LINKING *CIS*-REGULATORY REGIONS USING TRANSCRIPTION FACTOR BINDING SIGNATURES

by

Yueming (Michelle) Kang

B.Sc., The University of British Columbia, 2018

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

Master of Science

in

The Faculty of Graduate and Postdoctoral Studies

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

June 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Linking *cis*-regulatory regions using transcription factor binding signatures

submitted by Yueming (Michelle) Kang in partial fulfillment of the requirements for

the degree of Master of Science

in Bioinformatics

Examining Committee:

Wyeth Wasserman, Professor, Medical Genetics, UBC

Supervisor

Carolyn Brown, Professor, Medical Genetics, UBC

Supervisory Committee Member

Nansheng Chen, Associate Professor, Molecular Biology and Biochemistry, SFU

Supervisory Committee Member

Abstract

Linking cooperatively functioning *cis*-regulatory elements (CREs), specifically enhancers and promoters, is a challenging task. Current strategies include correlation of expression of RNA transcribed from the CREs, experimentally measured chromatin interactions (Promoter Capture Hi-C) or machine learning based computational predictions. However, all three approaches require the availability of experimental data, which is sparse for most cells and tissues. We propose a new similarity metric to link enhancers to their target promoters based on transcription factor (TF)-binding “signatures”. TF-binding signatures are binary string representations (*e.g.* 0011001...), where each position indicates binding (“1”) or not (“0”) of a TF to a CRE. We apply a cosine similarity metric to enhancer-promoter pairs linked in published studies involving CRISPRi-FlowFISH, co-expression (FANTOM), or experimental tiling-deletion (CREST-seq). We find a significant difference between TF signature similarities of linked promoter-enhancer pairs compared to unlinked pairs. Furthermore we observe that TF-binding similarity scores are CRR specific. Based on the results, new directions are proposed that may allow further improvement towards a reliable mapping of interacting CREs across the genome.

Lay Summary

The human body is composed of many different cell types that all come from the same DNA instructions. Neurons, white blood cells and muscle cells look and function very differently. The properties of each cell type depends on which genes that are turned on or off. Within the DNA are on/off switches for the genes, but they are spread out and it is not yet known which on/off switches work on which genes. In this thesis a method is created and tested that aims to determine which on/off switches act on which genes. The method is based on comparing characteristics of the on/off switches with the characteristics at the start of the genes, under the expectation that these characteristics should be similar if they work together.

Preface

This thesis contains original work performed at the UBC Centre for Molecular Medicine and Therapeutics at the BC Children's Hospital Research Institute under the supervision of Dr. Wyeth Wasserman. No text is taken from previously published material.

The TF binding signature design protocol was defined by myself and Dr. Oriol Fornes. I programmed scripts creating TF binding signatures and calculating similarity scores between signatures, and performed downstream statistical analysis.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank my supervisor Dr. Wyeth Wasserman for giving me the chance to work on this project and for all of his patience, support and encouragement throughout my studies. Thank you to Dr. Oriol Fornes for the endless hours of discussions, suggestions and comments throughout my studies and for this thesis, Manu Saraswat for all of our computational discussions and collaborations, Dora Pak for schedule management and overall support and the rest of the Wasserman lab for many helpful discussions. Additionally thanks to Allen Zhang for providing helpful suggestions for this thesis and moral support throughout my studies. I would also like to thank Devon Graham for having answers to all of my programming questions and the countless coffees throughout this time. A special thanks to my committee members, Dr. Carolyn Brown and Dr. Jack Chen for their comments and suggestions throughout my studies.

Dedication

To my mom, my dad and my sister for always giving me so much love and support.

Introduction

While the human body is composed of many different cell types (*e.g.* neurons, fibroblasts, leucocytes), they all originate from the same DNA. The morphology and function of each cell type depends on the specific subset of genes that are turned on (*i.e.* expressed) or off (*i.e.* repressed). Gene expression is a complex cellular process that is regulated at multiple levels. At the transcriptional level, regulation of gene expression largely depends on the coordinated action of *cis*-regulatory regions (CRRs) and transcription factors (TFs)[1]. CRRs are regions of DNA that regulate the spatiotemporal expression of target genes. Broadly, CRRs include promoters, *i.e.* proximal regulatory regions overlapping transcription start sites (TSSs) of their target genes, and distal regulatory regions or enhancers. TFs are proteins that bind to these CRRs in a sequence-specific manner to promote or repress gene expression[2]. To do so, they stabilize/block the binding of RNA polymerase II (RNAP2) to DNA, promote the modification (*e.g.* acetylation/deacetylation) of histones (see below), and/or recruit coactivators/corepressors. Additional layers of transcriptional regulation include aforementioned histone modifications, and the local three-dimensional (3D) architecture of DNA. New technologies and experimental methods have emerged to enable high-throughput profiling of different mechanisms of transcriptional control, including the accessibility of chromatin[3], genomic locations where TFs bind to DNA and histones are modified [4], and the 3D conformation of the genome[5]. Together, they enable the identification of CRRs and TF binding sites (TFBSs; *i.e.* specific genomic locations where TFs bind).

While substantial progress has been made in identifying CRRs, linking enhancers to their target genes remains challenging. Current strategies include correlating data across individual cells[6] or across multiple cell types and/or tissues, identifying chromatin interactions

between promoter and distal regions (*e.g.* Promoter Capture Hi-C[6,7]), or disrupting candidate enhancer regions by CRISPR technology. More recently, this task has caught the attention of computational biologists in the field of machine learning [8,9]. The development of machine learning and other computational methods is important. Experiments can only identify enhancer-promoter pairs in the cells and/or tissues analyzed and, for most of them, this data is sparse. However, current machine learning methods have limited performance and predicted enhancer-promoter pairs must still be experimentally validated[10].

Given the limitations of current predictive methods, **developing a method that correctly links enhancers to their target genes is key to furthering the current understanding of gene regulation.**

In the following sections I will describe the current state of understanding and research methods that inform the approaches taken in this thesis.

1.1 Transcription factors

Formally, the term transcription factor refers to proteins involved in the process of transcription. In this thesis, I focus on the subset of TFs that exhibit sequence-specific DNA-binding properties, and hereafter, the term TF will refer specifically to this subset. The initiation of transcription begins with a TF binding to its cognate site, followed by the sequential recruitment of coactivator proteins and ultimately RNAP2. Note that there is regulation of other RNAPs, for instance RNA polymerase III, which synthesizes small RNAs such as 5S rRNA and tRNA, but within this thesis I focus on RNAP2-mediated transcription. TFs are modular proteins composed of one or more functional domains. DNA-binding domains enable recognition and binding to short (6-20 bp) DNA motifs, trans-activating domains allow interactions between TFs

and transcriptional coactivators/corepressors, and signal-sensing domains respond to external stimuli to increase or repress expression of the transcribed gene. Activators (*i.e.* TFs that promote transcription) interact with other TFs and coactivator proteins to recruit RNAP2 for transcription initiation. In contrast, repressors inhibit RNAP2 from initiating transcription by blocking the binding of coactivators or impeding RNAP2 progress along the DNA strand. Depending on different protein-protein interactions and/or protein modifications, some TFs can reverse their role acting as an activator to function as a repressor. For example, MYC is a regulator for the transcription of ~15% of human genes. While MYC acts as an activator for essential genes involved in cell proliferation, cell growth and metabolism, it can also interact with other TFs (*e.g.* SP1 and MIZ1) to repress the expression of genes involved in the negative regulation of cell proliferation [11].

Each TF recognizes diverse DNA-binding sites, but only a few of these TF binding sites (TFBSs) are occupied *in vivo*. TF binding sites can be conserved or change across cells and tissues based on the varying affinity of TF-DNA interactions to specific DNA sequences[12]. CTCF, a ubiquitously expressed TF that can function as an activator, repressor or insulator, has relatively conserved binding sites. On the other hand, MYC can bind in a cell type-specific manner to lower affinity TFBSs, facilitating cell type-specific gene expression[13]. The difference in TF binding affinity between a variety of TFBSs is especially important during development, for instance to control spatiotemporal gene expression. Some TFs are constitutively active in all cell types, while others are expressed gradiently in specific timepoints or cell types and are important for development [14,15]. In addition, some TFs preferentially bind within promoter regions while others are enriched at enhancers [16, 17, 18]. Sp1, a TF involved in many cellular processes including cell differentiation, primarily binds within promoter regions while growth-factor-inducible TFs (AP-1) binding is enriched at enhancers [16, 19] .

1.1.1 Identification of TF binding sites

Development of high-throughput technologies to detect protein-DNA interactions *in vitro* and *in vivo* has enabled the identification of TFBSs. *In vivo* techniques measure TF-DNA interactions in the context of cellular chromatin, while most *in vitro* techniques measure TF interactions with "naked" DNA. The *in vitro* methods are powerful ways to understand the diversity of DNA sequences bound by a given TF. Since TF binding is affected by DNA accessibility and cofactors that induce conformational changes in the DNA, *in vitro* techniques are unable to accurately capture the high affinity binding sites that TFs preferentially bind to *in vivo* [20]. In the context of this thesis, we will therefore focus on chromatin immunoprecipitation followed by sequencing (ChIP-seq), an *in vivo* method to identify TFBSs.

ChIP-seq has become a standard technique to delineate the genomic locations where TFs bind to DNA. In a ChIP-seq experiment, proteins are first crosslinked to their interacting DNA. Next, DNA is fragmented and immunoprecipitated with antibodies targeting the TF of interest. Then, immunoprecipitated crosslinks are reversed and the purified DNA is sequenced and mapped back to genome coordinates using high-throughput methods. Finally, TF-bound regions are determined by peak-calling algorithms that identify genomic regions enriched in sequenced DNA fragments. Note that ChIP-seq can be applied to other DNA-binding proteins (e.g. RNAPII, p300) or to detect histone modifications (see section 1.2.3.1). Efforts have been made by consortia such as ENCODE [21,22] to generate collections of ChIP-seq datasets for hundreds of TFs in a variety of biological samples including tissues, primary cells, and immortalized cell lines. The ReMap database [22] is a public repository of human and *Arabidopsis* regulatory region data assembled through uniformly analyzing thousands of quality controlled, publically available, ChIP-seq experiments. The 2020 human version of the ReMap includes

5,798 high quality ChIP-seq datasets profiling regulatory regions covering a total of 1,135 TFs across 602 cells and tissues.

1.2 Cis-regulatory regions

1.2.1 Promoters

By definition, promoters are regulatory regions of DNA located proximal to and overlapping the TSSs of genes, and are essential for transcription of DNA to RNA. Each gene promoter must include TFBSs, allowing for assembly of the transcription machinery and recruitment of RNAP2. Genes commonly have multiple promoters [23]; the FANTOM5 project showed that more than 6,000 genes are regulated by multiple promoters [24]. Detailed characterization of promoters have shown that some promoters use one specific TSS, while other promoters produce transcripts from a variety of TSSs[25]. However, in eukaryotes, about a quarter of promoters include a conserved sequence called the TATA-box [26]. It has been shown *in vitro* that the TATA-binding protein (TBP) binds upstream of the TATA-box and is sufficient to initiate transcription of these genes [27]. Promoters that do not contain the TATA-box (TATA-less promoters) frequently contain other elements that allow general TFs to bind. The initiator element (Inr) and the downstream promoter element (DPE) are regions where general TFs, such as TBP Associated Factors (TAFs), are able to bind and interact with TBP to initiate transcription. [23]. Promoters featuring one specific TSS commonly have strong TATA-box motifs, while promoters using a diverse set of TSSs do not[25].

1.2.2 Enhancers

For many genes, the precise control of gene expression requires additional regulatory sequences beyond the promoter region. In such cases, enhancer-promoter interactions may be required for the precise control of gene expression levels [28]. By definition, enhancers are distal regulatory regions containing TFBSs involved in the spatio-temporal control of gene expression. Enhancers are frequently located in non-coding or intronic regions of the genome, and can be located hundreds of kilobase pairs upstream or downstream of their target genes[29]. Enhancers contribute to the expression of their target genes in a synergistic and partly redundant manner through chromatin looping facilitated by cohesin and other protein complexes. Such looping brings enhancers to the proximity of their target promoters in 3D-space[28].

Both promoters and enhancers have been shown to produce RNA transcripts (both transcribed by RNAP2). The FANTOM5 project found that enhancer RNA (eRNA) transcripts are bidirectionally transcribed [30], while promoter RNA transcripts have directional bias[24]. Most eRNAs are short (generally less than 350 bp), unstable and unspliced. In contrast, sense-strand promoter RNA transcripts are long (on average ~1,200 bp), and 80% of them are spliced [30].

Recent experimental evidence contradicts these classifying characteristics, as it has been shown that promoters are also bidirectionally transcribed, producing antisense promoter upstream transcripts (PROMPTs) that resemble eRNAs[31]. In addition, experimental evidence supports that some promoters can act as enhancers for nearby genes [32,33].

Thus it is not inappropriate to consider enhancers and promoters to be labels describing two ends of a continuous functional distribution of CRRs. In the context of this thesis, however, promoters are classified as regulatory regions overlapping the TSS(s) of a gene, and enhancers are distal regulatory regions identified by any of the methods discussed below.

1.2.3 Identification of cis-regulatory regions

Over the last decade, new experimental methods have been developed to enable the identification of CRRs. These methods include profiling genomic locations where histones are modified (*i.e.* ChIP-seq[4]), chromatin accessibility (*e.g.* DNase-seq[3]), and RNA transcription initiation (*e.g.* CAGE[34] ,GRO-seq[35]).

1.2.3.1 Histone modifications

In the nucleus, DNA is coiled around proteins called histones to facilitate genome organization. Histone proteins are subject to diverse post-translational modifications such as the addition of a methyl group (methylation) or an acetyl group (acetylation), at specific amino acids. Ultimately, such modifications dictate the accessibility of the DNA region for TF binding[36]. Genome-wide analysis of histone modifications identified certain modifications as markers of enhancer and TSS regions, respectively. For example, the tri-methylation (Me3) at lysine 4 (K4) of histone H3 (*i.e.* H3K4Me3) marks regions proximal to TSSs[37], while H3K4me1 marks enhancer regions, and the presence of H2K27ac distinguishes active from inactive enhancers. Histone modifications can be detected using methods such as ChIP-seq (see 1.1.1 Identification of TF Binding Sites). Over 3,000 uniformly processed histone ChIP-seq datasets, conducted on a variety of cell types and tissues, are available through the ENCODE Portal[38].

1.2.3.2 Chromatin accessibility

Since accessible DNA regions facilitate TF binding, they are also susceptible to cleavage by DNase I, an endonuclease with little sequence specificity [39]. DNase I hypersensitive sites (DHSs) can be identified experimentally using DNase-seq, a technique that involves digesting DNA using DNase I followed by sequencing of the cleaved regions. In 2012, the

ENCODE project performed DNase-seq on 125 different human cell and tissue types and identified 2,890,742 DHSs ~150 bp long. Downstream analysis of the identified DHSs found that 97.4% of 1,046 experimentally validated enhancer regions overlap a DHS. In addition, it was discovered that DHSs overlapping promoter regions are relatively conserved across cell types, while DHSs overlapping enhancer regions are more cell type-specific [40]. Today, over 1,000 DNase-seq datasets covering hundreds of human cell and tissue types are available through the ENCODE Portal.

1.2.3.3 Transcription initiation

Enhancer and promoter regions are transcribed into capped RNAs by RNAP2. Introduced in 2003 by Shiraki *et al.* [34], cap analysis of gene expression (CAGE) captures the capped 5' region of transcribed mRNAs and eRNAs. These captured transcripts (*i.e.* tags) are then sequenced and mapped back to a reference genome. The Functional Annotation of Mammalian Genome (FANTOM) consortium (http://fantom.gsc.riken.jp/) both generated and analyzed an extensive collection of CAGE data across the majority of human cell types, organs and immortalized cell lines to identify TSSs and enhancer regions. TSSs were identified by clustering CAGE tags that were strongly biased towards the sense direction of the gene, while enhancers were identified as regions enriched with bidirectional CAGE peaks [24],[30]. At the time of publication, the FANTOM5 project included CAGE data for 573 primary human cells, 152 human post mortem tissues and 250 cancer cell lines. From the CAGE data, the FANTOM5 project identified TSSs for 91% of human protein coding genes [24]and 43,011 enhancers[30]. While some are expressed ubiquitously, many of these identified enhancers are expressed in a cell type-specific manner[30].

Global run-on sequencing (GRO-seq) [35]captures nascent transcription from actively engaged RNA polymerases. While eRNA instability inhibits CAGE from detecting many enhancer regions, GRO-seq measurements are independent of the instability and high decay rate of eRNAs, resulting in improved sensitivity[41]. A recent improvement to GRO-seq's nuclear run-on based method, PRO-seq, replaces the classically used bromouridine substrate with biotin-labeled nucleotide triphosphates (biotin-NTPs) to achieve base-pair resolution [41]. Active enhancer and promoter regions can be identified from raw GRO-seq/PRO-seq data using software such as dREG, a machine learning tool that predicts active regulatory regions based on support vector regression. Due to expensive, time-consuming experimental procedures, the number of publically available GRO-seq/PRO-seq datasets is limited and sparse[42].

1.3 Linking enhancers to promoters

While identifying CRRs has become increasingly possible (both computationally and experimentally), linking enhancers to their target promoters remains challenging[43]. Enhancers can skip the nearest gene to regulate a more distal one, and the genomic distance between enhancers and promoters can be quite large[44]. Current strategies include correlation of DHS or expression data across multiple cells and/or tissues (*e.g.* CAGE data from the FANTOM5 project[30]), chromatin interactions (*e.g.* Promoter Capture Hi-C[7]), CRISPR perturbations and machine learning methods, each with underlying limitations. Linking CRRs by correlation requires data from many cell types and has a low accuracy for rarely expressed genes. While the identification of distal regions that interact with promoters is important in identifying enhancers, current Promoter Capture Hi-C methods are expensive (e.g. require a large number of cells) and have low resolution[7]. As an alternative to the physical interaction measured by chromatin capture methods, CRISPR-based methods allow detection of genetic interactions.

Emerging methods that rely on CRISPR technologies are limited by low throughput and are restricted to the subset of genes that can be assayed. Finally, machine learning methods often have inflated performance measures and require experimental validation to identify truly linked CRRs[45].

1.3.1 Correlation

Thurman *et al.*[46] observed that known cell type-specific enhancers become DNase hypersensitive synchronously with their target gene's promoters. Correlation analysis (using a simple Pearson correlation) between 1,454,901 DHSs and all promoters within 500 kilobase pairs (kb) of each DHS across 79 cell types identified in the ENCODE project was performed, resulting in 1,595,025 DHS-promoter linkages [46]. Assuming that each DHS is a candidate enhancer, this analysis found that on average a promoter is correlated with 22.8 enhancers, with 84% of promoters correlated with more than one enhancer. Moreover, DHS-correlated enhancer-promoter pairs were enriched for chromatin interactions [46].

In addition to the identification of CRRs, the FANTOM5 project linked enhancers to their target genes based on correlation of CAGE data. The Pearson correlation coefficient was calculated between CAGE tags per million (TPMs), across available samples, of all intra-chromosomal enhancer-TSS pairs within 500 kb; highly correlated enhancer-TSS pairs ($r > 0.7$) were identified as putative enhancer-TSS pairs. This method of linking enhancers to TSSs resulted in TSSs being associated to 4.9 enhancers each on average, with enhancers being linked to 2.4 TSSs each on average. Linking enhancers to promoters by correlation of expression appeared to be substantially more concordant than correlation of DHS after validation against ChIA-PET (RNAP2-mediated) interaction data from the ENCODE consortium

(20.6% vs 4.3% at the same threshold) [30]. This suggests that not all regions accessible to TFs are transcribed enhancers.

1.3.2 Promoter Capture Hi-C

It has been observed that distal enhancers are aided by DNA looping to control the expression of their target promoters. Consequently, Promoter Capture Hi-C (PCHi-C), an assay that captures long-range interactions of promoter regions, can be used to identify enhancers for target promoters[7]. In PCHi-C, interacting regions of the genome are crosslinked and digested to generate di-tags. Di-tags encompassing promoter regions are captured by specific RNA baits and sequenced. Misfud *et al.* used PCHi-C to identify distal interacting regions for 21,841 promoters in GM12878 and CD34+ cells. Distal regions found to interact with active promoters were enriched in DHS and enhancer-associated histone marks such as H3K4me1, H3K4me3, and H3K27ac. Furthermore, DNA fragments interacting with inactive or weakly transcribed promoters were enriched for H3K27me3, a repressing mark, and depleted for the activating marks present in fragments interacting with active promoters. The DNA regions interacting with promoters were cell type dependent. For instance, GM12878 cell type-specific enhancers were enriched in the promoter-interacting fragments of the GM12878 PCHi-C dataset.

1.3.3 CRISPR-based methods

In the classic CRISPR-Cas9 assay, cells expressing Cas9 are infected with a viral library of guide RNAs. Then, Cas9 cleaves DNA regions that are complementary to the guide RNA resulting in double strand breaks. Next, the breaks are joined either by a donor template or by the cell's double strand break repair machinery, which frequently introduces mutations that affect function at the target locus.

CRISPRi-FlowFISH[47] is a method that utilizes CRISPR technology to identify enhancer regions of target genes. KRAB-dCas9, a nuclease-deficient Cas9 bound to an inhibitor, is expressed in cells instead of Cas9 to minimize the varying effects of mutations. Cells expressing fluorescently tagged target genes are infected by a guide RNA library designed to target candidate enhancer regions (*i.e.* to deliver the KRAB inhibitor to the enhancer region). Then, RNA fluorescence *in situ* hybridization (FISH) followed by fluorescence-activated cell sorting (FACS) is performed to label and bin cells with different expression levels of the target gene. The effect of each guide RNA towards gene expression can be inferred after high-throughput sequencing of the resulting bins, and enhancer regions can be identified as guide RNAs that significantly decrease the target gene's expression.

While candidate enhancer regions are a prerequisite for designing the CRISPRi-FlowFISH guide RNA library, CRR scan by tiling-deletion and sequencing (CREST-seq) do not require such a prerequisite. In a CREST-seq experiment, cells expressing Cas9 are infected with a guide RNA library that introduces a large number of overlapping genomic deletions (~2 kb deletions, each overlapping by 1.9 kb). Then, FACS is performed to isolate cells with lowered expression of the target gene. Next, deletions resulting in lowered expression of the target gene are determined by high-throughput sequencing. Finally, enhancer regions are inferred from regions enriched in guide RNA-facilitated deletions.

1.3.4 Machine learning methods

Supervised machine learning methods have emerged as an alternative to experimental approaches to predict enhancer-promoter linkages in a cell type-specific manner. These methods rely on the analysis of integrated genomics data of candidate CRRs. Despite some

initial promise, independent benchmarking [10] showed that the performance of computational approaches are often little better than a simple pairing of enhancers with the closest promoters. TargetFinder is a popular tool that claims to accurately link enhancers to their target promoters based on distinguishing patterns of TF binding, histone modifications and DHSs between interacting and non-interacting enhancer-promoter pairs [8]. The method reported a false discovery rate (probability of false positive enhancer-promoter pairs) 15 times lower than the false discovery rate of linking enhancers to the most proximal promoter. Further analysis of the training and testing sets of TargetFinder by *Cao et al.* found that 53–76% of regions between interacting enhancer-promoter pairs overlapped with each other, while only 0.16% of regions between interacting enhancer-promoter pairs overlapped a window of a non-interacting enhancer-promoter pair[45]. After correcting for this bias, *Cao et al.* found that the accuracy of TargetFinder in predicting enhancer-promoter pairs decreased dramatically from 77-90% to 1.3-9.8% in the 6 different cell lines studied[45]. Furthermore, *Moore et al*.[10]compared TargetFinder's performance to a distance-based approach (*i.e.* linking enhancers to the closest promoter) and found that TargetFinder only slightly outperformed the distance-based approach when trained and tested on the same cell line, but performed worse than the distance-based approach when a trained model was tested on data from a different cell line.

PEP-motif[9](predicting enhancer-promoter interactions) is a machine learning tool that predicts enhancer-promoter pairs from sequence-based features (*i.e.* motifs). For each cell line (K562, GM12878, HeLa-S3, HUVEC, IMR90, NHEK), TF binding motifs of enhancers and promoters in TargetFinder's cell line-specific positive and negative sets are concatenated and used to train a supervised model. While PEP-motif reports a similar performance to TargetFinder with a weighted average precision and recall (F1) accuracy of 77-90%, *Moore et al.*[10] found that PEP-motif performs worse than the distance-based method. PEP-motif achieves

an area under the precision recall curve (AUPRC) of 0.3 while the distance-based method achieves an AUPRC of 0.43 after retraining the PEP-motif model with unbiased training and testing sets.

While machine learning methods for predicting enhancer-promoter interactions are anticipated to be impactful, current performance metrics indicate there is substantial opportunity for improvement.

1.4 Hypothesis

In depth analyses of the regulation of specific genes have revealed cases in which the same TF functionally binds to motifs within both the gene's promoter region and a distal enhancer region, such as the TF HNF1 in the GC gene[48]. Since enhancers act in conjunction with promoters to regulate gene expression, and because TFs recruit RNAP2 and accessory factors to promote transcription at these CRRs, we hypothesize that overlapping patterns of TF binding will be enriched in cooperatively functioning pairs of enhancers and promoters. Such enrichment would provide a mechanism to predict links between enhancers and their target promoters.

Past approaches based on the hypothesized relationship between TF binding patterns at enhancers and promoters have performed poorly, potentially due to the data features used. We suspect that the low performance of PEP-motif is due to using computationally predicted, rather than experimentally identified, TFBSs, especially since DNA accessibility is not considered in the approach. In addition, we hypothesize that TF binding similarity between enhancers and their target promoters is TSS-specific, with a signal that cannot be sufficiently captured in the global manner that PEP-motif is trained.

We have developed a correlation metric to link enhancers to promoters based on TF binding signatures (Figure 1). The signature is a binary string representation (*e.g.* 0011001...), where each position indicates the binding (“1”) or not (“0”) of a TF at the CRR. The correlation is calculated as the cosine similarity between signatures. We hypothesize that TF binding signatures will be more similar between enhancers and their target promoters compared to enhancers and promoters they do not target.

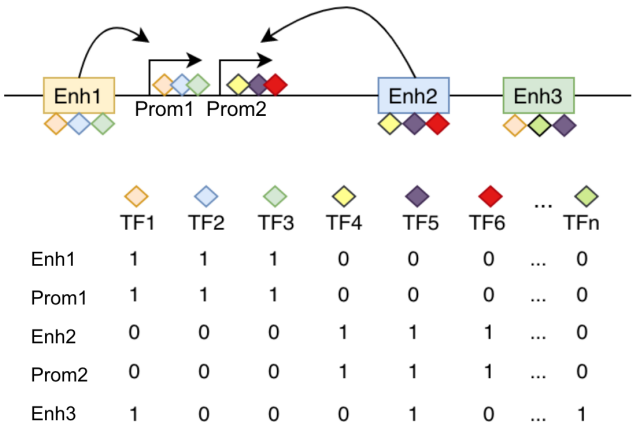


Figure 1. Schematic overview of enhancer-promoter linkage using TF binding binary vectors. The figure depicts a segment of DNA within which are observed two promoter regions (Prom1 and Prom2) delineated with right-angle arrows. In addition to these promoter regions, there are three distinct enhancer regions (Enh1, Enh2, and Enh3) indicated by coloured boxes. Curved arrows indicate functional roles of an enhancer acting upon a promoter. Coloured diamonds indicate specific types of TFs binding to the DNA at each region. Below a data matrix depicts the observed experimental ChIP-seq data that reports the binding of a TF to the DNA, where a 1 indicates observed binding in the cell type of interest, and a 0 indicates no observed binding.

Methods

2.1 Experimental data sources

Experimental data was obtained from published manuscripts, spanning several experimental approaches. CRISPRi-FlowFISH data for cell line K562, derived from bone marrow lymphoblasts, was obtained from[10]. CAGE-based data (cell lines K562, MCF-7 from breast epithelium, Hep-G2 from liver, and GM12878 from B-lymphocytes), including TSSs, enhancers and their linkage, were obtained from the FANTOM5 Consortium [30]; [24]. CREST-seq data (cell line H1 hESC) was obtained from[33]. Transcription initiation regions (TIRs) identified by dREG [42] were provided by Drs. Charles Danko and Zhong Wang (cell lines K562, MCF-7, and GM12878). TF binding data (cell lines K562, MCF-7, Hep-G2, GM12878 and H1 hESC) were obtained from ReMap2018[49].

All data is publicly available and relates to the build 37 of the Genome Reference Consortium human genome (hg19).

2.2 Software utilized

Bedtools (version v2.28.0) (https://bedtools.readthedocs.io/en/latest/index.html#) was used to overlap ChIP-seq-identified TFBSs with candidate CRRs. Python (version 3.7) with the *biopython* and *scipy* modules were used to generate TF binding signatures for each CRR and compare TF binding signatures between CRRs.

R (version 3.6.0) with the *Tidyverse* (https://joss.theoj.org/papers/10.21105/joss.01686) library was employed to perform statistical analyses and visualize TF binding similarities between CRRs.

2.3 Defining promoter regions

Promoter regions vary in size among genes, which makes the task of setting a constant promoter size nontrivial. Machine learning analysis of chromatin modification data in the 4 kb region centered around a TSS determined that most activating histone modifications are observed within the 2 kb region centered at each TSS[50]. Through positional analysis of ChIP-seq data, it was found that TFs binding sites were present in high concentration within the 300 bp upstream of the TSS and more uniformly spread across the 8 kb upstream and 2 kb downstream of regions tested [51]. Furthermore, GENCODE [52], which is part of the ENCODE project, and the Benchmark of candidate Enhancer-Gene Interactions (BENGI)[10] define promoters as 2 kb regions centered at TSSs. After extensive literature analysis, we decided to adopt the definition of promoter region used by GENECODE and BENGI (*i.e.* 2 kb region centered at the TSS). However, 58% of CAGE peaks generated by the FANTOM5 project occur in promoters with multiple TSSs[24]. Thus, the choice for the TSS upon which to center the promoter region is, again, nontrivial. The FANTOM5 project identified differentially regulated TSSs, some with proximity within 100 bp from each other, for 91% (94% at the permissive threshold) of human protein coding genes. TSSs were ranked and numbered sequentially by the total number of CAGE tags covering the region (for example *p1@RELA* corresponds to the TSS of *RELA* with the largest tag support); highest ranked TSSs will now be referred to as the strongest TSSs for a particular gene. In our analysis, **the promoter region of**

each gene was generated by centering a 2 kb region around the strongest TSS of that gene.

2.4 Defining enhancer regions

Positional specificity analysis of TF binding shows conserved patterns of TF motifs within enhancer regions[53]. Grossman *et al.* [53] interrogated 400 bp regions centered at nucleosome-depleted regions with strong DHS signals, and found that cell type-specific TF motifs are concentrated near the DHS signal peak, and that the majority are distributed between -100 bp and +100 bp from the peak max (*i.e.* the nucleotide position within the peak with the most mapped reads)[53]. Furthermore, the FANTOM5 project identified enhancer regions based on divergent transcription events. They observed that reverse and forward strand transcription initiation correspond to nucleosome boundaries and are separated by 180 bp on average [30]. Since TF binding is strongest at nucleosome-depleted regions, **we define enhancer regions to be 200 bp in length,** and adapt this definition to data type dependent requirements as discussed below.

2.4.1 Enhancer regions in CRISPRi-FlowFISH data

In the published CRISPRi-FlowFISH analysis, 30 genes located in five genomic regions ranging from 1.1Mbp to 4.0Mbp in length were screened for enhancer regions. All ENCODE DHSs within the 450 kb region surrounding each target gene (108-277 DHSs per gene totaling 884 unique DHSs) were expanded by 175 bp on each side, resulting in 500 bp candidate enhancer regions for testing. A total of 127 significant enhancer-gene linkages were identified, covering 93 of the 884 unique enhancer regions. Figure 2 shows that multiple enhancers were identified for the majority of target genes, and that the majority of identified enhancers showed

functional links to only one gene. Our analysis included all 884 unique candidate enhancers[47], but **the enhancer region was restricted to the central 200 bp** (to be consistent with the size of the FANTOM enhancers).

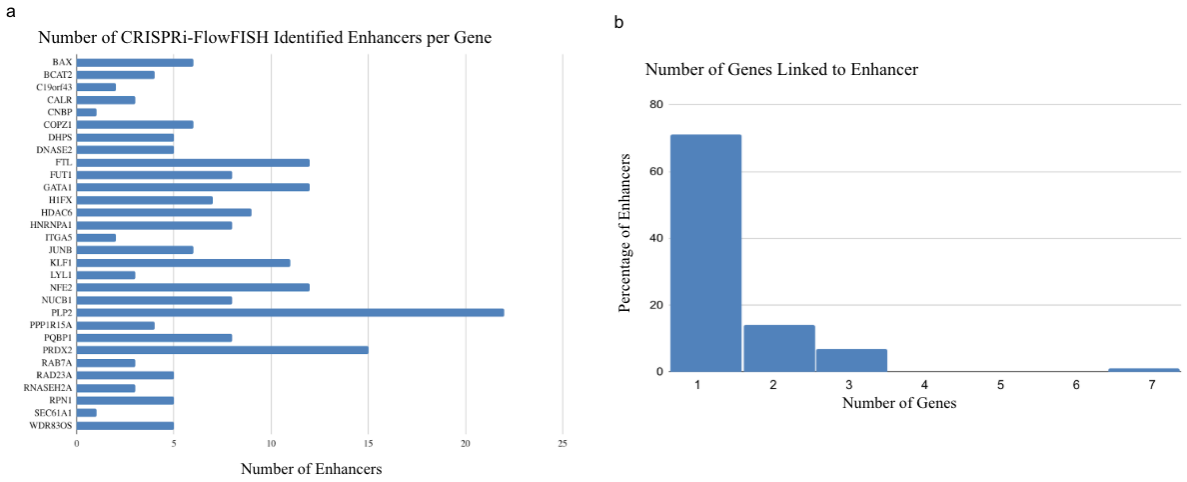


Figure 2. Overview of significant enhancer-promoter interactions identified by CRISPRi-FlowFISH. A reference data collection obtained from[47] was generated using the CRISPRi-FlowFISH technology. The histograms presented here depict: **a)** the number of identified enhancers reported per gene (where the genes are listed along the Y-axis); and **b)** The percentage of genes which are observed to be influenced per enhancer.

2.4.2 Enhancer regions in FANTOM5 CAGE data

The FANTOM5 project observed that while promoter RNAs are enriched in the sense direction, eRNAs display a pattern of similar transcript levels in both directions. On the basis that enhancer regions are bidirectionally transcribed, the FANTOM5 project identified 43,011 candidate enhancers across 808 human CAGE datasets covering 573 primary human cells, 152 human post mortem tissues and 250 cancer cell lines. Further analysis of these enhancers revealed that 89% are supported by ENCODE DHSs and 71% are supported by enhancer-marking histone modifications such as H3K4me1 or H3K27ac. Active enhancers in

K562, GM12878, MCF-7 and Hep-G2 cells were included in our analysis. Based on supporting evidence that weak enhancers (*i.e.* not displaying strong eRNA signals) are not enriched in TF binding[42], we generated two additional subsets for each set of FANTOM5-identified cell type-specific enhancers to filter out weak enhancers. The first subset excludes enhancers that have not been linked to a TSS in the FANTOM5 project. In addition, we generate a second subset of CAGE-identified enhancers for GM12878 and MCF-7 that are also supported by PRO-seq data [42] by overlapping FANTOM5-identified enhancers with dREG-identified TIRs. dREG[42] is a machine learning tool that identifies active cell type-specific TIRs from nascent transcription sequencing technologies. Figure 3 shows the differences in TIR sizes between cell lines. Both the GM12878 and MCF-7 TIRs were determined based on PRO-seq experiments and have similar mean lengths of 521 bp and 522 bp respectively. As the K562 TIRs were determined based on lower resolution GRO-seq experiments and have a mean TIR length of 2,878, we decided not to use them for this study. Table 1 summarizes the size and average enhancer length (defined by FANTOM5) of each cell type-specific enhancer set. The central 200 bp region of each enhancer was used in our analysis.

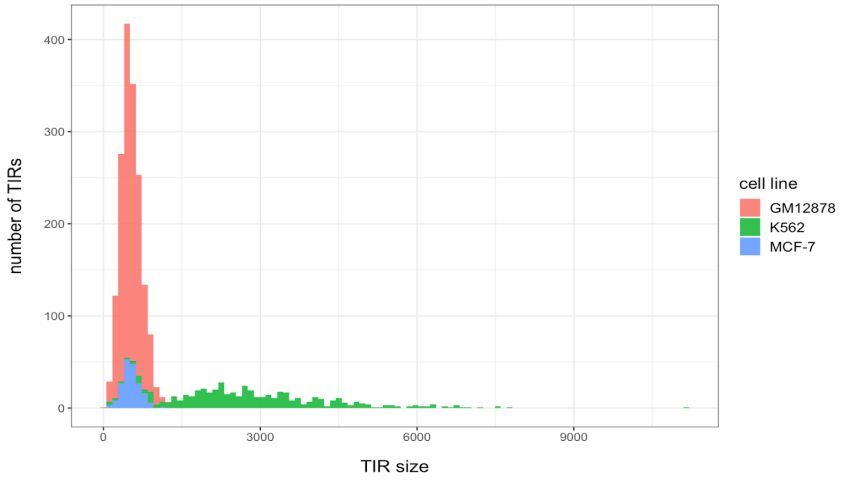


Figure 3. Overview of dREG-identified TIR lengths. dREG-predicted TIRs were provided by Drs Charles Danko and Zhong Wang (Cornell University, Ithaca, New York, USA), based on the method described in[42]. The histograms presented here depict: the distributions of TIR sizes for

each cell line where TIR lengths are along the X-axis and the number of TIRs for each length are on the Y-axis.

Cell Type	Total FANTOM5 Enhancers	Promoter-linked Enhancers	dREG peak overlapping Enhancers	Average FANTOM5 Length (bp)
K562	6,925	2,350	-	320
GM12878	12,783	4,499	1,463	326
MCF-7	3,151	1,161	189	330
Hep-G2	8,425	2,865	-	320

Table 1. Summary of CAGE identified enhancers. The number of enhancers in each cell type-specific set are denoted. The number of dREG overlaps was determined for Total Enhancers. K562 enhancers overlapping dREG peaks are excluded from our analysis due to the low resolution of dREG K562 TIRs. Hep-G2 enhancers overlapping dREG peaks are unavailable due to lack of PRO-seq data for HepG2 cells.

2.4.3 Enhancer regions in CREST-seq data

The published CREST-seq study was performed on human embryonic stem H1 cells (H1 hESC) to identify enhancer regions of the *POU5F1* gene. Within the 2Mbp region centered at the *POU5F1* gene, adjacent genomic regions averaging 2 kb and overlapping by 1.9 kb were deleted using CRISPR technology. An enrichment analysis was performed to identify sgRNAs resulting in deletions that significantly decreased *POU5F1* expression. A set of 44 enhancers, along with the promoter of *POU5F1*, were identified. CREST-seq-identified enhancer regions average 2,956 bp in length, and a majority (69%) are supported by ENCODE DHSs. Furthermore, identified enhancer regions are enriched in activating histone modifications H3K27ac (22%), H3K4me3 (31%), and H3K4me1 (22%), and depleted in repressive modifications H3K9me3 (6.7%) and H3K27me3 (6.7%). Due to the large size of CREST-seq-identified enhancers we generated multiple overlapping enhancer regions for each

of them. We split the 2Mbp region centered at *POU5F1* into 200 bp windows with a step-size of 100 bp using Bedtools[54] *makewindows*. Each window was overlapped with all CREST-seq-identified enhancer regions using Bedtools *intersect*, and all windows overlapping a CREST-seq enhancer were labelled as enhancer regions in our analysis.

2.5 Associating transcription factor ChIP-seq data to CRRs

TF ChIP-seq datasets for K562, GM12878, Hep-G2, MCF-7 and H1 hESC cells were obtained from ReMap[49]. Table 2 shows the number of ChIP-seq datasets and TFs covered for each cell line. All TF binding data was aggregated for each cell line and overlapped with candidate enhancer and promoter regions. TF binding was associated with enhancer regions based on the percentage overlap of the enhancer regions (200 bp) with the TF ChIP-seq peak (ranging from 50-300 bp). A TF was defined as binding an enhancer if >50% of the enhancer region overlapped a ChIP-seq peak for that TF. For promoter regions (2,000 bp), a TF was defined as bound if at least one ChIP-seq peak for that TF was completely encompassed in the promoter region.

Cell Type	Number of TFs	Number of Datasets
K562	204	530
GM12878	110	186
Hep-G2	103	287
MCF-7	85	142
H1 hESC	31	78

Table 2. Summary of the number of available human TF ChIP-seq datasets for each cell type. TF ChIP-seq datasets were obtained from ReMap2018[49]. The number of unique TFs and experimental datasets for each cell type are denoted.

2.6 Transcription Factor Binding Signatures

TF binding signatures are binary string representations (*e.g.* 0011001...), where each position indicates the binding (“1”) or not (“0”) of a TF to a CRR (or candidate CRR). The length of a signature differs between cell types and depends on the availability of cell type-specific TF ChIP-seq data (see Table 2), ranging between 31 in H1 hESC to 204 in K562 cells. For each cell type, the TF represented at each position is kept constant to enable comparisons between enhancer and promoter signatures.

2.7 Comparing signatures between enhancers and promoters

TF binding similarity between enhancer and promoter TF binding signatures was computed by means of cosine similarity, which measures the cosine of the angle between two non-zero vectors:

$$cos(\theta) = \frac{A \cdot B}{||A||\,||B||}$$

Where A and B are the TF binding signatures being compared. Cosine similarities were calculated using the “*distance.cosine”* function of the Python’s “*scipy”* package[55]. The cosine similarity between two signatures ranged between 0 and 1, with 1 indicating two identical signatures.

We used Z-scores to normalize cosine similarities to account for the variation in the number of TF binding events across promoters and enable comparison across different genes:

$$Z = \frac{x-\mu}{\sigma}$$

Where μ is the mean of all cosine similarities for a given promoter and σ the standard deviation. Z-scores were calculated using the "*stats.zscore*" function of the Python's "*scipy*" package.

Results

3.1 Positive and Negative Sets of enhancer-promoter pairs

To compare the TF binding similarities between linked and unlinked enhancer-promoter pairs, we established a set of experimentally supported enhancer-promoter pairs (*i.e.* “positive set”), and a corresponding “negative set” of enhancers which were assigned to genes that they do not regulate (*i.e.* unlinked). We generated three different positive and corresponding negative sets based on the *bonafide* enhancer-to-gene linkages identified by CRISPRi-FlowFISH and CREST-seq, and on the expression correlation of CAGE tags from FANTOM5.

3.1.1 CRISPRi-FlowFISH Positive and Negative Sets

The authors of the CRISPRi-FlowFISH dataset [47] perturbed hundreds of K562 DHSs (*i.e.* candidate enhancers) with the KRAB-dCas9 system, and quantified their effects on the expression of a target gene by RNA fluorescence *in situ* hybridization (FISH) and flow cytometry. Enhancers were defined as 200 bp regions centered on experimental K562 DHSs. For the target genes, the promoter was defined as the 2,000 bp region centered at the strongest TSS. KRAB-dCas9-inhibited enhancers that significantly decreased the expression of a given target gene (p-value < 0.05) were regarded as positive, otherwise they were regarded as negative (Figure 4). All enhancers overlapping the promoter region of their target genes were excluded.

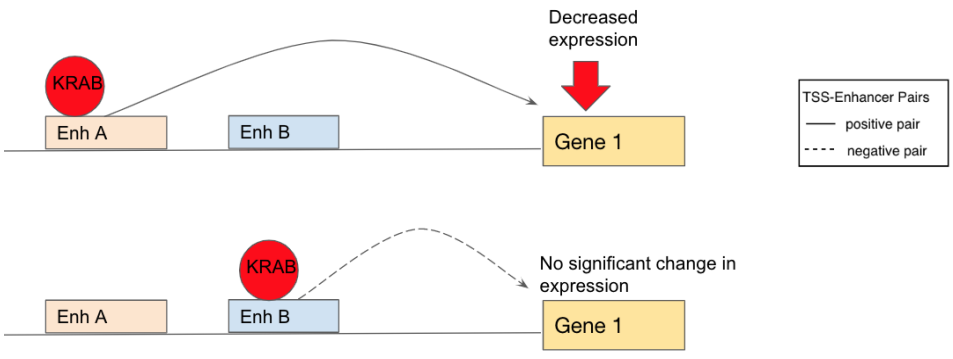


Figure 4. Schematic overview of CRISPRi-FlowFISH Positive and Negative Sets. Data obtained from [47] was generated using the CRISPRi-FlowFISH technology. The figure depicts a segment of DNA within which are observed two candidate enhancer regions (Enh A and Enh B) delineated with two small rectangles. In addition to these enhancer regions, there is one target gene (Gene 1) indicated by the large rectangle. The red circle indicates KRAB-dCAS9 blocking the accessibility of its bound region to TF binding, therefore inhibiting transcription at its bound region.

3.1.2 Nascent transcription-based Positive and Negative Sets

In the FANTOM5 project, enhancers and TSSs were linked based on correlation of CAGE tags by calculating the Pearson correlation coefficient between the expression levels of all enhancer and TSS pairs located within 500 kb from each other and expressed >1TPM in at least one sample. Enhancer-TSS pairs were linked if the Pearson correlation between their expression levels was significant (based on a Benjamini-Hochberg *FDR* <= 1e-5; adjusting the original empirically determined *p*-value).

We focused on K562 (for comparison with the CRISPRi-FlowFISH data) and three other well characterized cell lines (GM12878, Hep-G2 and MCF7), and established one positive set and two corresponding negative sets for each. Enhancers and promoters were defined as described in the Methods (see sections 2.3 Defining Promoter Regions and 2.4.2 Enhancer Regions in FANTOM5 CAGE Data). For each cell line, we generated all possible enhancer-promoter pairs where both the enhancer and TSS were expressed and located within 500 kb. Enhancer-promoter pairs that had been linked by FANTOM5 were included in the positive set, while the remaining enhancer-promoter pairs composed the first negative set (Figure 5a**)**. The second negative set, more restrictive, was obtained by filtering out any unlinked enhancers (*i.e.* enhancers not linked to any TSS in the positive set; Figure 5b).

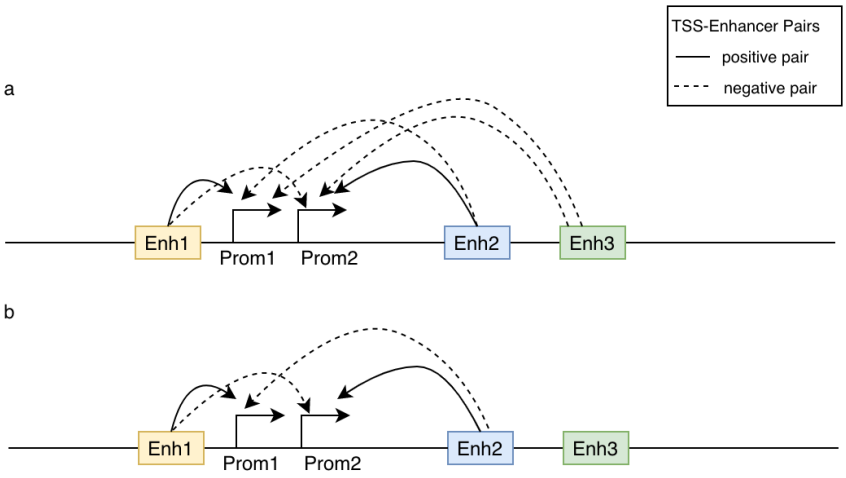


Figure 5. Depiction of FANTOM5 Positive and Negative Sets. The FANTOM5 consortium linked enhancers to TSSs by correlation of expression. The figure depicts a segment of DNA within which are observed three FANTOM5-identified enhancer regions (Enh1, Enh2 and Enh3) delineated with coloured rectangles. In addition, there are two distinct promoter regions (Prom1 and Prom2) delineated with right-angle arrows. Curved, solid arrows indicate positive enhancer-promoter pairs (enhancers and promoters that are linked by correlated FANTOM5 CAGE expression patterns), and dashed arrows indicate negative pairs. Panel **a)** depicts a negative set including all FANTOM5-identified cell type-specific enhancers., Panel **b)** depicts a negative set restricted to enhancers that are linked to at least one promoter in FANTOM5 (Enh3 is excluded from this negative set as it is not linked to any promoter).

The FANTOM5 method of linking enhancers to TSSs based on correlation of expression is expected to lead to the inclusion of false negative pairs to our negative set. For instance, false negative enhancer-promoter pairs were introduced when an enhancer was not linked to the strongest TSS of a target gene but to a weaker one (Figure 6). To avoid this, we integrated nuclear run-on assay-based (*i.e.* GRO- and PRO-seq) TIRs identified by dREG to our analysis. While CAGE captures the 5' transcripts of capped RNAs, GRO- and PRO-seq capture elongating RNAs. Figure 7 shows multiple CAGE TSSs overlapping the same TIR. Subsequently, to establish high-confidence positive and negative sets, for each cell line, we mapped FANTOM5 enhancer and TSS annotations to dREG-identified TIRs in that cell line. Enhancer-promoter pairs are in the dREG-filtered positive set if they are linked by FANTOM5 and they are in two separate dREG TIRs. Enhancer-promoter pairs are present in the

dREG-filtered negative set if they are within 500 kb and are located in separate dREG TIRs not linked in the positive set.

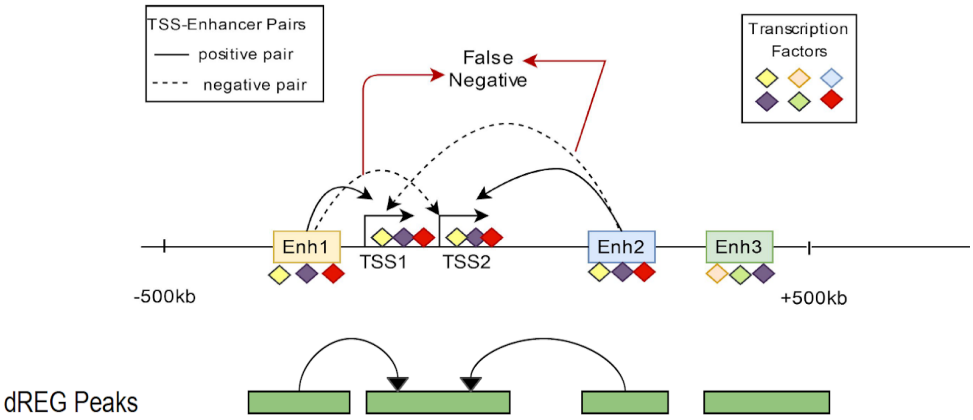
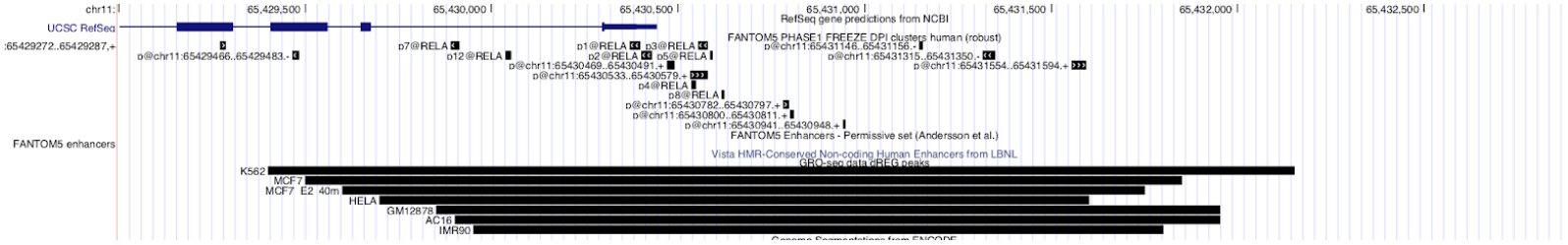


Figure 6. Correcting for false negative pairs by filtering for dREG overlapping enhancer and promoter regions. Three FANTOM5-identified enhancer regions (Enh1, Enh2 and Enh3) delineated with coloured rectangles and two FANTOM5-identified TSS regions(TSS1 and TSS2) delineated with right-angle arrows are shown in the DNA region above. Curved solid arrows indicate FANTOM5 Positive Set pairs and dashed arrows indicate FANTOM5 Negative Set pairs. dREG predicted TIRs overlapping FANTOM5-identified enhancer and TSSs are depicted with green bars below the DNA region. False negative pairs arise when a negative enhancer-TSS FANTOM5 pair overlaps the dREG TIRs of a positive enhancer-TSS pair (Enh1-TSS2 and Enh2-TSS1).

Figure 7. Genome browser view of the *RELA* promoter region. The short black bars represent



FANTOM5-identified TSSs and the long black bars represent dREG identified TIRs for each respective cell line.

3.1.3 CREST-seq Positive and Negative Sets

Diao *et al.*[33] used CREST-seq to interrogate the 2Mbp *POU5F1* locus in human embryonic stem cells H1, identifying 45 *POU5F1* enhancers. Identified enhancers (as described in Methods) were labelled as positive while the remaining windows were labelled as negative (Figure 8). The 10 kb region surrounding the strongest *POU5F1* TSS was excluded to prevent overlap between the promoter region and any enhancer.

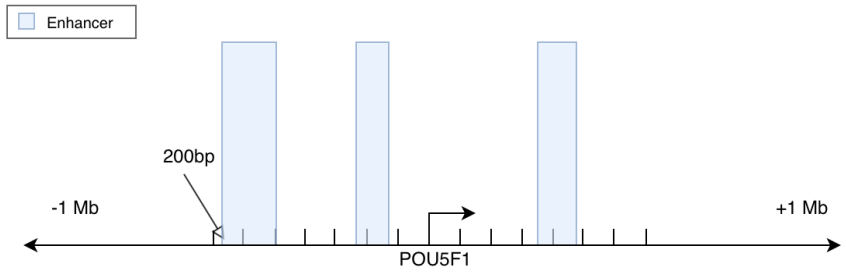


Figure 8. Generation of CREST-seq Positive and Negative set. Data was generated using CREST-seq technology[33]. The figure depicts the 2Mbp region of DNA centered at the promoter of the *POU5F1* gene (right angle arrow). The region is split into 200bp windows with a step size of 100bp. Windows overlapping identified enhancers by CREST-seq (blue bars) are in the Positive Set and all other windows are in the Negative Set

3.1.4 Window-based analysis of a representative gene

We decided to perform a computational analysis on a sample gene to depict anecdotally how the TF binding profile comparison could be used to suggest candidate relationships between enhancers and promoters. We selected *RELA* because it is active in the GM12878 cell

line (for which there is abundant TF binding data), and happened to be of interest to a member of the lab. We focused on the 1Mbp region centered at the strongest TSS of *RELA*, a TF critical for lymphoblastoid B cell (*e.g.* GM12878) development and function (Figure 9). Emulating the CREST-seq reference case, within the 1Mbp region, we generated 200 bp overlapping windows with a step-size of 100 bp. Windows overlapping a FANTOM5 enhancer linked to *RELA* in the GM12878 positive set were labelled as positive, while windows overlapping an enhancer linked to *RELA* in the GM12878 negative set were regarded as negative. Moreover, windows overlapping other GM12878 enhancers (*i.e.* not present in the positive or negative sets) were labelled as "unlinked", and windows overlapping other active GM12878 TSSs were labelled as "TSS". The remaining unlabelled windows were labelled as "none". To ensure that no windows overlapped with the promoter region of the gene, windows within 10 kb of the strongest TSS of *RELA* were excluded.

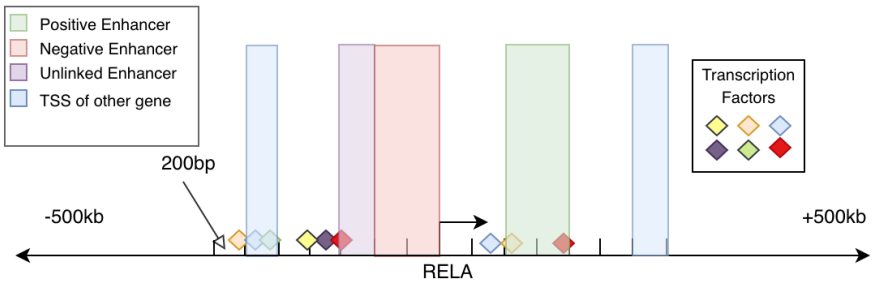


Figure 9. Generation of labelled *RELA* windows. The figure depicts the 1Mbp region of DNA centered at the *RELA* gene (right angle arrow). The region is split into 200bp windows with a step size of 100bp. Enhancers linked to *RELA* in the FANTOM5 Positive Set are denoted with green bars, enhancers linked to RELA in the FANTOM5 Negative Set are denoted with red bars, FANTOM5 identified enhancers that have not been linked are denoted with purple bars and FANTOM5 identified TSSs of other genes are denoted with blue bars. Windows overlapping each feature are labelled respectively, and windows that do not overlap any feature are labeled "none"

3.2 TF binding similarity comparisons between enhancer-promoter pairs

3.2.1 TF binding similarity comparison—linkage by CRISPRi-FlowFISH

We observed a significant difference in the distributions of cosine similarity scores between enhancer-TSS pairs in the positive and negative CRISPRi-FlowFISH sets (Figure 10a; K-S test statistic = 0.34; *p*-value = 9.9e-11). Hypothesizing that the signal is TSS-specific and cannot be compared across genes (*i.e.* a cosine similarity of 0.6 can be the best for one gene, and can represent a negative result in another), we performed Z-score normalization to enable comparison. Although Z-score normalization did not increase the difference between the two distributions (Figure 10b; K-S test statistic = 0.34; *p*-value = 8.8e-12), the majority of negative pairs had a negative Z-score while the majority of positive pairs had a positive Z-score.

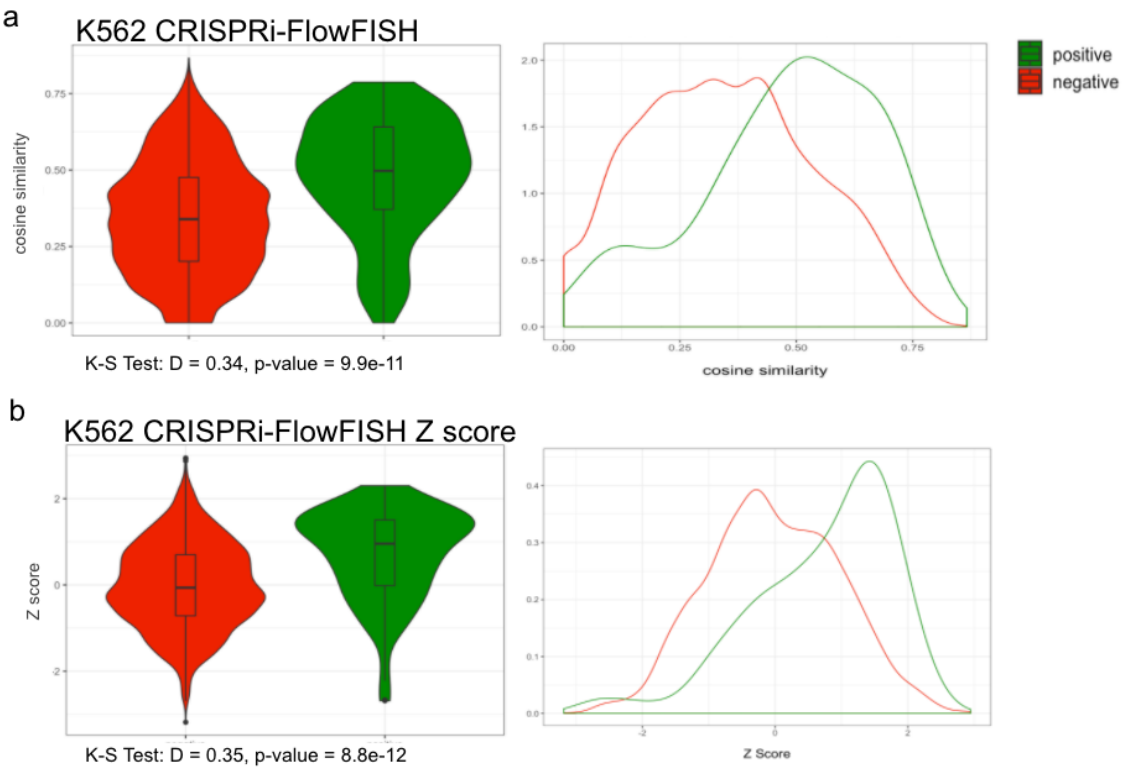


Figure 10. TF binding similarity comparison of CRISPRi-FlowFISH supported positive and negative sets. Cosine similarity scores were calculated between binary TF binding vectors for all enhancers and promoters in the CRISPR-FlowFISH reference data set. The results depicted in **a)** on the left are violin plots in which the y-axis represents the observed cosine similarity and the x-direction depicts the frequency of observations within the range of observed scores. The green distribution is for scores between functionally linked enhancers and promoters (*i.e.* from the positive set), while the red distribution represents scores between promoter and enhancer pairs from the negative set. The results on the right are the same data, but plotted in a smoothed histogram (where the x-axis is the cosine similarity and the y-axis is the frequency of observations). Z-score normalization was performed on all cosine similarity scores described above. The results of normalization are depicted in **b)** and consistent with the formatting of (A)

We analyzed whether the observed differences could be explained by a bias in distance (linear distance in terms of bp) between enhancers and promoters. We compared the distance between positively and negatively labelled enhancers to their target genes (Figure 11a), and observed that positive enhancers tend to be closer to their target genes. We corrected for the

difference in distance by calculating the distance between all enhancer-gene pairs and matching each positive enhancer-gene pair with the most similar negatively labelled enhancer-gene pair in distance (Figure 11b). After correcting for distance, the positive and negative sets still displayed a significant difference in cosine similarity distributions (K-S test D = 0.42, p-value = 7.4e-9) Figure 11c.

Moreover, we analysed whether the observed differences could be explained by a bias in GC composition. We observed a significant difference in the GC content distributions (K-S test D = 0.24, p-value = 2.1e-6) between positive and negative enhancers, and found that positively labeled enhancers have higher GC content than negatively labeled enhancers (Figure 11a). In order to correct for such discrepancy, we matched each positively labelled enhancer-gene pair with a negatively labelled enhancer-gene pair (keeping the gene constant) with the most similar GC content. Like for distance, the same observed signal (though less significant) could be seen after correcting by GC content; negative pairs displayed a significantly lower cosine similarity than positive pairs.

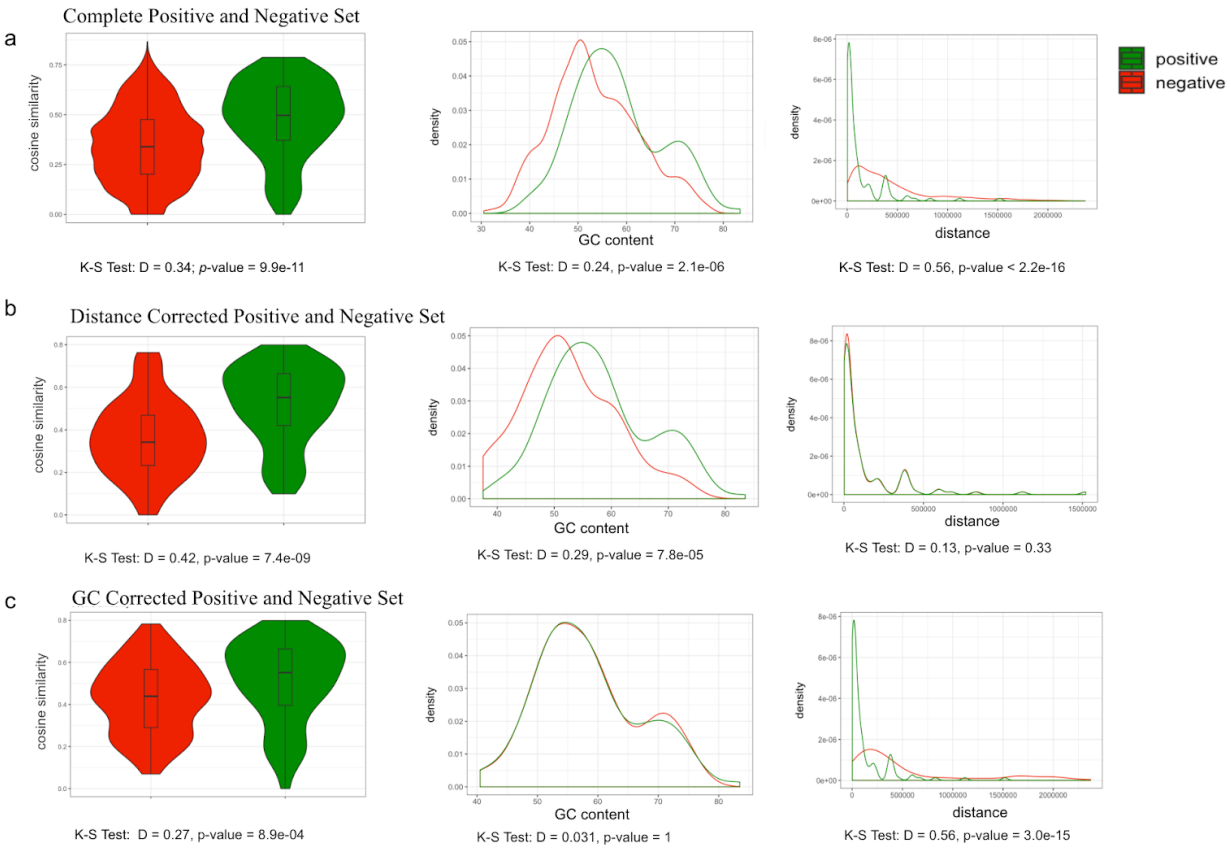


Figure 11. CRISPRi-FlowFISH Positive and Negative Set comparison after correcting for distance and GC content. Cosine similarity scores were calculated between binary TF binding vectors for all enhancers and promoters represented in the CRISPRi-FlowFISH reference data set. The results depicted in **a)** on the left are violin plots in which the y-axis represents the observed cosine similarity and the x-direction depicts the frequency of observations within the range of observed scores. The green distribution is for scores between functionally linked enhancers and promoters (*i.e.* from the positive set), while the red distribution represents scores between promoter and enhancer pairs from the negative set. The results in the middle compares GC content between enhancer and promoter pairs from the positive (green) and negative (red) sets (where the x-axis is the percent GC composition and the y-axis is the density of observations). The results on the right compares the distances between enhancer and promoter pairs from the positive (green) and negative (red) sets(where the x-axis is the distance in bp and the y-axis is the density). The results in **b)** represent comparisons made in **a)** after correcting for distance by pairing each positive enhancer-promoter pair to the negative pair most similar in distance. The results in **c)** represent comparisons after GC content is corrected by pairing each positive enhancer-promoter pair to the negative pair with the most similar GC composition.

3.2.2 TF binding similarity comparison—linkage by FANTOM5

To investigate whether the observed signal could be generalized beyond K562 cells, we compared TF binding signatures between FANTOM5-derived positive and negative sets in four cell lines: K562 (to enable comparison with the CRISPRi-FlowFISH data), GM12878, MCF-7 and Hep-G2. As shown in Figure 12, the difference in distributions of cosine similarities between the positive and negative sets were significant in all four cell lines.

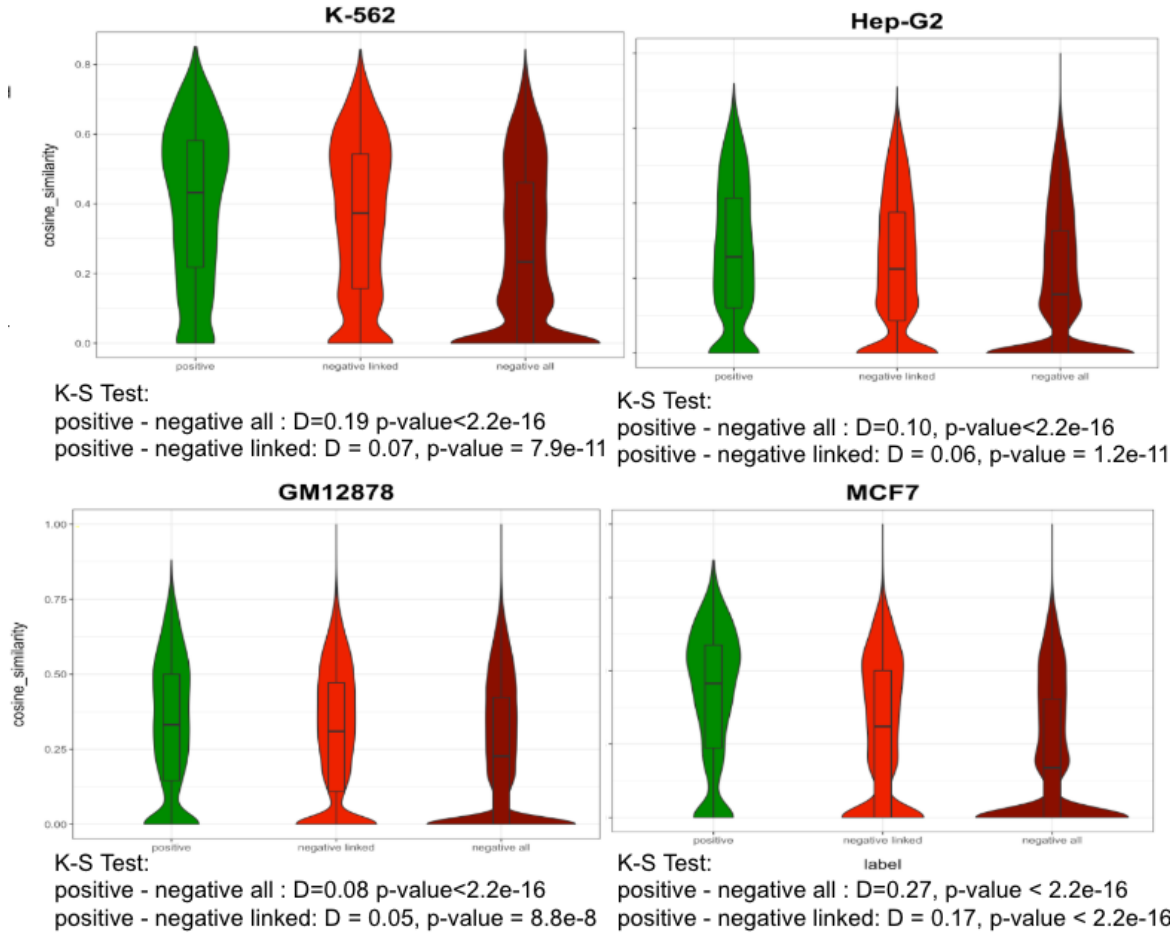


Figure 12. **Cosine similarity comparison between FANTOM5 Positive and Negative Sets.** Cosine similarity scores were calculated between binary TF binding vectors for all enhancers and promoters represented in each (K562, Hep-G2, GM12878, MCF7) FANTOM5 dataset. Violin plots were generated for each cell line in which the y-axis represents the observed cosine similarity and the x-direction depicts

the frequency of observations within the range of observed scores. The green distribution is for scores between enhancers and promoters linked by FANTOM5. The dark red distribution represents scores between promoters and all cell type specific enhancers that are not reported to be correlated by FANTOM5. The bright red distribution represents scores between promoters and cell type specific enhancers that are not reported to be correlated but are linked in FANTOM5.

We then compared TF binding similarity between enhancer-gene pairs in positive and negative sets at the promoter level. For each cell line, we compared promoters with both positively and negatively linked enhancers and found that for the majority of promoters, the average cosine similarity score is higher for positively labelled enhancers when compared to negatively labelled enhancers. (Figure 13).

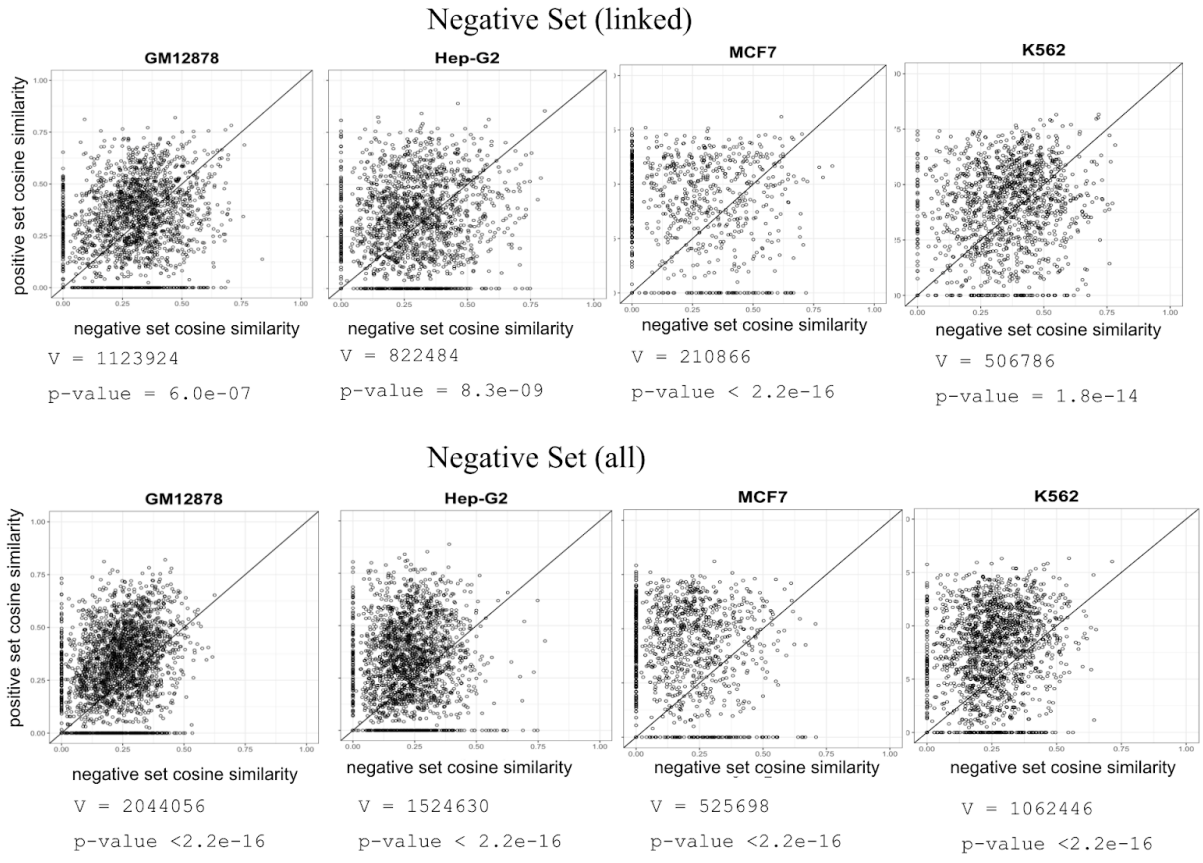


Figure 13. Average Positive Set vs average Negative Set cosine similarity comparison. In each cell

line-specific plot, each point represents a unique promoter. The Y coordinate represents the average cosine similarity score calculated between binary TF binding vectors of a promoter and the enhancers linked to it by FANTOM5 (situated within 500kbp). The X coordinate represents the average cosine similarity score between the promoter and cell type-specific enhancers linked elsewhere in FANTOM5 (Negative Set linked) or all cell type-specific enhancers in FANTOM5 that are not linked to the promoter (Negative Set all). Corresponding Wilcoxon Test statistics are reported for each graph.

In an attempt to correct for the potential inclusion of false positive and negative pairs by the correlation method used in FANTOM5, we filtered the GM12878 and MCF-7 positive and negative sets for pairs overlapping dREG-predicted TIRs. After filtering, a similar proportion of promoters had a higher average cosine similarity for positively labelled enhancers than for negatively labelled enhancers (Figure 14b).

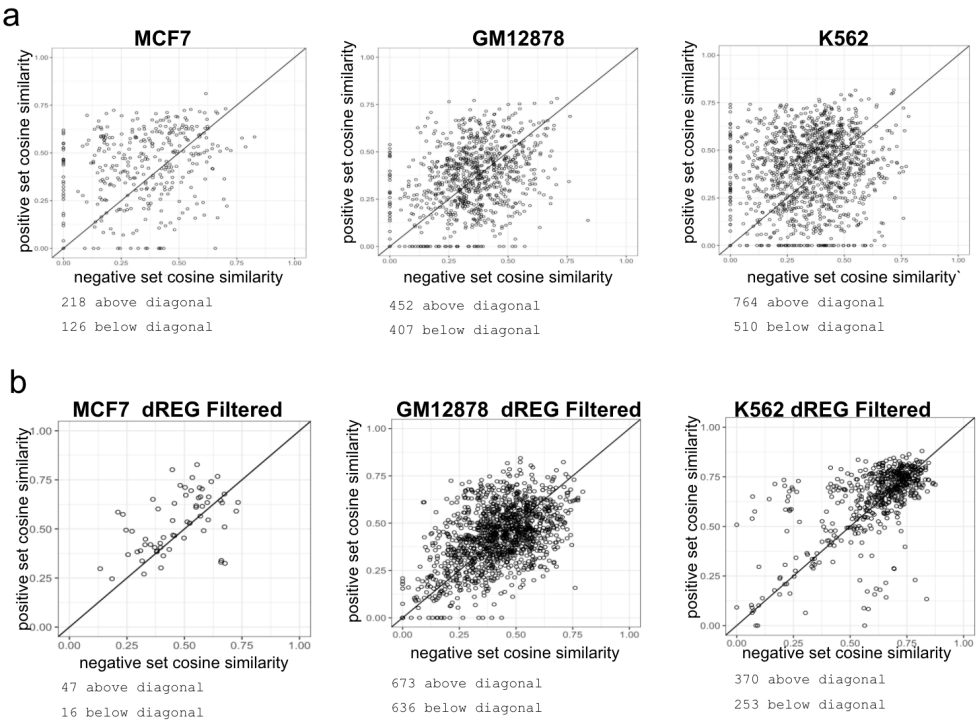


Figure 14. Promoter level comparison of TF binding similarity between dREG-filtered Positive and Negative Sets Cosine similarity scores were calculated between binary TF binding vectors for all enhancers and promoters in dREG-corrected Positive and Negative Sets. The results depicted in **a)** are scatterplots where each point represents a unique promoter with the x-coordinate being the average

negative set cosine similarity and the y-coordinate being the average positive set cosine similarity of the promoter before filtering with dREG. The results after dREG filtering are depicted in **b)** where each point represents a unique promoter with the x-coordinate being the average negative set cosine similarity and the y-coordinate being the average positive set cosine similarity of the promoter after dREG filtering. The diagonal line represents the axis where the average Positive Set score equals the Negative. The number of promoters above and below the diagonal is reported for each graph.

At the enhancer level, differences between TF binding similarity between enhancer-gene pairs in positive and negative sets were more discernible. After filtering, a greater proportion of enhancers had a higher average cosine similarity for positively labelled promoters than for negatively labelled promoters (Figure 15).

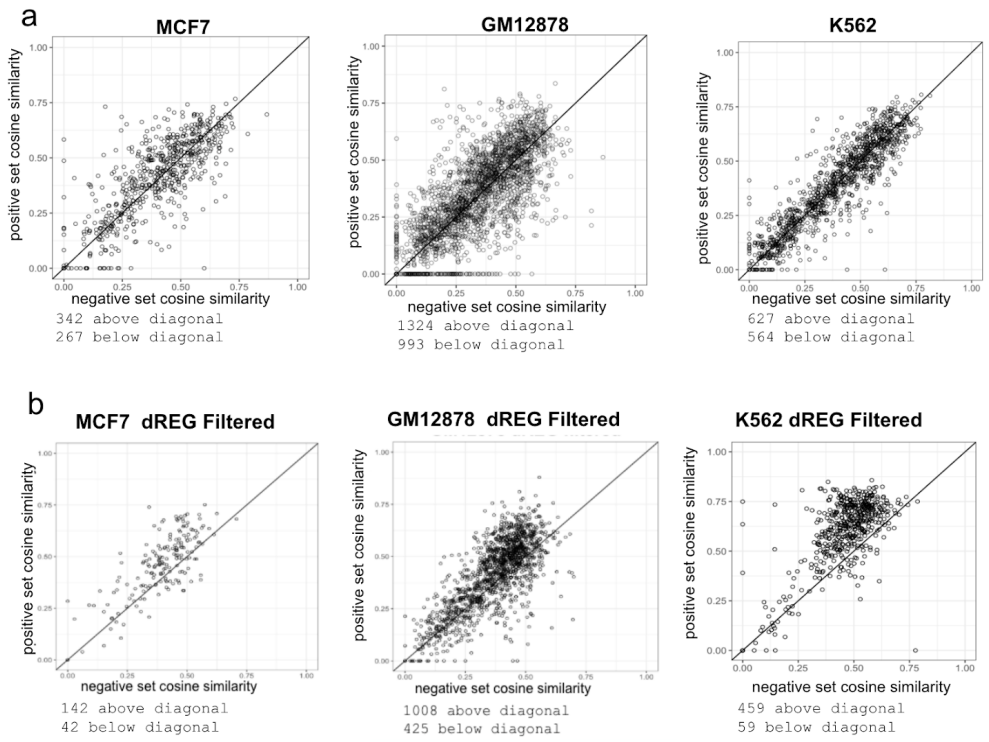
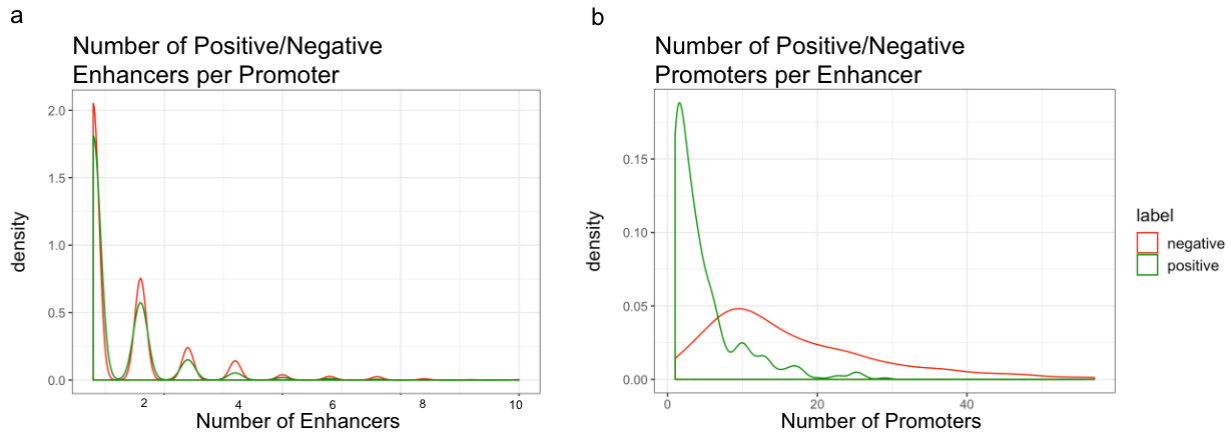


Figure 15. Enhancer Level Comparison of TF binding similarity between dREG-filtered Positive and Negative Sets Cosine similarity scores were calculated between binary TF binding vectors for all enhancers and promoters in dREG-corrected Positive and Negative Sets. The results depicted in **a)** are scatterplots where each point represents a unique enhancer with the x-coordinate being the average

negative set cosine similarity and the y-coordinate being the average positive set cosine similarity of the enhancer before filtering with dREG. The results after dREG filtering are depicted in **b)** where each point represents a unique enhancer with the x-coordinate being the average negative set cosine similarity and the y-coordinate being the average positive set cosine similarity of the promoter after dREG filtering. The diagonal line represents the axis where the average Positive Set score equals the Negative. The number of enhancers above and below the diagonal is reported for each graph.

We analyzed the compositions of dREG filtered FANTOM5 positive and negative sets and found that while each promoter has a balanced set of positive and negative enhancers, enhancers are heavily biased to have negative promoters linked to them (Figure 16). This bias appears to be due to an imbalance in the number of active promoters versus enhancers in each cell type. Each set of cell type specific promoter regions identified by FANTOM5 is more than



five times larger than than the corresponding FANTOM5 identified enhancer region set.

Figure 16. Promoter and Enhancer Set Compositions. The results depicted in **a)** compares the number of positive (green) and negative (red) enhancers linked to each promoter in the dREG filtered positive and negative sets. The x-axis represents the number of enhancers and the y-axis depicts the density of observations. The results depicted in **b)** compares the number of positive (green) and negative (red) promoters linked to each enhancer in the dREG filtered positive and negative sets. The x-axis represents the number of promoters linked to each enhancer and the y-axis represents the density of observations.

To balance the number of positive and negative promoters linked to each enhancer in the dREG filtered positive and negative sets, we matched each positive enhancer-promoter pair with a negative enhancer-promoter pair most similar in distance. The same observed enhancer specific signal is present in the resulting balanced sets (Figure 17).

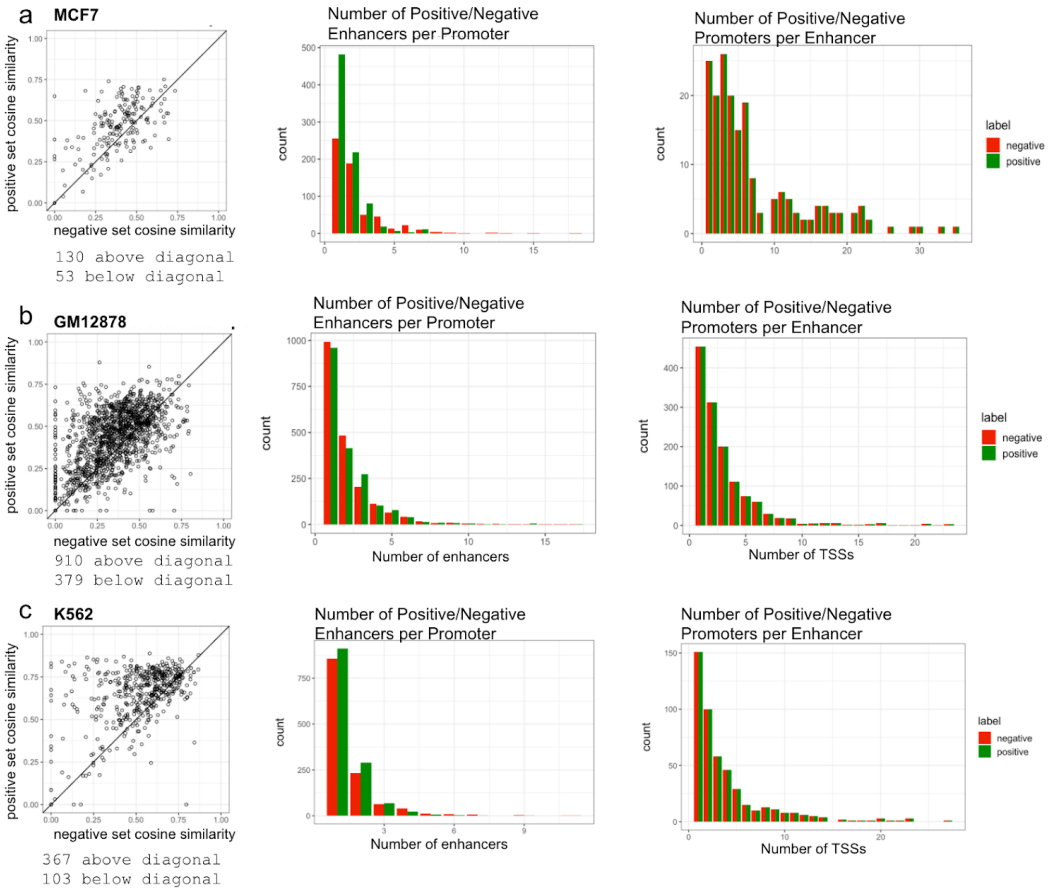


Figure 17. Balanced enhancer level comparison of TF binding similarity between dREG-filtered Positive and Negative Sets The number of positive and negative promoters linked to each enhancer in dREG filtered Positive and Negative Sets were balanced by matching each positive enhancer-promoter pair with a negative enhancer-promoter pair most similar in distance. The results of balancing the MCF7 set is depicted in **a)**. On the left is a scatter plot where each point represents a unique enhancer with the x-coordinate being the average negative set cosine similarity and the y-coordinate being the average positive set cosine similarity of the enhancer. The diagonal line indicates where the average Positive Set score equals the Negative. The number of enhancers above and below the diagonal is reported. In the center is a bar graph comparing the number of positive (green) and negative (red) enhancers linked to

each promoter after balancing. The x-axis represents the number of enhancers and the y-axis depicts the number of promoters. On the right is a bar graph comparing the number of positive (green) and negative (red) promoters linked to each enhancer after balancing. The x-axis represents the number of promoters and the y-axis depicts the number of enhancers. The results depicted in **b)** and **c)** represent the comparisons made in **a)** but for GM12878 and K562 datasets respectively.

3.2.3 TF binding similarity comparison—individual promoters

We hypothesized that TF binding similarity scores between enhancer-gene pairs is CRR specific rather than directly comparable between pairs involving different promoters within the same gene (or between genes). While the FANTOM5 enhancer-promoter linkages did not show a strong signal at the promoter level on a genome wide scale, under the hypothesized promoter-specific scoring we would expect better performance on the local level, so we sought some case examples of deep data on individual promoters.

We compared the TF binding signatures between the CREST-seq *POU5F1* positive and negative sets. We observed that most windows not overlapping with a *POU5F1* enhancer have a negative Z-score (Figure 18a). These windows correspond to genomic regions without coverage for TF binding data. Therefore, we omitted these regions from our analysis. The resulting plot, Figure 15b, still shows a difference in the distribution of Z-scores, with enhancers having higher Z scores.

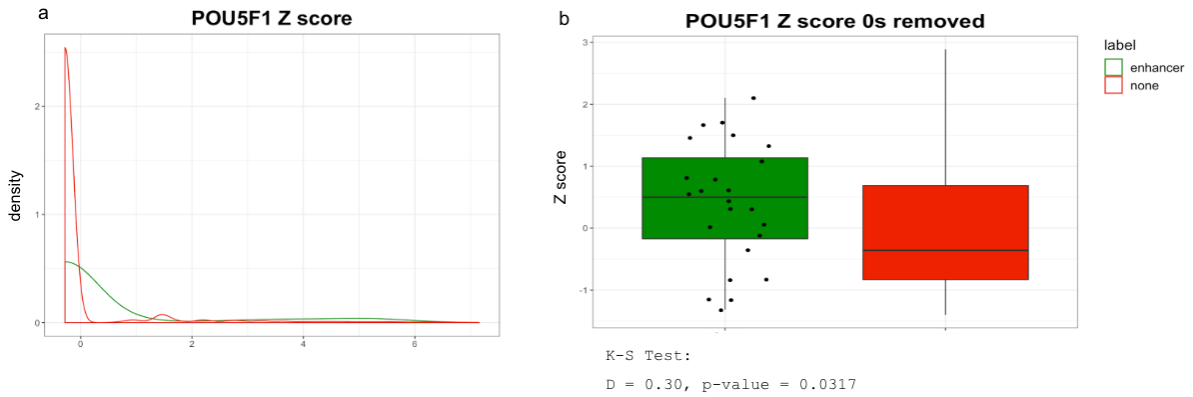


Figure 18. Comparison of Z-scores between CREST-seq-identified *POUF51* enhancer regions and neighbouring TF binding regions. Z-scores were calculated from cosine similarity scores between binary TF binding vectors for all identified enhancers by CREST-seq and the *POU5F1* promoter, as well as neighbouring regions and the *POU5F1* promoter. The results depicted in **a)** compares the distribution of Z-scores between enhancer regions (green) and surrounding non-enhancer regions (red) where the x-axis is the Z-score and the y-axis is the density of observations. In **b)** non-enhancer regions without TF binding activity are removed and Z-scores are compared between enhancer and non-enhancer regions.

Similarly we compared TF binding similarity between the *RELA* promoter and each labelled 200 bp window (described in 3.1.4.). Figure 19 shows the different distributions of cosine similarities after Z-score normalization. We observed that positive and TSS-labelled windows had the highest Z-scores when compared to the rest. Our findings are in concordance with the observation made by Diao *et al.*[33] that promoters of proximal genes can function as enhancers.

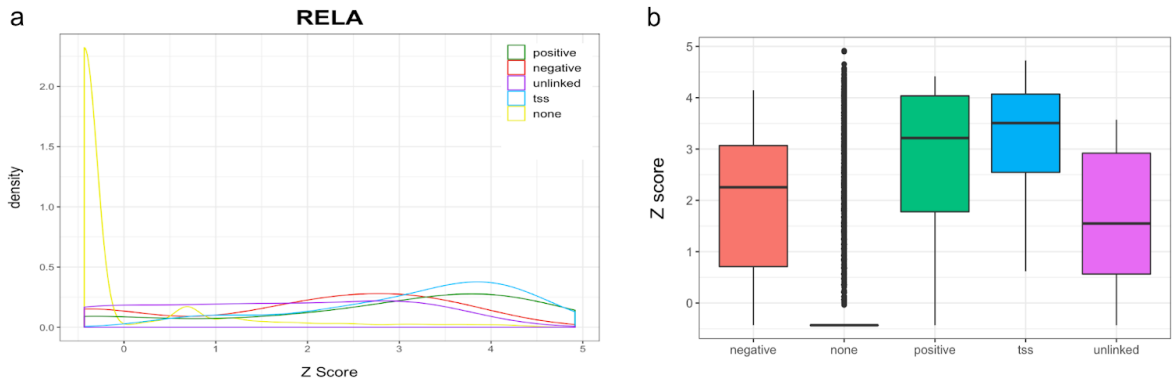


Figure 19. Comparison of distribution of Z scores between labeled windows surrounding *RELA*. Z-scores were calculated from cosine similarity scores between binary TF binding vectors for all FANTOM5-identified enhancers and the *RELA* promoter. Enhancers linked to *RELA* in the FANTOM5 Positive Set are represented in green, enhancers linked to *RELA* in the FANTOM5 Negative Set are represented in red bars, FANTOM5-identified enhancers that have not been linked to any TSS are represented in purple, FANTOM5-identified TSSs of other genes are represented in blue, and regions that do not overlap any feature (*i.e.* none) are represented in yellow. The results depicted in **a)** compares the distribution of Z-scores between each labelled region with the Z-score on the x-axis and the density of observations on the y-axis. The results in **b)** are the same data as **a)** but presented as boxplots with the y-axis representing the observed Z scores.

Discussion

We have developed a novel metric to link enhancers to their target genes based on similarity of TF binding. In our analysis, we show that enhancers share more commonalities in TF binding with their target genes compared to genes they do not regulate. While there are statistically significant differences between score distributions, the distributions remain overlapping, and on a genome scale. By coupling additional information about active regulatory regions into the process, we demonstrate that pairing of promoters to a specific enhancer of interest can be informative, while the mapping of enhancers to a specific promoter of interest is not improved by the filtering process. Case studies of individual genomic regions support the observation that the scores are CRR specific, and therefore the method may be most immediately useful for local analyses of individual genes. The TF binding profile comparison introduced in this thesis represents one of many approaches (both computational and experimental) currently being pursued to determine relationships between enhancers and promoters. Two computational methods, TargetFinder and PEPmotif, learn general features of enhancer-promoter pairs, but they do not capture enhancer-specific patterns. Based on the research in the thesis, it appears that local CRE-specific characteristics will be important to optimize the mapping success. Fulco *et al.*[43] developed the activity-by-contact (ABC) model to computationally link enhancers to their promoters based on integrating the presence of overlapping DHS and H3K27ac markers at each enhancer with Hi-C identified promoter regions interacting with each enhancer. In the published results involving the three models, ABC appears to be performing slightly better, but direct comparison is difficult because the methods require distinct input data. In the future, it may be possible to combine the approaches in the ABC model with the TF binding profile comparisons to improve performance.

Improvement of the TF binding profile comparison approach should be feasible. The use of cosine similarity as the metric for comparisons was based on trying a range of vector comparison options. We anticipate that there may be further improvements possible with alternative approaches. In recent work in my laboratory (Saraswat, unpublished), the scoring system was adjusted by dividing its cosine similarity by the sum of cosine similarities of all enhancers (positive and negative) linked to that promoter, in an attempt to resolve the local nature of the original metric. We will continue to evaluate this alternative, which appears promising. Training machine learning methods with the TF binding profiles, rather than performing vector comparisons as in this thesis, may also allow improved performance. Following our observations in the FANTOM5 and CREST-seq positive and negative sets that TF-binding similarity is a CRE-specific metric rather than a continuous scale, we believe that our novel bioinformatics approach should be incorporated as a feature to future machine learning models to improve the performance of computationally-predicted enhancer-promoter pairs rather than be used on its own.

There are distinct limitations of the TF binding profile comparisons, as it depends on experimental TF ChIP-seq data. For example, data sparsity affects our analysis in H1 cells; TF binding data is available for 31 different TFs in H1 cells in comparison to over 80 TFs for other cell lines in our analysis. Future developments to the TF binding profile method will attempt to mitigate experimental data dependencies by focusing on computational prediction based on a combination of experimental data and DNA sequence using machine learning methods. In particular, there are indications that comparisons on the sequence level based on gapped k-mers may have utility.

The challenge of reliably identifying which CREs are functionally interacting represents

the current challenge in a long-term effort to unravel the complex regulatory networks governing transcriptions. Dramatic advances in recent years have allowed the individual CREs to be annotated across the genome, but the higher order relationships between them remain challenging to predict. This thesis explored a new algorithmic approach to predicting relationships between enhancers and promoters based on a hypothesis that functionally interacting members of the two classes of CREs will tend to be bound by the same TFs. The research showed the approach to be promising, and provides direction for further work focused on adjusting the comparisons to account for the specific properties of individual promoters or enhancers.

Bibliography

1. Gene Expression and Regulation | Learn Science at Scitable. https://www.nature.com/scitable/topic/gene-expression-and-regulation-15/.
2. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **175**, 598–599 (2018).
3. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
4. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**, 1497–1502 (2007).
5. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
6. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858–871.e8 (2018).
7. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
8. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).
9. Yang, Y., Zhang, R., Singh, S. & Ma, J. Exploiting sequence-based features for predicting enhancer-promoter interactions. *Bioinformatics* **33**, i252–i260 (2017).
10. Moore, J. E., Pratt, H. E., Purcaro, M. J. & Weng, Z. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol.* **21**, 17 (2020).
11. Herkert, B. & Eilers, M. Transcriptional repression: the dark side of myc. *Genes Cancer* **1**, 580–586 (2010).

12. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
13. Lee, B.-K. *et al.* Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Res.* **22**, 9–24 (2012).
14. Siggers, T. & Gordân, R. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res.* **42**, 2099–2111 (2014).
15. Transcription Factors. (2014) doi:10.1016/B978-0-12-801238-3.05466-0.
16. Dynan, W. S. & Tjian, R. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* **35**, 79–87 (1983).
17. Dufour, S., Broders-Bondon, F. & Bondurand, N. Chapter 13 - β1-Integrin Function and Interplay during Enteric Nervous System Development. in *Neural Surface Antigens* (ed. Pruszak, J.) 153–166 (Academic Press, 2015).
18. Vandel, J., Cassan, O., Lèbre, S., Lecellier, C.-H. & Bréhélin, L. Probing transcription factor combinatorics in different promoter classes and in enhancers. *BMC Genomics* **20**, 103 (2019).
19. Vierbuchen, T. *et al.* AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection. *Mol. Cell* **68**, 1067–1082.e12 (2017).
20. Xie, Z., Hu, S., Qian, J., Blackshaw, S. & Zhu, H. Systematic characterization of protein-DNA interactions. *Cellular and Molecular Life Sciences* vol. 68 1657–1668 (2011).
21. ENCODE (Encyclopedia of DNA Genetic Elements). *Encyclopedia of Genetics, Genomics, Proteomics and Informatics* 602–602 (2008) doi:10.1007/978-1-4020-6754-9_5275.
22. Chèneby, J. *et al.* ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids*

Res. **48**, D180–D188 (2020).

23. Schier, A. C. & Taatjes, D. J. Structure and mechanism of the RNA polymerase II transcription machinery. *Genes Dev.* **34**, 465–488 (2020).
24. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
25. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.* **8**, 424–436 (2007).
26. Yang, C., Bolotin, E., Jiang, T., Sladek, F. M. & Martinez, E. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* **389**, 52–65 (2007).
27. Gene Expression. in *Cell Biology* 165–187 (Elsevier, 2017).
28. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
29. Karnuta, J. M. & Scacheri, P. C. Enhancers: bridging the gap between gene control and human disease. *Hum. Mol. Genet.* **27**, R219–R227 (2018).
30. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
31. Preker, P. *et al.* PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res.* **39**, 7179–7193 (2011).
32. Nguyen, T. A. *et al.* High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* **26**, 1023–1033 (2016).
33. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* **14**, 629–635 (2017).

34. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15776–15781 (2003).
35. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
36. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).
37. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
38. Jou, J. *et al.* The ENCODE Portal as an Epigenomics Resource. *Curr. Protoc. Bioinformatics* **68**, e89 (2019).
39. Chen, Y. & Chen, A. Unveiling the gene regulatory landscape in diseases through the identification of DNase I-hypersensitive sites. *Biomed Rep* **11**, 87–97 (2019).
40. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, (2018).
41. Lopes, R., Agami, R. & Korkmaz, G. GRO-seq, A Tool for Identification of Transcripts Regulating Gene Expression. *Methods Mol. Biol.* **1543**, 45–55 (2017).
42. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods* **12**, 433–438 (2015).
43. Wang, J. *et al.* HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.* **47**, D106–D112 (2019).
44. Yao, L., Berman, B. P. & Farnham, P. J. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Crit. Rev. Biochem. Mol. Biol.* **50**, 550–573 (2015).
45. Cao, F. & Fullwood, M. J. Inflated performance measures in enhancer-promoter

interaction-prediction methods. *Nature genetics* vol. 51 1196–1198 (2019).

46. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
47. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
48. Krivan, W. & Wasserman, W. W. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* **11**, 1559–1566 (2001).
49. Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* **46**, D267–D275 (2018).
50. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).
51. Koudritsky, M. & Domany, E. Positional distribution of human transcription factor binding sites. *Nucleic Acids Res.* **36**, 6795–6805 (2008).
52. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
53. Grossman, S. R. *et al.* Positional specificity of different transcription factor classes within enhancers. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E7222–E7230 (2018).
54. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
55. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).