
- Natarajan, Anirudh, Galip Gürkan Yardimci, Nathan C. Sheffield, Gregory E. Crawford, and Uwe Ohler. 2012. “Predicting Cell-Type-Specific Gene Expression from Regions of Open Chromatin.” *Genome Research* 22 (9): 1711–22.
- Ng, Pauline C., and Steven Henikoff. 2003. “SIFT: Predicting Amino Acid Changes That Affect Protein Function.” *Nucleic Acids Research* 31 (13): 3812–14.
- Nicolae, Dan L., Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan, and Nancy J. Cox. 2010. “Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS.” *PLoS Genetics* 6 (4): e1000888.
- Niwa, Hitoshi. 2014. “The Pluripotency Transcription Factor Network at Work in Reprogramming.” *Current Opinion in Genetics & Development* 28: 25–31.
- Nolis, Ilias K., Daniel J. McKay, Eva Mantouvalou, Stavros Lomvardas, Menie Merika, and Dimitris Thanos. 2009. “Transcription Factors Mediate Long-Range Enhancer-Promoter Interactions.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (48): 20222–27.
- Noonan, James P., and Andrew S. McCallion. 2010. “Genomics of Long-Range Regulatory Elements.” *Genomics Human Genetics* 11: 1–23.
- Oakes, Christopher C., Marc Seifert, Yassen Assenov, Lei Gu, Martina Przekopowicz, Amy S. Ruppert, Qi Wang, et al. 2016. “DNA Methylation Dynamics during B Cell Maturation Underlie a Continuum of Disease Phenotypes in Chronic Lymphocytic Leukemia.” *Nature Genetics* 48 (3): 253–64.
- Oldoni, Federico, Jutta Palmen, Claudia Giambartolomei, Philip Howard, Fotios Drenos, Vincent Plagnol, Steve E. Humphries, Philippa J. Talmud, and Andrew J. P. Smith. 2016. “Post-GWAS Methodologies for Localisation of Functional Non-Coding Variants: ANGPTL3.” *Atherosclerosis* 246 (March): 193–201.
- Pan, Guangjin, and James a. Thomson. 2007. “Nanog and Transcriptional Networks in Embryonic Stem Cell Pluripotency.” *Cell Research* 17: 42–49.
- Panne, Daniel. 2008. “The Enhanceosome.” *Current Opinion in Structural Biology* 18: 236–42.
- Pan, Yongping, Chung-Jung Tsai, Buyong Ma, and Ruth Nussinov. 2010. “Mechanisms of Transcription Factor Selectivity.” *Trends in Genetics: TIG* 26 (2): 75–83.
- Parker, Stephen C. J., Michael L. Stitzel, D. Leland Taylor, Jose Miguel Orozco, Michael R. Erdos, Jennifer a. Akiyama, Kelly Lammerts van Bueren, et al. 2013. *Chromatin Stretch Enhancer States Drive Cell-Specific Gene Regulation and Harbor Human Disease Risk Variants*. Vol. 110.
- Patwardhan, Rupali P., Joseph B. Hiatt, Daniela M. Witten, Mee J. Kim, Robin P. Smith, Dalit May, Choli Lee, et al. 2012. “Massively Parallel Functional Dissection of Mammalian Enhancers in Vivo.” *Nature Biotechnology* 30 (3): 265–70.
- Pennacchio, Len A., Nadav Ahituv, Alan M. Moses, Shyam Prabhakar, Marcelo A. Nobrega, Malak Shoukry, Simon Minovitsky, et al. 2006. “In Vivo Enhancer Analysis of Human Conserved Non-Coding Sequences.” *Nature* 444 (7118): 499–502.
- Pennacchio, Len A., Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano. 2013. “Enhancers: Five Essential Questions.” *Nature Reviews. Genetics* 14 (4): 288–95.

- Piens, Marie, Marc Muller, Morgan Bodson, Gregory Baudouin, and Jean-Christophe Plumier. 2010. "A Short Upstream Promoter Region Mediates Transcriptional Regulation of the Mouse Doublecortin Gene in Differentiating Neurons." *BMC Neuroscience* 11: 64.
- Pietersen, Alexandra M., and Maarten van Lohuizen. 2008. "Stem Cell Regulation by Polycomb Repressors: Postponing Commitment." *Current Opinion in Cell Biology* 20 (2): 201–7.
- Pollard, Katherine S., Sofie R. Salama, Bryan King, Andrew D. Kern, Tim Dreszer, Sol Katzman, Adam Siepel, et al. 2006. "Forces Shaping the Fastest Evolving Regions in the Human Genome." *PLoS Genetics* 2 (10): e168.
- Ponomarenko, Julia V., Galina V. Orlova, Anatoly S. Frolov, Mikhail S. Gelfand, and Mikhail P. Ponomarenko. 2002. "SELEX\_DB: A Database on in Vitro Selected Oligomers Adapted for Recognizing Natural Sites and for Analyzing Both SNPs and Site-Directed Mutagenesis Data." *Nucleic Acids Research* 30 (1): 195–99.
- Ponting, Chris P., and Ross C. Hardison. 2011. "What Fraction of the Human Genome Is Functional?" *Genome Research* 21 (11): 1769–76.
- Poulter, M., E. Hollox, C. B. Harvey, C. Mulcare, Katri Peuhkuri, Kajsa Kajander, M. Sarner, Riitta Korpela, and D. M. Swallow. 2003. "The Causal Element for the Lactase Persistence/non-Persistence Polymorphism Is Located in a 1 Mb Region of Linkage Disequilibrium in Europeans." *Annals of Human Genetics* 67 (4): 298–311.
- Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.
- Raghavan, Avanthi, Derek Peters, Nicolas Kuperwasser, Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Tarjei Mikkelsen, and Kiran Musunuru. 2014. "Abstract 242: Functional Characterization of a Cis-eQTL Locus for Plasma Cholesterol Using CRISPR/Cas Genome Editing in Human Pluripotent Stem Cells." *Arteriosclerosis, Thrombosis, and Vascular Biology*.  
[https://www.ahajournals.org/doi/abs/10.1161/atvb.34.suppl\\_1.242](https://www.ahajournals.org/doi/abs/10.1161/atvb.34.suppl_1.242).
- Rajasekhar, Vinagolu K., and Martin Begemann. 2007. "Concise Review: Roles of Polycomb Group Proteins in Development and Disease: A Stem Cell Perspective." *Stem Cells* 25 (10): 2498–2510.
- Rana, Anita, Sudhir Jain, Nitin Puri, Meenakshi Kaw, Natalie Sirianni, Deniz Eren, Brahma Raju Mopidevi, and Ashok Kumar. 2017. "The Transcriptional Regulation of the Human Angiotensinogen Gene after High-Fat Diet Is Haplotype-Dependent: Novel Insights into the Gene-Regulatory Networks and Implications for Human Hypertension." *PloS One* 12 (5): e0176373.
- Ranciaro, Alessia, Michael C. Campbell, Jibril B. Hirbo, Wen-Ya Ko, Alain Froment, Paolo Anagnostou, Maritha J. Kotze, et al. 2014. "Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa." *American Journal of Human Genetics* 94 (4): 496–510.
- Reményi, Attila, Katharina Lins, L. Johan Nissen, Rolland Reinbold, Hans R. Schöler, and Matthias Wilmanns. 2003. "Crystal Structure of a POU/HMG/DNA Ternary Complex Suggests Differential Assembly of Oct4 and Sox2 on Two Enhancers." *Genes & Development* 17 (16): 2048–59.
- Reményi, Attila, Hans R. Schöler, and Matthias Wilmanns. 2004. "Combinatorial Control of Gene Expression." *Nature Structural & Molecular Biology* 11 (9): 812–15.

- Reynolds, Nicola, Aoife O'Shaughnessy, and Brian Hendrich. 2013. "Transcriptional Repressors: Multifaceted Regulators of Gene Expression." *Development* 140 (3): 505–12.
- Ridefelt, Peter, and Lena D. Håkansson. 2005. "Lactose Intolerance: Lactose Tolerance Test versus Genotyping." *Scandinavian Journal of Gastroenterology* 40 (7): 822–26.
- Robasky, Kimberly, and Martha L. Bulyk. 2011. "UniPROBE, Update 2011: Expanded Content and Search Tools in the Online Database of Protein-Binding Microarray Data on Protein-DNA Interactions." *Nucleic Acids Research* 39 (Database issue): D124–28.
- Roche, Olga, María Laura Deguiz, María Tiana, Clara Galiana-Ribote, Daniel Martinez-Alcazar, Carlos Rey-Serra, Beatriz Ranz-Ribeiro, et al. 2016. "Identification of Non-Coding Genetic Variants in Samples from Hypoxemic Respiratory Disease Patients That Affect the Transcriptional Response to Hypoxia." *Nucleic Acids Research* 44 (19): 9315–30.
- Roeder, R. G. 1996. "The Role of General Initiation Factors in Transcription by RNA Polymerase II." *Trends in Biochemical Sciences* 21 (9): 327–35.
- Roeder, Robert G. 2003. "Lasker Basic Medical Research Award. The Eukaryotic Transcriptional Machinery: Complexities and Mechanisms Unforeseen." *Nature Medicine* 9 (10): 1239–44.
- Roukos, Dimitrios H. 2009. "Personal Genomics and Genome-Wide Association Studies: Novel Discoveries but Limitations for Practical Personalized Medicine." *Annals of Surgical Oncology* 16 (3): 772–73.
- Rubinstein, Marcelo, and Flávio S. J. de Souza. 2013. "Evolution of Transcriptional Enhancers and Animal Diversity." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 368: 20130017.
- Saint Pierre, Aude, and Emmanuelle Génin. 2014. "How Important Are Rare Variants in Common Disease?" *Briefings in Functional Genomics* 13 (5): 353–61.
- Sandelin, Albin, Wynand Alkema, Pär Engström, Wyeth W. Wasserman, and Boris Lenhard. 2004. "JASPAR: An Open-Access Database for Eukaryotic Transcription Factor Binding Profiles." *Nucleic Acids Research* 32 (Database issue): D91–94.
- Sandelin, Albin, Piero Carninci, Boris Lenhard, Jasmina Ponjavic, Yoshihide Hayashizaki, and David a. Hume. 2007. "Mammalian RNA Polymerase II Core Promoters: Insights from Genome-Wide Studies." *Nature Reviews. Genetics* 8 (6): 424–36.
- Sanyal, Amartya, Bryan R. Lajoie, Gaurav Jain, and Job Dekker. 2012. "The Long-Range Interaction Landscape of Gene Promoters." *Nature* 489 (7414): 109–13.
- Schork, Nicholas J., Sarah S. Murray, Kelly A. Frazer, and Eric J. Topol. 2009. "Common vs. Rare Allele Hypotheses for Complex Diseases." *Current Opinion in Genetics & Development* 19 (3): 212–19.
- Segal, Eran, Tali Raveh-Sadka, Mark Schroeder, Ulrich Unnerstall, and Ulrike Gaul. 2008. "Predicting Expression Patterns from Regulatory Sequence in *Drosophila* Segmentation." *Nature* 451 (7178): 535–40.
- Shen, Yin, Feng Yue, David F. McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, et al. 2012. "A Map of the Cis-Regulatory Sequences in the Mouse Genome." *Nature* 488 (7409): 116–20.
- Singh, Upinder, Rene H. Quintanilla, Scott Grecian, Kyle R. Gee, Mahendra S. Rao, and Uma Lakshmipathy. 2012. "Novel Live Alkaline Phosphatase Substrate for Identification of

- Pluripotent Stem Cells.” *Stem Cell Reviews* 8 (3): 1021–29.
- Spielman, Stephanie J., and Sergei L. Kosakovsky Pond. 2018. “Relative Evolutionary Rate Inference in HyPhy with LEISR.” *PeerJ* 6 (February): e4339.
- Spitz, François, and Eileen E. M. Furlong. 2012. “Transcription Factors: From Enhancer Binding to Developmental Control.” *Nature Reviews. Genetics* 13 (9): 613–26.
- Stormo, G. D., and D. S. Fields. 1998. “Specificity, Free Energy and Information Content in Protein-DNA Interactions.” *Trends in Biochemical Sciences* 23 (3): 109–13.
- Stranger, Barbara E., Eli A. Stahl, and Towfique Raj. 2011. “Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics.” *Genetics* 187 (2): 367–83.
- Swallow, Dallas M. 2003. “Genetics of Lactase Persistence and Lactose Intolerance.” *Annual Review of Genetics* 37: 197–219.
- Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors.” *Cell* 126 (4): 663–76.
- Tak, Yu Gyoung, and Peggy J. Farnham. 2015. “Making Sense of GWAS: Using Epigenomics and Genome Engineering to Understand the Functional Relevance of SNPs in Non-Coding Regions of the Human Genome.” *Epigenetics & Chromatin* 8 (December): 57.
- Tewhey, Ryan, Dylan Kotliar, Daniel S. Park, Brandon Liu, Sarah Winnicki, Steven K. Reilly, Kristian G. Andersen, et al. 2016. “Direct Identification of Hundreds of Expression-Modulating Variants Using a Multiplexed Reporter Assay.” *Cell* 165 (6): 1519–29.
- Thiel, Gerald, Michael Lietz, and Mathias Hohl. 2004. “How Mammalian Transcriptional Repressors Work.” *European Journal of Biochemistry / FEBS* 271 (14): 2855–62.
- Thurman, Robert E., Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield, et al. 2012. “The Accessible Chromatin Landscape of the Human Genome.” *Nature* 489 (7414): 75–82.
- Tishkoff, Sarah A., Floyd A. Reed, Alessia Ranciaro, Benjamin F. Voight, Courtney C. Babbitt, Jesse S. Silverman, Kweli Powell, et al. 2007. “Convergent Adaptation of Human Lactase Persistence in Africa and Europe.” *Nature Genetics* 39 (1): 31–40.
- Uhl, Juli D., Arya Zandvakili, and Brian Gebelein. 2016. “A Hox Transcription Factor Collective Binds a Highly Conserved Distal-Less Cis-Regulatory Module to Generate Robust Transcriptional Outcomes.” *PLoS Genetics* 12 (4): e1005981.
- Ulirsch, Jacob C., Satish K. Nandakumar, Li Wang, Felix C. Giani, Xiaolan Zhang, Peter Rogov, Alexandre Melnikov, et al. 2016. “Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits.” *Cell* 165 (6): 1530–45.
- Viñuelas, José, Gaël Kaneko, Antoine Coulon, Guillaume Beslon, and Olivier Gandrillon. 2012. “Towards Experimental Manipulation of Stochasticity in Gene Expression.” *Progress in Biophysics and Molecular Biology* 110 (1): 44–53.
- Visel, Axel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-frick, et al. 2009. “ChIP-Seq Accurately Predicts Tissue-Specific Activity of Enhancers.” *Nature* 457 (7231): 854–58.
- Walters-Sen, Lauren C., Sayaka Hashimoto, Devon Lamb Thrush, Shalini Reshmi, Julie M.

- Gastier-Foster, Caroline Astbury, and Robert E. Pyatt. 2015. "Variability in Pathogenicity Prediction Programs: Impact on Clinical Diagnostics." *Molecular Genetics & Genomic Medicine* 3 (2): 99–110.
- Wang, Jianlong, Sridhar Rao, Jianlin Chu, Xiaohua Shen, Dana N. Levasseur, Thorold W. Theunissen, and Stuart H. Orkin. 2006. "A Protein Interaction Network for Pluripotency of Embryonic Stem Cells." *Nature* 444 (November): 364–68.
- Wang, Jie, Jiali Zhuang, Sowmya Iyer, Xin-Ying Lin, Melissa C. Greven, Bong-Hyun Kim, Jill Moore, et al. 2013. "Factorbook.org: A Wiki-Based Database for Transcription Factor-Binding Data Generated by the ENCODE Consortium." *Nucleic Acids Research* 41 (Database issue): D171–76.
- Wang, Jie, Jiali Zhuang, Sowmya Iyer, Xinying Lin, Troy W. Whitfield, Melissa C. Greven, Brian G. Pierce, et al. 2012. "Sequence Features and Chromatin Structure around the Genomic Regions Bound by 119 Human Transcription Factors." *Genome Research* 22 (9): 1798–1812.
- Weingarten-Gabbay, Shira, and Eran Segal. 2014. "The Grammar of Transcriptional Regulation." *Human Genetics* 133 (6): 701–11.
- Weirauch, Matthew T., Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R. Riley, Julio Saez-Rodriguez, et al. 2013. "Evaluation of Methods for Modeling Transcription Factor Sequence Specificity." *Nature Biotechnology* 31 (2): 126–34.
- White, Michael A. 2015. "Understanding How Cis-Regulatory Function Is Encoded in DNA Sequence Using Massively Parallel Reporter Assays and Designed Sequences." *Genomics* 106 (3): 165–70.
- White, Michael A., Jamie C. Kwasnieski, Connie A. Myers, Susan Q. Shen, Joseph C. Corbo, and Barak A. Cohen. 2016. "A Simple Grammar Defines Activating and Repressing Cis-Regulatory Elements in Photoreceptors." *Cell Reports* 17 (5): 1247–54.
- White, Michael A., Connie A. Myers, Joseph C. Corbo, and Barak A. Cohen. 2013. "Massively Parallel in Vivo Enhancer Assay Reveals That Highly Local Features Determine the Cis-Regulatory Function of CHIP-Seq Peaks." *Proceedings of the National Academy of Sciences* 110 (29): 11952–57.
- Whyte, Warren a., David a. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard a. Young. 2013. "Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes." *Cell* 153 (2): 307–19.
- Wijchers, Patrick J., Peter H. L. Krijger, Geert Geeven, Yun Zhu, Annette Denker, Marjon J. A. M. Verstegen, Christian Valdes-Quezada, et al. 2016. "Cause and Consequence of Tethering a SubTAD to Different Nuclear Compartments." *Molecular Cell* 61 (3): 461–73.
- Williams, David C., Mengli Cai, and G. Marius Clore. 2004. "Molecular Basis for Synergistic Transcriptional Activation by Oct1 and Sox2 Revealed from the Solution Structure of the 42-kDa Oct1·Sox2·Hoxb1-DNA Ternary Transcription Factor Complex." *The Journal of Biological Chemistry* 279 (2): 1449–57.
- Xian, Hai-Qing, Kelly Werth, and David I. Gottlieb. 2005. "Promoter Analysis in ES Cell-Derived Neural Cells." *Biochemical and Biophysical Research Communications* 327 (1): 155–62.

- Xi, Hualin, Hennady P. Shulha, Jane M. Lin, Teresa R. Vales, Yutao Fu, David M. Bodine, Ronald D. G. McKay, et al. 2007. "Identification and Characterization of Cell Type-Specific and Ubiquitous Chromatin Regulatory Structures in the Human Genome." *PLoS Genetics* 3 (8): e136.
- Yie, J., K. Senger, and D. Thanos. 1999. "Mechanism by Which the IFN-Beta Enhanceosome Activates Transcription." *Proceedings of the National Academy of Sciences of the United States of America* 96 (Track II): 13108–13.
- Zentner, Gabriel E., and Peter C. Scacheri. 2012. "The Chromatin Fingerprint of Gene Enhancer Elements." *The Journal of Biological Chemistry* 287 (37): 30888–96.
- Zhang, Feng, and James R. Lupski. 2015. "Non-Coding Genetic Variants in Human Disease." *Human Molecular Genetics* 24 (R1): R102–10.
- Zhang, Xiaofei, Juan Zhang, Tao Wang, Miguel a. Esteban, and Duanqing Pei. 2008. "Esrrb Activates Oct4 Transcription and Sustains Self-Renewal and Pluripotency in Embryonic Stem Cells." *The Journal of Biological Chemistry* 283: 35825–33.
- Zhao, Yue, David Granas, and Gary D. Stormo. 2009. "Inferring Binding Energies from Selected Binding Sites." *PLoS Computational Biology* 5 (12): e1000590.
- Zhou, Harry Y., Yulia Katsman, Navroop K. Dhaliwal, Scott Davidson, Neil N. Macpherson, Moorthy Sakthidevi, Felicia Collura, and Jennifer A. Mitchell. 2014. "A Sox2 Distal Enhancer Cluster Regulates Embryonic Stem Cell Differentiation Potential," 2699–2711.
- Zhou, Hufeng, Stefanie C. S. Schmidt, Sizun Jiang, Bradford Willox, Katharina Bernhardt, Jun Liang, Eric C. Johannsen, et al. 2015. "Epstein-Barr Virus Oncoprotein Super-Enhancers Control B Cell Growth." *Cell Host & Microbe* 17 (2): 205–16.
- Zhou, Stanley, Aislinn E. Treloar, and Mathieu Lupien. 2016. "Emergence of the Noncoding Cancer Genome: A Target of Genetic and Epigenetic Alterations." *Cancer Discovery* 6 (11): 1215–29.
- Zhao, Yue, David Granas, and Gary D. Stormo. 2009. "Inferring Binding Energies from Selected Binding Sites." *PLoS Computational Biology* 5 (12): e1000590.  
<https://doi.org/10.1371/journal.pcbi.1000590>.