Sackler Faculty of Exact Science, Blavatnik School of Computer Science

Methods for inferring and utilizing enhancer-promoter networks

THESIS SUBMITTED FOR THE DEGREE OF

"DOCTOR OF PHILOSOPHY"

By

Tom Aharon Hait

The work on this thesis has been carried out

under the supervision of

Prof. Ron Shamir and **Prof. Ran Elkon**

Submitted to the Senate of Tel-Aviv University

September 2023

Acknowledgments

I wish to express my sincere gratitude to my two supervisors, Prof. Rani Elkon and Prof. Ron Shamir. During my MSc and PhD Rani and Ron gave me the freedom to pursue my research interests under their professional guidance and extensive knowledge. Their constant encouragement and support helped me to get to the finish line. Thank you, Ron and Rani!

I would like to thank all my friends and the past and present lab members for their support and cooperation. I would like to thank all my students that I have taught in different courses for providing me with moments outside of research.

I would like to thank the Edmond J. Safra Center for Bioinformatics for the support in my MSc and PhD. Special thanks to Gilit Zohar-Oren for being always there to help and cheer up.

Last but definitely not least, I would like to thank my dear family for their support during my studies.

Preface

This thesis is based on the following papers:

1. Hait TA, Amar D, Shamir R, Elkon R**. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map.** *Genome Biol* 2018, 19:56. https://pubmed.ncbi.nlm.nih.gov/29716618/
2. Hait TA, Elkon R, Shamir R. **CT-FOCS: a novel method for inferring cell type-specific enhancer–promoter maps.** *Nucleic Acids Research*, Volume 50, Issue 10, 10 June 2022, Page e55, https://doi.org/10.1093/nar/gkac048.
3. Hait TA, Elkon R, Shamir R**. Inferring transcriptional activation and repression maps in single-nucleotide resolution using deep-learning.** Submitted.

The following works completed during the Ph.D. are not covered in the thesis:

4. Chandran A.E.J, Finkler A, Hait TA, Kiere Y, David S, Pasmanik-Chor M, Shkolnik D. **Calcium regulation of the Arabidopsis Na+/K+ transporter HKT1;1 improves seed germination under salt stress**. Submitted.
5. Grigg I, Ivashko-Pachima I*, Hait TA*, Korenková V, Touloumi O, Lagoudaki R, Van Dijck A, Marusic Z, Anicic M, Vukovic J, Kooy RF, Grigoriadis N, Gozes I. **Tauopathy in the young autistic brain: novel biomarker and therapeutic target.** *Translational Psychiatry* (2020), Volume 10, Article number 228. *contributed equally. https://pubmed.ncbi.nlm.nih.gov/32661233/
6. Hait TA, Maron-Katz A, Sagir D, Amar D, Ulitsky I, Linhart C, Tany A, Sharan R, Shiloh Y, Elkon R, Shamir R. **The EXPANDER Integrated Platform for Transcriptome Analysis.** *J. Mol. Biol.* (2019), Volume 431, Issue 13, Pages 2398-2406. https://pubmed.ncbi.nlm.nih.gov/31100387/

Abstract

We live in an exciting era where massive biological and biomedical datasets have been produced, allowing us to discover novel biological insights on genome regulation, which describes how the cell controls the amount and exact composition of proteins it produces from each gene in a given circumstances. A key element in this effort is the computational tasks of discovering noncoding regulatory elements and how they are spatially organized in a 3D genome structure to control transcription. Moreover, understanding how cells obtain specific 3D structures can provide valuable insights into the cell type-specific transcriptional events that ultimately dictate cell fate decisions. In this thesis, we studied the practical and statistical aspects of the regulatory elements and their spatial organization from a broad variety of data sources covering diverse cell types. By utilizing techniques from probabilistic modeling, and statistical and deep learning we were able to handle complex large scale data.

In this work, we developed novel methods for inferring enhancer-promoter interactions. The first method, expanded from the MSc studies, infers interactions showing high correlation between the enhancer and promoter activity patterns across many cell types. The second method dissect which of those interactions are cell type-specific. We showed that both methods outperform existing methods and provide novel biological insights. Lastly, using deep learning techniques, we answered on various questions on how to properly predict functional silencers and what defines them epigenetically.

Contents

Overview of the projects included in this thesis

1. **Predicting global enhancer-promoter maps**

   Recent sequencing technologies enable joint quantification of promoters and their enhancer regions, allowing inference of enhancer–promoter links. We show that current enhancer–promoter inference methods produce a high rate of false positive links. We introduce FOCS, a new inference method, and by benchmarking against ChIA-PET, HiChIP, and eQTL data show that it results in lower false discovery rates and at the same time higher inference power. By applying FOCS to 2630 samples taken from ENCODE, Roadmap Epigenomics, FANTOM5, and a new compendium of GRO-seq samples, we provide extensive enhancer–promotor maps (http://acgt.cs.tau.ac.il/focs). We illustrate the usability of our maps for deriving biological hypotheses.

   This study was published as:
   FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map
   Tom Aharon Hait, Ron Shamir and Ran Elkon
   Genome Biol 2018, 19:56

2. **Predicting cell-type specific enhancer-promoter maps**

   Spatiotemporal gene expression patterns are governed largely by the activity of enhancer elements, which engage in physical contacts with their target genes. Identification of enhancer–promoter links that are functional only in a specific subset of cell types is a key challenge in understanding gene regulation. We introduce CT-FOCS (cell type FOCS), a statistical inference method that uses linear mixed effect models to infer EP links that show marked activity only in a single or a small subset of cell types out of a large panel of probed cell types. Analyzing 808 samples from FANTOM5, covering 472 cell lines, primary cells and tissues, CT-FOCS inferred such EP links more accurately than recent state-of-the-art methods. Furthermore, we show that strictly cell type-specific EP links are very uncommon in the human genome.

   This study was published as:
   CT-FOCS: a novel method for inferring cell type-specific enhancer–promoter maps
   Tom Aharon Hait, Ran Elkon and Ron Shamir
   Nucleic Acids Research, Volume 50, Issue 10, 10 June 2022, Page e55

3. **Inferring transcriptional activation and repression activity maps in single-nucleotide resolution using deep-learning**

   Recent computational methods for inferring cell type-specific functional regulatory elements have used sequence and epigenetic data. Active regulatory elements are characterized by open-chromatin state, and the novel experimental technique ATAC-STARR-seq couples ATAC-seq assays, which capture such genomic regions, with a

functional assay (STARR-seq) to selectively examine the regulatory activity of accessible DNA. ATAC-STARR-seq may thus provide data that could improve the quality of computational inference of active enhancers and silencers. Here, we propose a novel regression-based deep learning (DL) model that utilizes such data for predicting single nucleotide activation and repression maps. We found that while models using only sequence and epigenetics data predict active enhancers with high accuracy, they generally perform poorly in predicting active silencers. In contrast, models building also on data of experimentally identified enhancers and silencers do substantially better in the identification of active silencers. Our model predicts many novel enhancers and silencers in the model lymphoblastoid cell line GM12878. Epigenetic signatures of the novel regulatory elements detected by our model resemble the ones shown by the experimentally validated enhancers and silencers in this cell line. ChIP-seq enrichment analysis in predicted novel silencers identify a few significant enriched transcriptional repressors such as SUZ12 and EZH2, which compose the PRC2 repressive complex. Intersection with GWAS data found that the novel predicted enhancers are specifically enriched for risk SNPs of the Lupus autoimmune disease. Overall, while silencers are still poorly understood, our results show that our DL-model can be used to complement the experimental results on regulatory element discovery.

A manuscript summarizing this study is available as:
Inferring transcriptional activation and repression maps in single-nucleotide resolution using deep-learning
Tom Aharon Hait, Ran Elkon and Ron Shamir
Research Square, Preprint, 23 August 2023

1. Introduction

1.1. Biological background

This chapter introduces the biological concepts and definitions needed to understand the goals and computational problems addressed in this thesis. For more information on basic biology, please refer to Alberts et al. [1]. For gene regulation, enhancers, silencers, and epigenetics, please refer to Shlyueva et al. and Segert et al. [2,3]. We also discuss current experimental technologies that allow us to systematically identify genomic regions of interest and measure their signals for our computational analyses.

1.1.1 Fundamentals of Cellular Biology

Living organisms consist of cells, which serve as the fundamental units of life. The field of cell biology focuses on examining the structure, function, and behavior of cells. The regulation of cellular and organismal functions and development is governed by deoxyribonucleic acid (DNA). DNA is housed within the cell's nucleus and contains functional segments known as genes. Genes can be categorized into two main groups: coding genes and non-coding genes. Coding genes contain instructions for protein synthesis, whereas non-coding genes contain instructions for the production of non-coding ribonucleic acids (ncRNAs), which do not undergo protein translation but regulate various cellular functions.

Non-coding regions proximal and upstream to genes are known as *promoters*. They encompass sequences capable of binding proteins and regulating gene transcription. *Enhancers*, another explored category of regulatory elements (REs), are distal regions from the genes that also bind activator proteins to stimulate gene transcription. *Silencers*, an understudied type of REs, are proximal and distal regions from the gene that bind repressor proteins to decrease gene transcription.

Inter-individual differences in coding sequences, as well as in non-coding regulatory regions, such as enhancers and silencers, contribute to the genetic variation observed in human DNA sequences. *Single nucleotide polymorphisms* (SNPs) are variations occurring at a single nucleotide position. Non-coding regions exhibit a higher frequency of SNPs compared to coding regions [4]. These variations have the potential to impact an individual's susceptibility to diseases, their response to pathogens, chemicals, and other factors.

By comprehending the mechanisms through which non-coding regulatory regions interact with genes and impact gene transcription in a cell type specific manner, we can establish cell type specific connections between variations in non-coding regions and the transcription levels of specific genes. This deeper understanding of genetic factors affecting

disease susceptibility can contribute to advancements in understanding disease-related genetic predispositions.

1.1.2 Gene regulation

Gene regulation is a fundamental process in which cells increase or decrease the amounts of gene products, RNAs, and proteins. This process is crucial for the adaptation and versatility of organisms in response to their environment, allowing cells to express specific RNAs and proteins when required. The *lac operon*, discovered by Jacques Monod in 1961, was the first studied instance of gene regulation. The expression of some enzymes involved in lactose metabolism in Escherichia coli bacteria is only triggered by the presence of lactose and the absence of glucose. In multicellular organisms, gene regulation drives cellular development and differentiation in embryos, resulting in different cell types that possess unique gene expression profiles originating from the same genome sequence. Variations in gene expression profiles can lead to differences in RNA/protein abundance and ultimately impact the phenotypic traits of the cells.

The genetic information model, also known as *the central dogma* of molecular biology (see **Figure 1a**), explains the transfer of genetic information from DNA to RNA (through transcription) and from RNA to protein (through translation). This transfer of genetic information typically follows a unidirectional flow, although some information flow can occur from RNA to DNA.

Transcriptional regulation of gene expression is controlled by sequence-specific proteins known as transcription factors (TFs), which bind to specific regulatory regions in the genome. The promoter region is a central regulatory element (RE) that initiates gene transcription and is located near the transcription start site (TSS) of the gene (see **Figure 1b** for gene structure). The initial product of gene transcription is a pre-mRNA consisting of 5' and 3' untranslated regions (UTRs), exons, and introns. The pre-mRNA is then spliced to generate mature RNA by removing introns and UTRs and joining exons together (see **Figure 1a**). The splicing process is also regulated by specific sequences within introns and exons and by sequence-specific proteins [5].

Enhancers are a type of REs that positively control transcription by forming a physical link with specific promoters, facilitated by co-factor proteins (as shown in **Figure 2a**). While the promoter sequence near the TSS can assemble the RNA-POL-II (hereafter referred to as POL2) machinery, transcription is usually low without enhancer-promoter (EP) links that help to stabilize the POL2 machinery [2].

Silencers are REs that negatively control transcription, namely decrease their target gene expression levels. While REs such as enhancers, insulators, and promoters, have been studied extensively over the past decade, silencers received less attention, mainly because it has

been harder to validate them experimentally [3,6]. Distal silencers can form physical links with specific promoters as enhancers do [7]. Characterizing functional silencers is currently an area of great interest with potential impact on lineage development and disease studies [3,8].

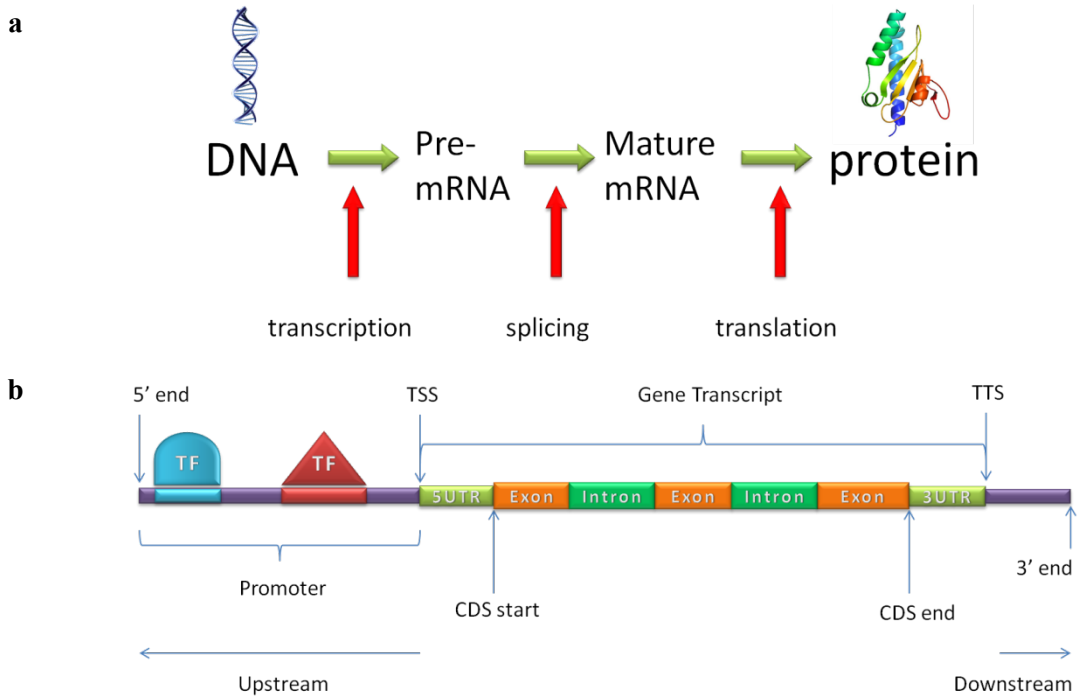


Figure 1. Gene structure and regulation**.** (a) The central dogma of molecular biology. The genetic information flow starts from the DNA and ends with the protein product. The gene regulation process controls the transcription step. (b) Gene structure. The promoter is bound by sequence specific TFs proximal to the gene. TSS – transcription start site, TTS – transcription termination site, CDS – coding DNA sequence, TF – transcription factor, 5UTR – 5 prime un-translated region, 3UTR – 3 prime un-translated region.

1.1.3 Chromatin organization

Chromatin is a complex mixture of macromolecules that resides in the nuclei of cells. It is composed of DNA, RNA, and protein, and serves several essential functions, including compacting and condensing DNA, facilitating cell division, protecting DNA from damage, and regulating gene expression and DNA replication. The organization of chromatin is regulated by a variety of factors, and this organization is critical for the proper execution of these functions.

In 1974, Don and Ada Olins discovered the *nucleosomes*, which play a crucial role in organizing the chromatin [9]. Nucleosomes consist of eight *histone* proteins that form an octameric core around which the DNA is wrapped twice, resulting in a unit of approximately 147 base pairs in length [10]. Series of higher order structures eventually form a *chromosome*, providing an additional layer of regulation for gene expression [11,12]. The folding and unfolding of DNA around nucleosomes are regulated by chemical modifications to the nucleosomes and by nucleosome displacement, as shown in **Figure 2b**.

The nucleosome core is composed of eight histones, two copies each of the histones H2A, H2B, H3, and H4. The degree to which DNA wraps around nucleosomes is regulated by chemical modifications made to specific sites on the histone proteins. For example, the event of histone H3 acetylation at lysine site 27 is denoted by H3K27ac. A single H3 methylation at lysine 4 is denoted H3K4me1. The function, such as an enhancer or a promoter, of those histone post translational modifications (PTMs) is mediated by specific complexes, which can read them and their combinatorial action [13]. For example, histone PTMs occur often in nucleosomes located near active enhancer regions, as depicted in **Figure 2b.b**. Active promoters that are bound by POL2 are surrounded by nucleosomes that carry the H3K27ac and H3K4me3 modification, as shown in **Figure 2b.c**. These histone modifications can be quantified using high-throughput techniques like chromatin immunoprecipitation sequencing (ChIP-Seq).

1.1.4 Enhancers

Enhancers are short, 50-1500 base-pairs (bp), regions of DNA that when bound by TFs increase gene's transcription [14,15]. These TFs recruit co-factor proteins acting as activators or repressors. The combination of all of these TFs and co-factors determines the enhancer activity in regulating specific genes. Activity of enhancers has been shown to correlate with specific markers of the chromatin (see **Fig. 2b**), which control DNA packaging and accessibility for transcription.

Enhancers were traditionally identified using enhancer trap techniques using reporter gene assays or by comparative sequence analysis between multiple species. For example, in flies, lacZ gene was used as a reporter and fused into the fly genome. If the reporter gene is fused near an enhancer then the lacZ expression reflects the expression pattern driven by that enhancer [16].

The emergence of more advanced genomic and epigenetic technologies allowed large-scale identification of enhancers. Next generation sequencing (NGS) methods (see below) enable the large-scale identification of TF binding sites, characterization of extensive epigenetic profiles across many cell types, and detection of ncRNAs. Therefore, accurate computational regulatory region discovery and linking such regions to their target genes are now attainable goals. An example of NGS-based method is DNase I hypersensitive sites sequencing (DNase-Seq), which enables identification of nucleosome-depleted, or open chromatin regions that can contain regulatory elements [17]. Computational methods for NGS data analysis include comparative genomics via sequence conservation of non-coding regions [18,19], clustering of known or predicted TF-binding sites [20], and supervised machine-learning approaches trained on known regulatory regions [21]. All of these methods have proven effective for regulatory

region discovery, but each has its own limitations, and each leads to some false-positive identifications [22].

Several consortia provided different high-throughput (HT) datasets covering hundreds of cell types and tissues for measuring enhancers and gene expression. The FANTOM5 consortium provided data from cap analysis of gene expression (CAGE) deep-sequencing HT method [23]. CAGE is used for measuring enhancer and gene's TSS expression. The ENCODE and Roadmap consortia provided DNase-Seq and ChIP-Seq HT data for the detection of nucleosome depleted genomic regions (i.e., open chromatin regions) and measuring histone modifications in the flanking nucleosomes of these regions [24,25].

FANTOM5 and ENCODE used their resources to predict functional interactions between enhancers and promoters of their target genes using pair-wise correlations. However, this method does not take into consideration the possibility that multiple enhancers contribute to enhancing the expression of the same gene [26]. In addition, this method is not capable of detecting cell type-specific EP links (i.e., EP links that are functional in few different cell types). Other projects seek EP links from contact interactions in the 3D genome architecture, which can be captured by Hi-C and ChIA-PET HT techniques [27–29]. However, currently a limited number profiles is available from Hi-C and ChIA-PET, and therefore, the confidence in the predicted EP links from these experiments is limited. It is important to note that functional association of an enhancer with a promoter (e.g., an EP link discovered using DHS-seq data) does not necessarily mean physical interaction (e.g., an EP interaction discovered using ChIA-PET data), and vice versa.

Elucidating cell type-specific EP links is a challenging task requiring more sophisticated methods that take advantage of multiple replicates per cell type instead of using naïve correlation-based methods. Detection of cell type-specific EP links can expand our understanding on gene regulation and may suggest different disease treatment approaches targeting enhancers in addition to the traditional gene-based therapies.

1.1.5 Silencers

Silencers are the repressive counterparts of enhancers. Silencers are DNA elements that when bound by repressive TFs reduce transcription from their target genes [3]. Although discovered more than 30 years ago [30,31], silencers have received less attention compared to enhancers, mainly because they were hard to identify experimentally [3,6].

In 2020, two seminal studies have advanced the experimental discovery of silencers in human and mouse cells [32,33]. Pang and Snyder developed the repressive ability of silencer elements (ReSE) method. In ReSE, candidate open chromatin DNA elements are cloned upstream of a promoter driving expression of a pro-apoptotic protein, Caspase-9. With silencing activity of the cloned DNA element, the promoter will not be able to drive Caspase-9 expression, not triggering apoptosis, and the cloned cell survives. Using this approach the

authors identified more than 2,500 and 1600 silencers in K562 myeloid and HepG2 hepatocyte cell lines, respectively. In a second study, Ngan et al. showed that target gene expression can be reduced as a result of chromatin interaction with distal silencers in mouse embryonic stem cells.

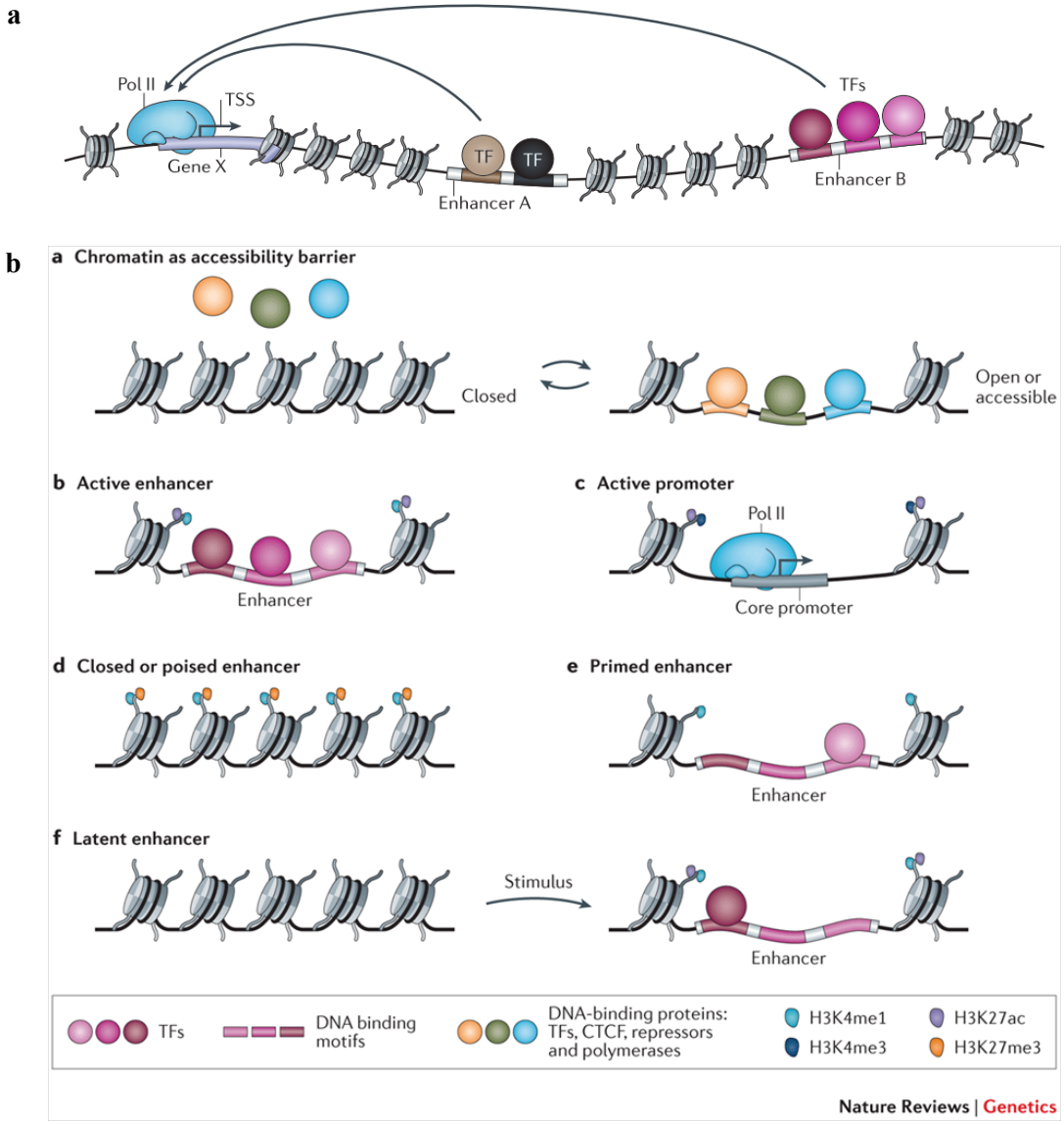


Figure 2. Enhancers and chromatin accessibility controlled by histone marks. (a) Enhancers located distal from gene X are linked with POL2 via co-factor TFs. Enhancers contain binding sites for sequence specific TFs. Nucleosomes are located in regions between enhancers and gene X. (b) Chromatin accessibility controlled by histone marks. These marks (H3K4me1/ H3K4me3/ H3K27ac/ H3K27me3) are found on the nucleosomes flanking open regulatory regions (enhancers or promoters). Open regions contain DNA binding sites for sequence specific TFs. Source: [2].

To identify silencer-promoter (SP) links, the authors pulled down chromatin interactions mediated by the Polycomb repressive complex 2 (PRC2), a key inducer of gene silencing, using the ChIA-PET HT technology. Deletion of certain PRC2-mediated interactions resulted with transcriptional derepression of their interacting genes with noticeable phenotypes such as embryonic lethality. The authors also showed that some of the PRC2-bound silencer elements can transition into active enhancers in other tissues during development. Later on, additional studies published experimentally identified silencers using reporter assays [34,35]. Jayavelu et al. identified 3,001 K562 silencers using the massive parallel reporter assay (MPRA) method [34]. Hansen and Hodges identified 21,125 silencers in B cells by testing the repression activity of open chromatin regions using the self-transcribing active regulatory region sequencing (STARR-Seq) method [35].

In spite of the abovementioned studies on silencers, there is still no established 'silencer epigenetic signature' in the distribution of epigenetic marks and the binding of repressor proteins that is common to all silencers; instead they may fall into various subclasses, acting in distinct mechanisms [3]. Given the vast knowledge on enhancer and promoter chromatin signature, comparing them to those computationally found within putative silencers could expand our current knowledge on silencers.

1.1.6 High-throughput omic technologies and the omics era

With the completion of the human genome project (HGP) in April 2003, generating the first sequence of the human genome, a new era has emerged referred as "The omics era". Following the HGP, the realization that the DNA is not the sole component regulating complex biological processes led to the rapid development of multiple molecular techniques to analyze additional layers of genomics, epigenomics, transcriptomics, proteomics, and metabolomics, altogether referred to as "Omics" (**Figure 3**). These techniques allow scientists to investigate complex biological systems and improve our understandings of disease outcomes.

Various HT technologies have been developed for investigating the information stored in biological molecules such as DNA, RNA and proteins. Each HT technology generates thousands to millions of values in a single experiment. Biological hypotheses can be drawn based on a single-omic analysis using a single HT technology, or based on a multi-omic analysis combining multiple HT technologies conducted at different cellular levels (**Figure 3**). Many datasets produced using HT technologies are freely available in public repositories, such as the ones provided by the National Center for Biotechnology and Information (NCBI). These datasets often require first extensive preprocessing to cope with high noise levels and then application of statistical and algorithmic methods to extract meaningful biological findings.

Next Generation Sequencing (NGS) is a general name for a plethora of very deep HT sequencing technologies developed over the past two decades. NGS can provide millions of short sequences in a single run. NGS has made genomic research several orders of magnitude

faster and cheaper than it was with Sanger sequencing [36], which was used in the human genome project.

In the next sections we introduce the relevant high-throughput techniques used in this thesis. We also describe in general how NGS data is preprocessed and used for RE identification and quantification.

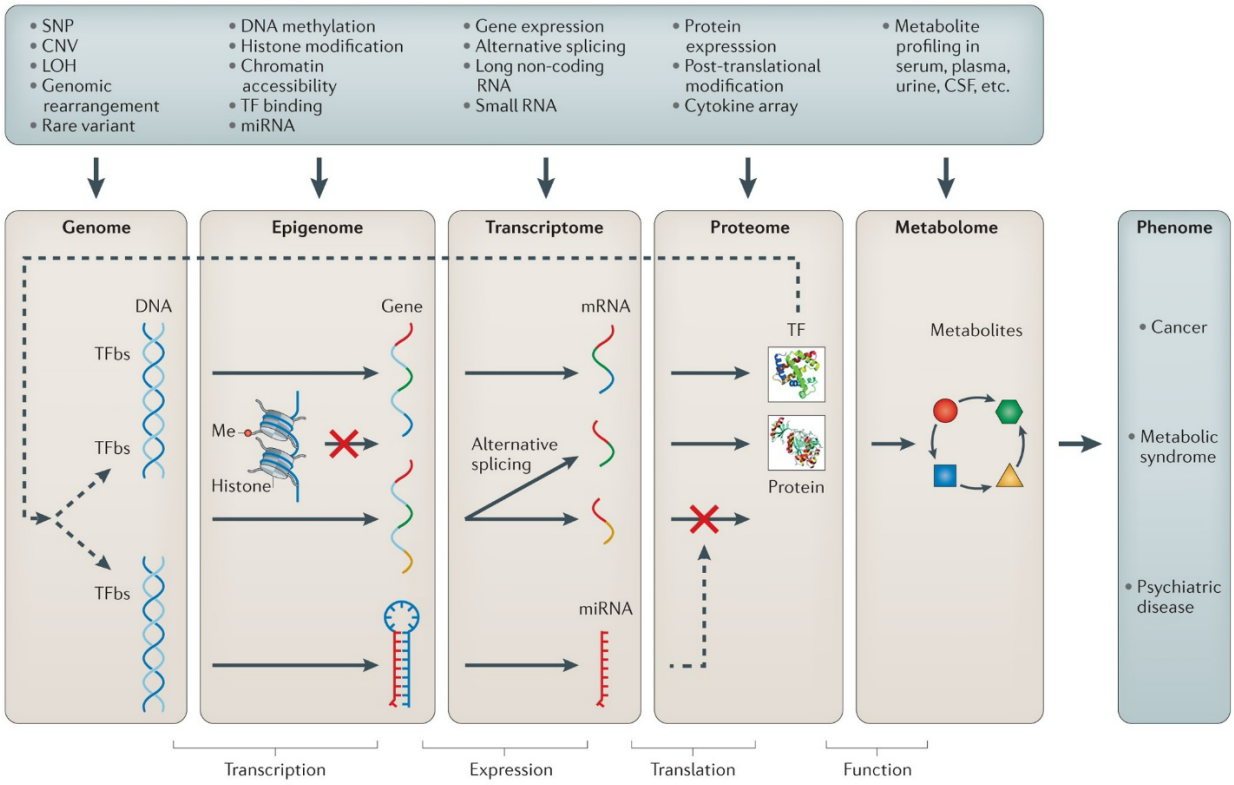


Figure 3. Omics data. The diagram shows diverse omics, from the genome, epigenome, transcriptome, proteome and metabolome to the phenome. The main diagram shows a simplified information flow from the lowest genomic level to the highest metabolomics level in a cellular system. The top part lists the different data types that can be measured in each level. Red crosses indicate inactivation of transcription or translation. SNP, single-nucleotide polymorphism; CNV, copy number variation; LOH, loss of heterozygosity; TF, transcription factor; miRNA, microRNA, CSF, cerebrospinal fluid; ME, methylation; TFbs, transcription factor binding sites. Source: [37].

1.1.6.1 Chromatin Immunoprecipitation sequencing (ChIP-Seq)

ChIP-Seq is an NGS method that identifies single protein attachments along the DNA [38]. ChIP-Seq extracts DNA fragments attached to a certain protein using chromatin immunoprecipitation (ChIP) and then performs DNA deep sequencing to identify binding sites (BSs) of the DNA-associated protein (**Figure 4**).

The detection of DNA-protein binding sites from ChIP-Seq read count data, commonly known as "peak calling", necessitates the development of computational tools. Among these tools, MACS stands out as a popular method [39]. MACS employs an empirical approach to model the shift size between two ChIP-Seq peaks flanking a given binding site on opposite

DNA strands. This modeling technique enhances the spatial resolution of predicted binding sites, as illustrated in **Figure 5**.

ChIP-Seq is a valuable technique that can be employed to detect histone modification sites across the genome. For example, targeting histone marks such as H3K4me1 and H3K27ac for detecting active enhancers, or H3K4me3 and H3K27ac for detecting active promoters (**Figure 2b**). In addition, ChIP-Seq can be used to detect BSs of P300 and POL2 proteins within enhancers and promoters, respectively. ChIP-Seq is extensively utilized for comparative analyses between different cell types by dissecting BS preferences for one or more TFs. For instance, in a seminal study, nine chromatin marks were mapped across nine cell types to pinpoint REs and establish connections between enhancers and their target genes [40].

1.1.6.2 Chromatin accessibility assays

Chromatin accessibility assays are methods for identifying nucleosome depleted genomic regions (also known as open chromatin regions). DNase-Seq, FAIRE-Seq and ATAC-Seq are such assays [16] [40–41]. For example, the DNase-Seq technique is briefly described in **Figure 6**. The inferred locations (also called peaks) from these assays are widely used for enhancer and promoter identification since these regions are known to be open and bound with activator proteins. These methods were extensively used to computationally predict global maps of enhancer promoter (EP) interactions [43–46].

Unlike the ChIP-Seq technique that can identify direct BSs of a single known protein, chromatin accessibility techniques can identify short sequence segments that contain potential BSs (also known as footprints) of multiple unknown proteins [47–49]. For example, Vierstra et al. detected a total of 4.5 million footprints from chromatin accessibility across 243 human cell and tissue types [48]. These footprints can be further utilized to narrow down genomic intervals searched for enriched DNA sequences called motifs in a process called motif finding as done in Hait et al. [50].

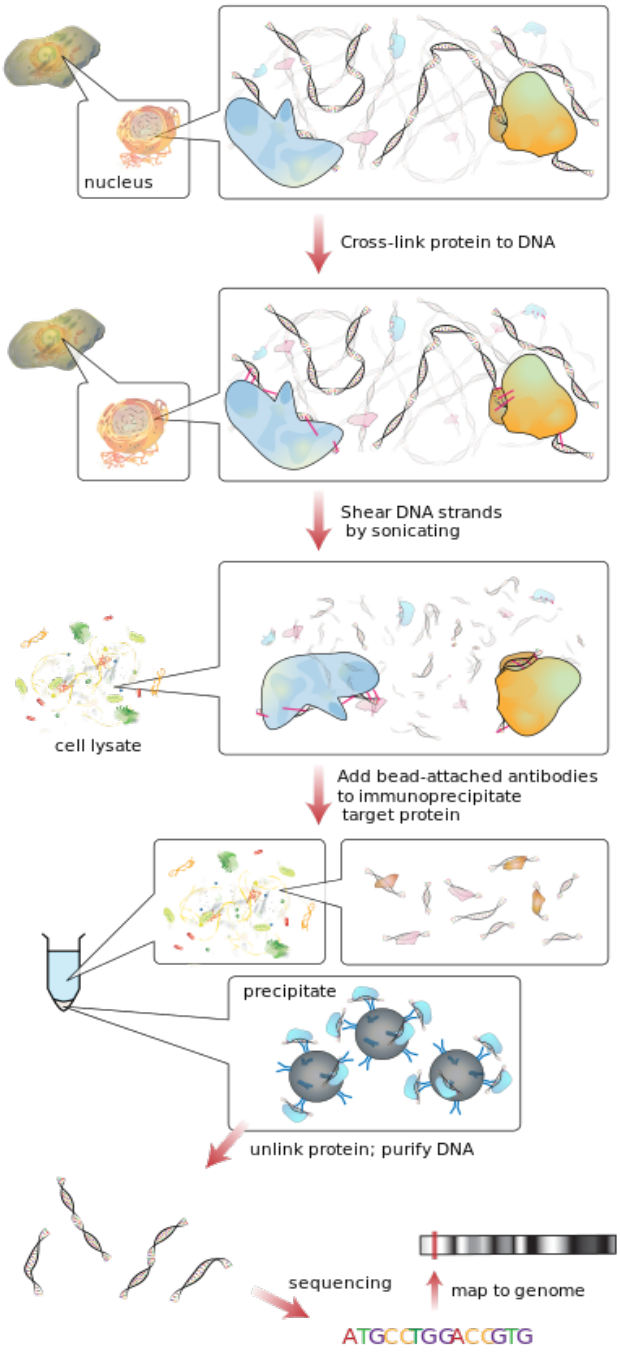


Figure 4. ChIP-Seq workflow. First, the DNA is isolated from the nucleus and cross linked to the protein to prevent detaching during sonication process. Second, the DNA is sheared by sonication. Third, a protein-specific antibody is used to immunoprecipitate the target protein and to select only DNA fragments attached to the protein. In the final step, the proteins are separated from the DNA fragments, which are then subjected to sequencing and alignment against a reference genome. Source: https://en.wikipedia.org/wiki/ChIP_sequencing

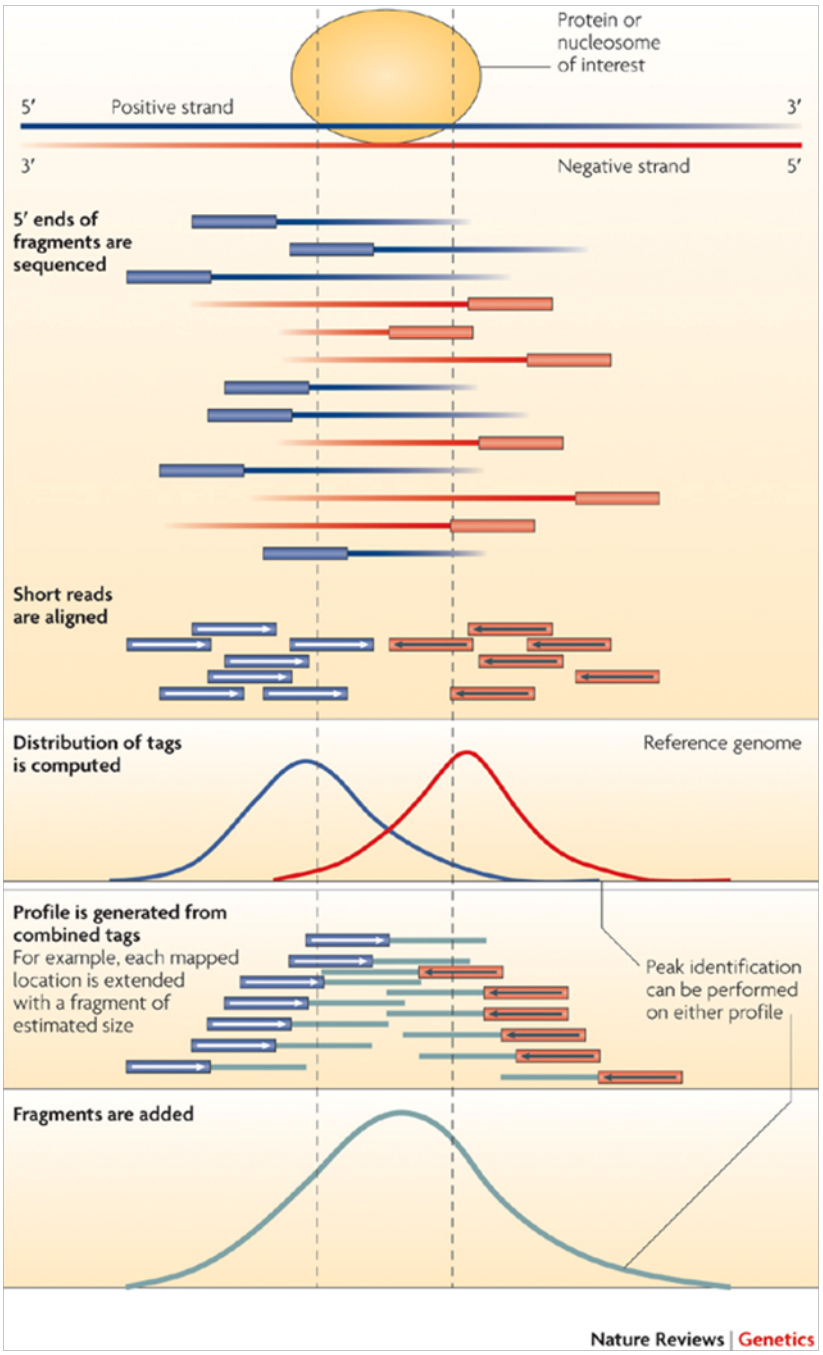


Figure 5. ChIP-Seq peak calling. In a ChIP-Seq experiment, DNA fragments are sequenced from the 5' end and aligned to the genome, resulting in two tag distribution peaks (one on each strand) that flank the binding location of the protein or nucleosome of interest. This specific pattern, which is specific to each strand, enables the detection of regions enriched with the protein of interest. To approximate the overall distribution of all fragments, each tag location can be extended based on an estimated fragment size in the correct orientation. By counting the number of fragments at each position, a representation of the fragment distribution can be generated. Source: [51].

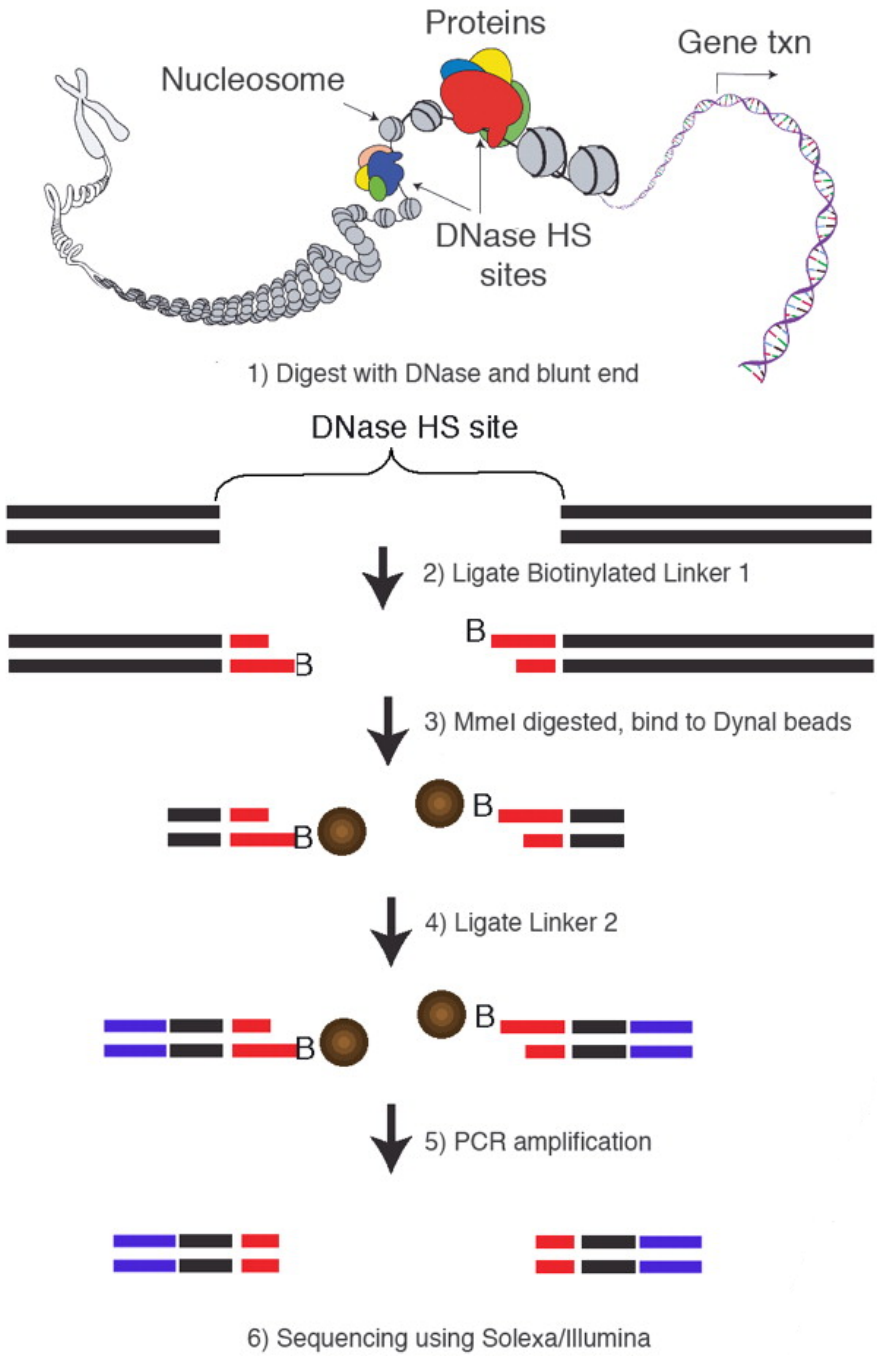


Figure 6. DNase-Seq workflow. The nuclei are released from the cells and are digested with optimal concentrations of DNase I enzyme. The DNA is then blunt-ended, extracted, and ligated to biotinylated linker 1 (red bars). Biotinylated fragments (linker 1 plus 20 bases of genomic DNA) are digested with MmeI enzyme and captured by streptavidin-coated Dynal beads (brown balls). Linker 2 (blue bars) is ligated to the 2-base overhang generated by MmeI, and the ditagged 20-bp DNAs are amplified by PCR, sequenced and aligned to the reference genome. Source: [52].

1.1.6.3 Methylation assays

Methylation assays quantify methylation levels at cytosine-guanine (CG) nucleotide sites (also known as CpG sites) within the genome. DNA methylation has a significant role in chromatin regulation and structure during development [53]. Changes in DNA methylation level have also been shown to contribute to cancer development and various diseases [54]. For example, a tumor suppressor gene silencing resulting from hypermethylation of CpG sites within its promoter is associated with tumorigenesis [54]. In addition, difference in methylation levels at enhancer elements between normal and tumor lung tissues has also been shown to contribute to tumorigenesis [55]. Methylation is also known to affect the DNA binding specificity of many TFs [56]. These changes and functions are governed by three classes of proteins that can write, erase, and read DNA methylation [57].

There are two types of methylation assays: chip-based and NGS-based. The chip-based technique measures a fixed set of CpG site probes from the genome. The number of probes in a chip varies from 27K (e.g., Illumina Infinium HumanMethylation27 BeadChip) to more than 850K (e.g., Illumina Infinium MethylationEPIC BeadChip). The whole genome bisulfite sequencing (WGBS) and the reduced-representation bisulfite sequencing (RRBS) are NGS-based techniques. RRBS technique is more focused towards sequence regions of high CpG content while WGBS captures whole genome CpG sites. In these methods, sodium bisulfite is used convert unmethylated cytosines into uracil. This enables methylation detection by distinguishing the methylated cytosines, which resist bisulfite treatment, from uracils. During sequence amplification, uracils are converted to thymines and methylated cytosines are converted to cytosines. Software tools, such as the PASH [58], identify genomic locations of the sequence (or, read) by comparing the bisulfite treated and original sequence. From the number of converted and unconverted reads at each individual CpG site, the total coverage and fractional methylation are reported.

1.1.6.4 Chromatin interaction capture methods

Chromatin interaction capture (also known as chromosome conformation capture – 3C) methods identify three-dimensional DNA-DNA physical contacts (**Figure 7**). Using these techniques one can analyze the spatial organization of the chromatin in a cell.

Among these techniques, the chromatin interaction analysis by paired-end tag (ChIA-PET) identifies interactions mediated by a protein of interest [59]. This technique incorporates ChIP-based enrichment, chromatin proximity ligation, paired-end tags, and sequencing (**Figure 7**). POL2 protein is often used in the ChIP step to identify enhancer-promoter interactions as it can indicate active gene expression. HiChIP is a similar method to ChIA-PET requiring 100-fold less input material [60].

ChIA-PET and HiChIP interaction data allow the study of gene regulation that depends on the 3D genome structure. Using such information one can validate EP links that were computationally predicted based on one dimensional HT technologies.

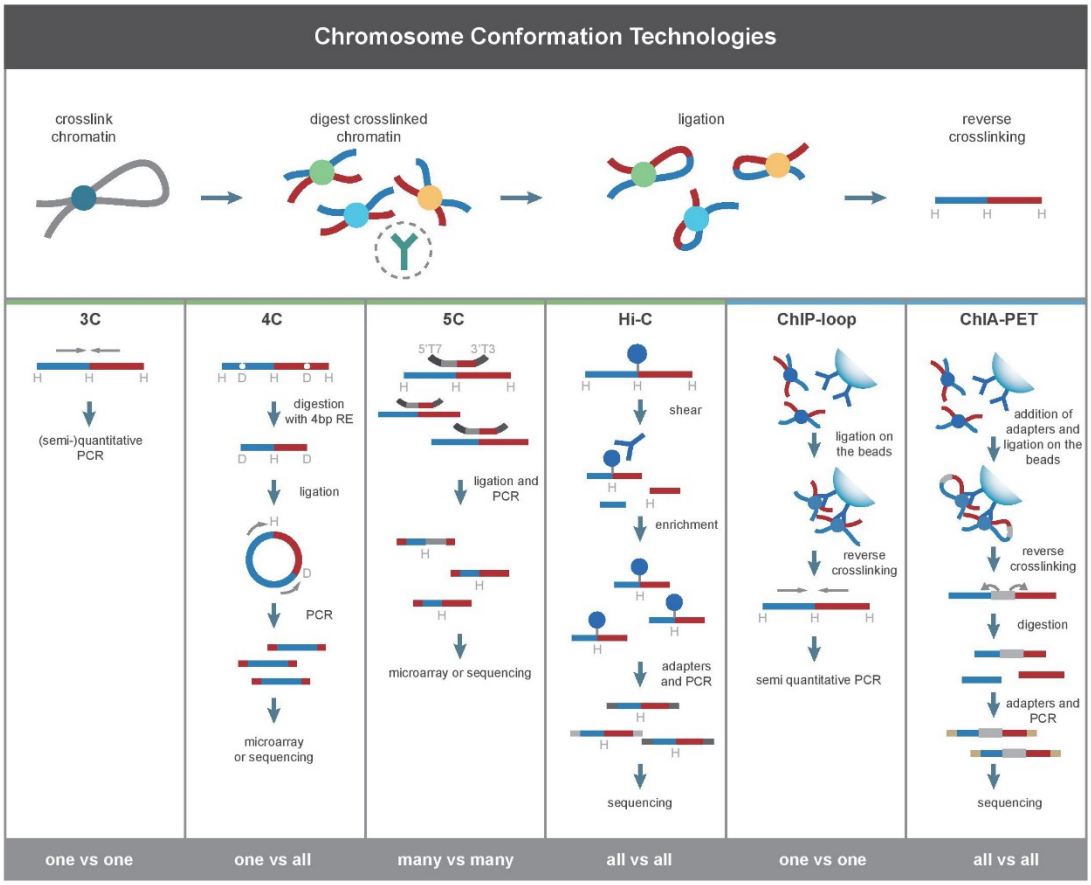


Figure 7. Chromatin interaction capture methods. The top panel shows the experimental steps common to all 3C methods: cross-linking, digestion, and ligation steps. The vertical panels show the specific steps required in each technique. Source: [61]

1.1.6.5 Limitations of the omics techniques used

1. DHS-seq or ATAC-seq identify regions of open chromatin. These signals are used as proxies for active regions, but an open state is not enough to establish activity. Thus, computationally inferred EP links using these technologies are not necessarily active.
2. TF ChIP-Seq profiles the binding of the TF to chromatin. It does not distinguish between functional and non-functional binding events.
3. 3C-type methods based on restriction enzymes are often considered to perform poorly in detecting enhancer-promoter interaction as they suffer from low resolution (recent variations based on MNase are significantly better for this task) [62].

1.2 Computational background

This chapter provides the computational foundation for this thesis. Each section discusses a different computational aspect. More information on the computational problems addressed is provided in the references for each section.

1.2.1 Data representation

In this section we lay out the data structures used in this thesis.

1.2.1.1 Signal data

The signal data of genomic features (e.g., enhancers and promoters) is represented as a real matrix $D \in \mathbb{R}^{n\,x\,m}$, where n is the number of genomic features and m is the number of samples. The rows contain the signal pattern of the epigenomic features across the samples, and the columns contain the signal profile of the samples across the genomic features. Samples can represent different cell types, conditions, or individuals.

Each entry, $D_{i,j}$, in the matrix represents counts or normalized values, depending on the computational method used. Normalized values are computed by dividing the $D_{i,j}$ count number by the genomic size of feature i and by the library size (that is the total number of mapped reads) of sample j. Normalized values are usually measured in units of Read per Kilobase exon per Million mapped reads (RPKM).

Samples in D are sometimes annotated with auxiliary information containing one or more labels such as cell type, disease, or treatment. In our analysis, we map each sample to its cell type.

1.2.1.2 Genomic position data

Genomic positions of features are defined by three fields: chromosome, start position, and end position. By definition, start position is always smaller than the end position. The position can be positively or negatively stranded, or un-stranded. For example, if the genomic feature is a gene then in terms of gene transcription, a positive strand denotes transcription from the start to the end whereas the negative strand denotes the reverse. Regulatory elements such as enhancers are un-stranded since they do not show transcription preference to any direction.

Genomic positions are usually stored in a Browser Extensible Data (BED) file format. A BED file contains, in addition to the mentioned three fields of genomic positions and the strand, other informative fields such as the feature name and the BS score (e.g., p-value or intensity computed using a peak caller tool), and other fields controlling the genomic feature visualization in UCSC genome browser.

1.2.2 Genomic data analysis

In this section we lay out the common downstream analyses used in this thesis.

1.2.2.1 Enrichment analysis

Enrichment analyses were originally developed to detect over-represented classes of genes within a set of genes produced by a certain analysis [63,64]. In this thesis, we used the same concept to detect over-represented classes of protein BSs or SNPs overlapping a set of genomic regions. The classes reflect biological knowledge or experimental results, e.g., a set of SNPs associated with a disease from Genome Wide Association Studies (GWAS) [65], or a set of whole genome BSs of a protein produced by a ChIP-Seq experiment.

The hypergeometric statistical test is used for enrichment analysis. For convenience, we assume that we have classes of genes, and we would like to find whether one or more of these classes is over-represented in our target set of genes.

Let G be the set of all genes (the background gene set), T be a set of genes of interest (the target set), and A be a set of genes that are known to be involved in a particular biological process (the class). We can test the significance of the intersection $I = T \cap A$ by comparing $|I|$ to the number of genes that are expected to be in the intersection by chance.

The null hypothesis of the test is that the genes in T were randomly selected without replacement from G. Under the null hypothesis, the size of the intersection $|T \cap A|$ follows a hypergeometric (HG) distribution. The probability that exactly x selected genes belong to A is:

$$P_{hg}(|G|, |A|, |T|, x) = \frac{\binom{|A|}{x}\binom{|G| - |A|}{|T| - x}}{\binom{|G|}{|T|}}$$

Thus, the p-Value is:

$$\Pr(X \geq |T \cap A|) = \sum_{x \geq |T \cap A|}^{\min(|T|,|A|)} P_{hg}(|G|, |A|, |T|, x)$$

If there are N gene groups and M a priori gene sets, then there will be $N * M$ statistical tests. This means that there is a high chance of false positives (FPs), therefore it is important to use multiple testing correction methods (such as Bonferroni [66] or FDR [67]) to control the number of FPs.

1.2.2.2 Motif finding

Motif finding tools aim to identify short DNA sequences (usually of 6-20 bp in size; also known as motifs) that are enriched within a set of genomic regions. Such tools can identify motifs de-novo, e.g., AMADEUS and MEME [68,69], or by scanning for occurrences of known TFBSs motifs, e.g., FIMO [70].

Scanning the genomic regions for known TFBSs motifs requires a database of position weight matrices (PWMs), each representing a TFBS preference (**Figure 8**). Popular databases of PWMs are HOCOMOCO and JASPAR [71,72]. HOCOMOCO, for example, assembled its database by inferring TFBSs from ChIP-Seq experiments.

A typical motif finding analysis is usually done on a small set of genes' promoters. However, in this thesis, finding all TF motif occurrences in a large set of enhancers that were computationally linked to promoters, each hundred of bases long, is prone to high false-positive rate. Therefore, to limit the search space, one can restrict the motif finding to digital genomic footprints (DGF) regions [48], which are very short segments that are more likely to contain true TFBSs, found within enhancers.

	1	2	3	4	5	6
A	0.1	0.8	0	0.7	0.2	0
C	0	0.1	0.5	0.1	0.4	0.6
G	0	0	0.5	0.1	0.4	0.1
T	0.9	0.1	0	0.1	0	0.3

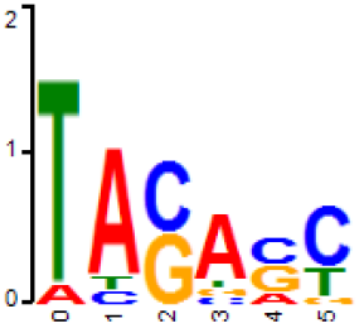


Figure 8. A PWM example with its logo illustration. The matrix shows the TF binding site preference to different nucleotides in each position in terms of the probability of occurrence of each nucleotide in each position. The logo is a visualization of the matrix.

1.2.3 Linear mixed effect models (LMMs)

Computationally linking enhancers to their target promoters is usually done using correlation based methods. For example, ENCODE and Roadmap Epigenomics predicted the mapping of enhancers to their target promoters using pairwise correlations across many samples [25,43]. FANTOM5 expanded the pairwise correlation methodology to map k nearest enhancers to each promoter by using a linear regression [23]. These methods tried to map enhancers to their target promoters in a non-cell-type specific manner. To infer cell-type specificity of EP links, one could simply add additional parameters to the linear regression model (in addition to the k nearest enhancers) such as the sample's cell type label. However, adding hundreds of different cell types to the model will make the model poor and hard to interpret since the number of parameters would be close to the number of samples. In addition, when applying standard linear regression, it is assumed that the samples are independent. However, this is not true when multiple samples belong to the same cell type. To cope with these problems, linear mixed effects models can be used.

A linear mixed model (LMM) is a type of statistical model that can be used to analyze data that has a hierarchical structure. This means that the data can be grouped into nested levels, such as students within classrooms or patients of specific doctors. In this thesis, samples are grouped into C different cell types. In addition to the variability between samples, mixed effect models allow taking into account also the variability between groups of samples.

A LMM is defined as $y = X\beta + Z\gamma + \epsilon$ where X is an $n \times p$ matrix of the predictor variables, β is the $p \times 1$ fixed-effect regression coefficient, $Z\gamma$ is the random effect and ϵ is a random error. γ is a C-long vector of random effects to be predicted, and Z is a nxC design matrix that groups the samples by their cell types. We assume that γ and ϵ are normally distributed. When the data is zero-inflated, one could use the generalized linear mixed effect models (GLMMs) assuming the random parameters are zero inflated negative-binomial distributed. Fitting LMMs is done using various approaches such as the Expectation-Maximization (EM) and the Newton-Raphson algorithms [73].

While most of the C predicted values in a random effect vector are expected to be close to zero since γ is normally distributed with mean zero, values that deviate from zero (known as outliers) are the interesting ones and one could use them to detect cell type-specific characteristics of the fitted LMM model. This intuition is extensively used in the second paper of this thesis.

1.2.4 Deep learning

Deep learning (DL) has become a dominant paradigm in science, industry and technology in recent years, spanning numerous applications in diverse domains. Here we give only a very short introduction. For more on the topic see, e.g., the book of Goodfellow, Bengio and Courville [74] on mathematical and conceptual background, and Keras [75] or PyTorch [76] for hands-on practice on DL.

DL is a class of machine learning algorithms that allows computers to learn and process datasets in a way that is inspired by the human brain anatomy. DL models are used to recognize complex patterns in pictures, text, videos, sounds, and other data and to infer accurate predictions. DL models contain multiple interconnected layers of nodes (called neurons) and apply repeated exchange of signals among them to progressively extract meaningful representations of the raw input (**Figure 9**). The word "deep" in "deep learning" refers to the number of layers through which the raw data is transformed and processed.

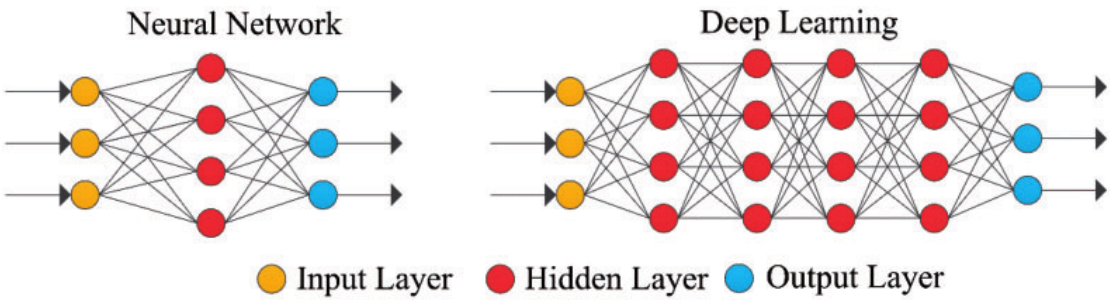


Figure 9. Deep learning diagram. Left: a neural network defined by one red hidden layer. Right: a deep learning network defined by multiple consecutive red hidden layers. Each circle is a neuron. Source: [77].

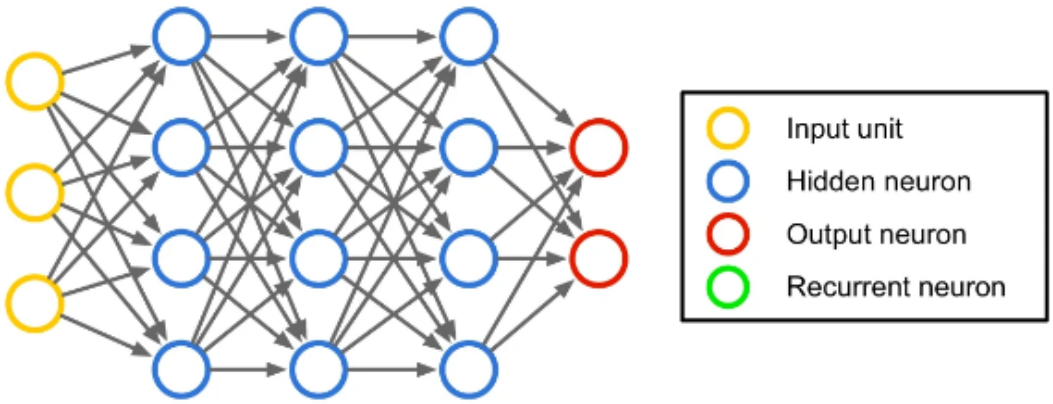
DL has many applications. Self-driving cars use DL to detect road signs and pedestrians [78]; defense and urban systems use DL to flag areas of interest in satellites images [79]; medical image screens use DL to detect patients with cancer [80]; DL is extensively used to analyze electronic health records (EHRs) in order to predict future disorders (e.g., medGPT [81]), and many more. These use cases can be broadly grouped into four categories: computer vision, speech recognition, natural language processing (NLP), and recommendation engines. Computer vision aims to derive computationally meaningful information from images and videos. Speech recognition analyzes human and other species speech to infer different patterns, tone, language, and accent. NLP gathers insights and meaning from text data and documents. Recommendation engines track user activity and develop personalized recommendations.

DL networks typically have three basic components (**Figure 9**): an input layer composed of neurons that accept the raw data. The hidden layers receive processed data from the input layer and further process the information at different levels. The output layer provides the prediction. A DL model has numerous different parameters to be learned while training. For example, edges connecting the neuros between layers can have different weights as parameters. DL models that output "yes" or "no" answers have only one neuron in the output layer. Those that output a wider range of answers have more output neurons.

As in many ML techniques, hyperparameter tunning (e.g., number of neurons in each layer) involves identifying the most effective hyperparameter values for a learning algorithm and utilizing the optimized algorithm on diverse datasets. This set of hyperparameters aims to enhance the model's performance by minimizing a predetermined loss function, leading to improved outcomes with reduced errors. A common technique to find the best possible configuration of hyperparameters is grid search: We select a set of concrete values in the range of each hyperparameter, and the set of these values across the parameters constitutes a grid. Subsequently, we systematically explore all possible combinations of grid values, and identify the set that results in the minimum loss function.

There are three common families of architectures that are used to connect neuronal layers in DL (**Figure 10**): feed-forward (FF), convolutional and recurrent [82]. FF is the simplest architecture, where every neuron of layer i is connected to some or all neurons in layer $i + 1$, and edges connecting between two layers have different weights, which are the model's scalar parameters to be learned (**Figure 10a**). For each neuron, the sum of the products of the incoming edge weights and their inputs is calculated and is the output propagated on each outgoing edge from the neuron. FF is used for generic prediction when there is no special relationship between different parts of the input data. In a convolutional neural network (CNN), the neurons in one or more hidden layers perform convolutions. For example, for 2D matrix of an image that serves as input the neurons scan each position of the input matrix computes a local weighted sum of its neighborhood and output a value [83] (**Figure 10b**). Here, the weights on the edges connecting the convolutional layers are convolution kernels (or, filters) to be learned. CNNs are useful in tasks where the input data has some spatially invariant patterns, i.e., they are not sensitive to the object's position in the matrix (e.g., a CNN will recognize a cat face in any position in a picture). In Recurrent neural networks (RNNs), derived from FF networks, connections between neurons can create a cycle, allowing the output of some neurons to affect subsequent input to the same neurons. RNNs are designed for sequential or time series data. The hidden layers of the RNN can be thought of as memory states that retain information from the input sequence that has been observed so far (**Figure 10c**). These memory states are updated at each time step. RNNs are useful for learning relationships between different parts of the input data (examples are shown in the next section). A DL model architecture can be a combination of one or more the three families.

a



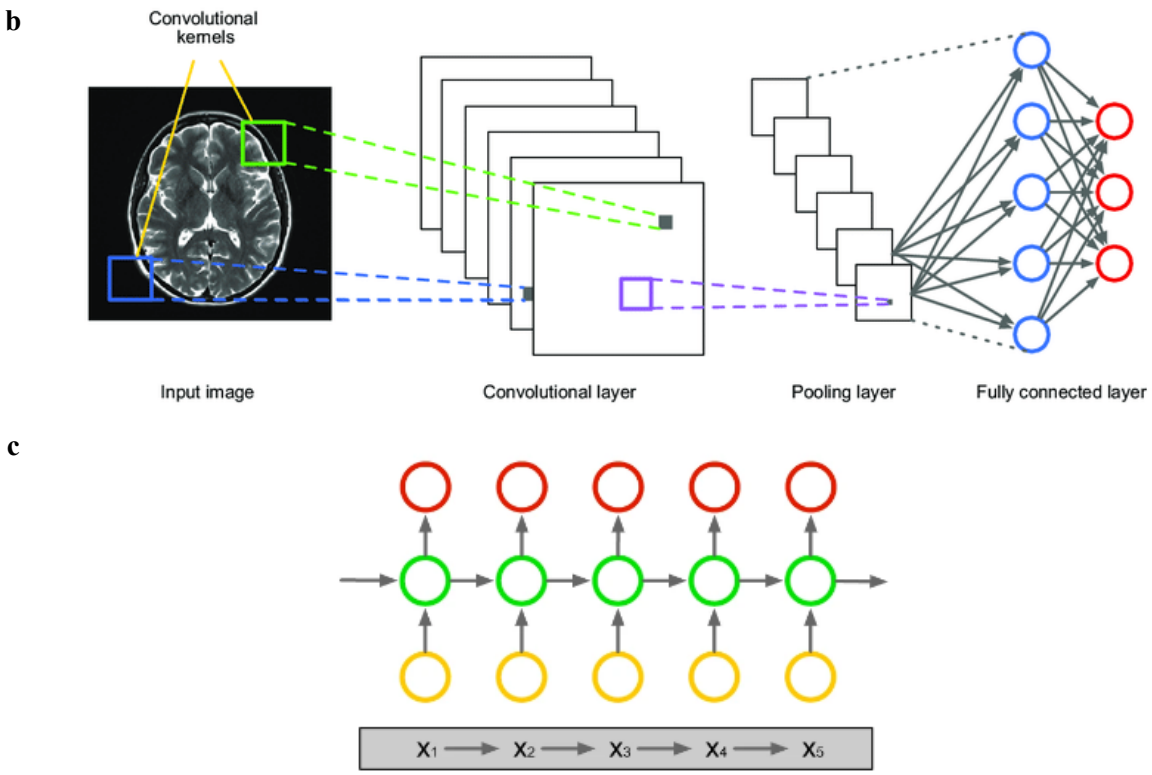


Figure 10. Common deep learning families. (a) Feed-Forward (FF) deep neural network. The information moves in one direction from the input layer (yellow circles) through one or more hidden layers (blue circles), to the output layer (red circles). (b) Convolutional neural network (CNN). A CNN typically consists of three types of layers contained between the input (e.g., a neuroimage input) and output layers. The convolutional layer generates feature maps by moving convolutional kernels across the preceding layer. The pooling layer serves to reduce the dimensionality of the prior convolutional layer. Lastly, the fully connected layer is responsible for making predictions based on the processed data. (c) A recurrent neural network (RNN) is designed to handle sequential data. Each recurrent neuron (green circles) in the network is tasked with encoding historical data by accepting both the current input element and the state vector from the preceding neuron. This produces a hidden state which is then passed on to the succeeding neuron. The RNN architecture, therefore, encodes not just individual information but also the dependency between the elements of a sequence like x1 → x2 → x3 → x4 → x5. Source: [84].

Training a DL for a prediction task is done using as input a labeled dataset in the form of (X_i, Y_i) where each X_i is the ith input and Y_i is the output label. Each training point X_i is fed to the network, the output label Y_i' is computed by the network, evaluated against the true label Y_i, and a loss function $L(Y_i', Y_i)$, quantifying the cost of error is computed. The network's parameters are trained by calculating the gradient of the loss function with respect to the parameters, and the parameters are slightly adjusted in the direction of the gradient in a process called back-propagation, which minimizes the loss function. After a sufficient number of

iterations, the network is expected to converge to a minimum loss function, producing the final trained DL model.

Successful use of DL raises several challenges. First, DL algorithms tend to give better results when they are trained on large amounts of high-quality data. Incorrect or irrelevant data can have a negative impact on DL models. For example, if non-animal images were accidentally added to a dataset of animal images, a deep learning model might classify an airplane as a bird. To avoid inaccurate results, one must perform data preprocessing, cleaning and processing large amount of data, before training a DL model. This process requires a large amount of work and data storage capacity. In addition, DL algorithms require large processing power since they are computationally intensive.

1.2.4.1 Deep learning in genomics

Classical Machine learning (ML) algorithms, e.g., support vector machines and logistic regression, have been extensively used in genomics research for decades [85]. The difference between DL and the standard ML methods used in genomics, is that DL models are much more flexible and capable to extract non-linear relationships between different features in the input data. DL requires great care to train on and to interpret the underlying biology in genomics [82].

Input data to DL model should first be transformed to a matrix of real values. In genomics, one input is usually a DNA sequence, in which the nucleotides A, C, G, and T are encoded as [1,0,0,0], [0,1,0,0], [0,0,1,0], and [0,0,0,1] [82]. Neurons that are the first to read the input constitute the input layer. Subsequent hidden layers further process the transformed input. The output layer of the DL is the prediction of interest (e.g., the probability that the input DNA sequence is a promoter).

Accuracy of the prediction is usually measured by calculating the precision, recall, F1 and the area under precision-recall curve (AUPRC). AUPRC values are preferable to the area under the ROC curve, since genomic datasets are often highly imbalanced (e.g., there are many more SNPs that are not disease causing than SNPs that are disease causing). In some computer vision tasks, DL models with more than 100 layers have been proven useful. However, in genomics applications, fewer than five layers is usually sufficient [82]. The most important factor for a successful DL model is the number of labeled examples, which should be at least several thousands [82].

Here are some examples of using DL in genomics. Luo et al. built a DL model to predict N^6-methyladenosine (m^6A) sites, used in silico mutagenesis and discovered that cis-element motifs that govern the m^6A deposition are located largely within the 50 nt downstream of the m^6A sites [86]. Fudenberg et al. built a CNN model to predict 3D genome structure, and used in silico mutagenesis to reveal that the CCCTC-binding factor (CTCF) BSs are the most important elements for 3D structure establishment [87]. Routhier et al. built a CNN model to predict nucleosome positioning in Saccharomyces cerevisiae directly from the DNA, and used

in silico mutagenesis to evaluate the effect of every single mutation in the genome on the nucleosome positioning [88].

A hard task in DL is model interpretation. Genomic researchers are usually more interested in the underlying biological mechanisms discovered by the trained DL model rather than prediction accuracy itself. For example, when one wishes to build a DL model that predicts EP links based on chromatin data, then the hope is to discover a novel gene-regulation grammar encoded in the trained model. DL models can achieve state-of-the-art predication accuracy, but, it is more challenging to interpret them compared to standard machine learning models.

There are a few methods that allow DL model interpretation in genomics. The simplest one is similar to in silico mutagenesis. Given a particular data point, each feature of the data point can be systematically varied while the rest of the features are held fixed (e.g., changing a nucleotide from A to C in one position in the input DNA sequence). Precomputing the DL model for each such variation allows us to track how the network's output changes in response to changes in the data point. This approach is easy to implement, but it can be computationally expensive, as the network must be re-evaluated for every mutation of the data point. A computationally tractable approximation to this approach is to take the derivative of the network's output with respect to each feature of the data point. This can be done using back-propagation, and it conveys the sensitivity of the output to small perturbations in the input features. Features with large positive or negative derivatives may have more influence on the outcome. Several variations of the derivative-based interpretation are available such as the integrated gradients [89] and DeepLift [90]. DL interpretation in genomics is currently an active area of research.

In the last few years there is a growing number of studies utilizing DL in genomics. The use of DL for regulatory genomics tasks, such as modeling TFBSs, has been particularly popular [91]. Examples in this field include tools predicting sequence specificity of DNA- and RNA-binding proteins and of enhancer and cis regulatory elements, gene expression, methylation and alternative splicing. These tools usually use DNase-Seq, ATAC-Seq, ChIP-Seq and more as input [92–95]. Enhancers were identified based on epigenetic profiles available from the ENCODE consortium [96–98]. DNA methylation state influencing gene expression has been inferred by using data from 3D DNA-DNA contacts [99]. Most of these tools used CNN or RNN for their tasks, which are well suited for modeling regulatory elements. Overall, using DL in genomics holds great promise to achieve more accurate predictions than standard ML methods and to decipher the gene-regulatory grammar.

2. Predicting global enhancer-promoter maps

Spatiotemporal gene expression patterns are governed to a large extent by the activity of enhancer elements, which engage in physical contacts with their target genes. Chromatin conformation capture assays, from which enhancer-promoter (EP) links can be derived, are still not available for many cell types and tissues. In contrast, large consortia, such as ENCODE, Roadmap Epigenomics and FANTOM5 [23–25], have produced numerous epigenetic datasets (e.g., open chromatin assays such as DNase-seq) covering hundreds of cell types. These datasets can be used to develop computational methods for linking enhancers to their target genes.

This study started during my MSc and was continued in the first year of my PhD research. In my MSc thesis we developed FOCS (*F*DR-corrected *O*LS with *C*ross-validation and *S*hrinkage) [100], a computational method for predicting EP links based on correlated activity patterns across many samples covering hundreds of different cell types. FOCS infers global EP links, i.e., EP links with significant high correlation between enhancer and promoter activity patterns across many samples. We applied the method on a small set of 246 samples (covering 23 cell types) from Global-Run-On sequencing (GRO-seq) data, which experimentally identifies functional enhancers and promoters.

During my PhD studies we expanded the scope and the validation of FOCS dramatically by using 10-fold more experimental data to train it. We applied the method and compared it to extant methods on 2,384 additional DNase-seq and CAGE samples collected from ENCODE, Roadmap Epigenomics, and FANTOM5 consortia [23–25]. In ENCODE dataset, analysis was not restricted to protein-coding genes. We also added HiChIP chromatin interaction data (in addition to ChIA-PET and eQTL data from the MSc studies) as an external source for evaluating the performance of our method.

In the following chapter we describe both the original FOCS algorithm that was part of the MSc thesis and the extended analysis and validation that were part of the PhD study. The complete project was published in *Genome Biology* in 2018 [101]. Some large supplementary files that were part of the analysis are available on the journal's website and links are provided in the chapter.

2.1. Results

The FOCS procedure for predicting enhancer-promoter links

We set out to develop an improved statistical framework for prediction of EP links based on their correlated activity patterns measured over many cell types. As a test case, we first focused on ENCODE's DHS profiles [43], which constitute 208 samples measured in 106 different cell lines (**Methods**). This rich resource was previously used to infer EP links based on pairwise correlation between DHS patterns of promoters and enhancers located within a distance of ±500 kbp. Out of ~42M pairwise comparisons, ~1.6M pairs showed Pearson's correlation>0.7 and were regarded as putatively functional EP links [43]. However, Pearson's correlation is sensitive to outliers and thus may be prone to high rate of false positive predictions. This is especially exacerbated in cases of sparse data (zero inflation), which are prevalent in enhancer activity patterns, as many of the enhancers are active only in a limited set of conditions. In addition, the combinatorial nature of transcriptional regulation in which a promoter is regulated by multiple enhancers is not considered by such pairwise approach.

To address these points we developed a novel statistically-controlled regression analysis scheme for EP mapping, that we dubbed FOCS. Specifically, FOCS uses regression analysis to learn predictive models for promoter's activity from the activity levels of its k closest enhancers, located within a window of ±500 kb around the gene's TSS. (Throughout our analyses we used $k = 10$). Importantly, to avoid overfitting of the regression models to the training samples, FOCS implements a leave-cell-type-out cross validation (LCTO CV) procedure, as follows. In a dataset that contains samples from C different cell-types, for each promoter, FOCS performs C iterations of model learning. In each iteration, all samples belonging to one cell-type are left out and the model is trained on the remaining samples. The trained model is then used to predict promoter activity in the left-out samples (**Fig. 2.1**).

We implemented and evaluated three alternative regression methods: ordinary least squares (*OLS*), generalized linear model with the negative binomial distribution (*GLM.NB*) [102] and zero-inflated negative binomial (*ZINB*) [103]. GLM.NB accounts for unequal mean-variance relationship within subpopulations of replicates. ZINB is similar to GLM-NB but also accounts for excess of samples with zero entries (**Methods**). For each promoter and regression method, the learning phase yields an activity vector, containing the promoter's activity in each sample as predicted when it was left out. FOCS applies two non-parametric tests, tailored for zero-inflated data, to evaluate the ability of the inferred models (consisting of the k nearest enhancers) to predict the activity of the target promoter in the left-out samples. The first test is a "*binary test*" in which samples are divided into two sets, positive and negative, containing the samples in which the promoter was active or not, respectively, based on their measured signal

(We used a signal threshold of 1.0 RPKM for this classification). Then, Wilcoxon signed-rank test is used to compare the predicted promoter activities between these two sets (**Fig. 2.1**). The second test is an "*activity level test*", which examines the agreement between the predicted and observed promoter's activities using Spearman's correlation. In this test, only the positive samples (that is, samples in which the measured promoter signal is ≥1.0 RPKM) are considered. Gene models with good predictive power should discriminate well between positive and negative samples (the binary test) and preserve the original activity ranks of the positive samples (the activity level test), and models that pass these tests are regarded as statistically cross-validated. Of note, these two validation tests evaluate each promoter model non-parametrically without assuming any underlying distribution on the data when inferring significance. Next, FOCS corrects the p-values obtained by these tests for multiple testing using the Benjamini and Yekutieli (BY) FDR procedure [104] with q-value<0.1. The BY FDR procedure takes into account possible positive dependencies between tests while the more frequently used Benjamini and Hochberg (BH) FDR procedure [67] assumes the tests are independent.

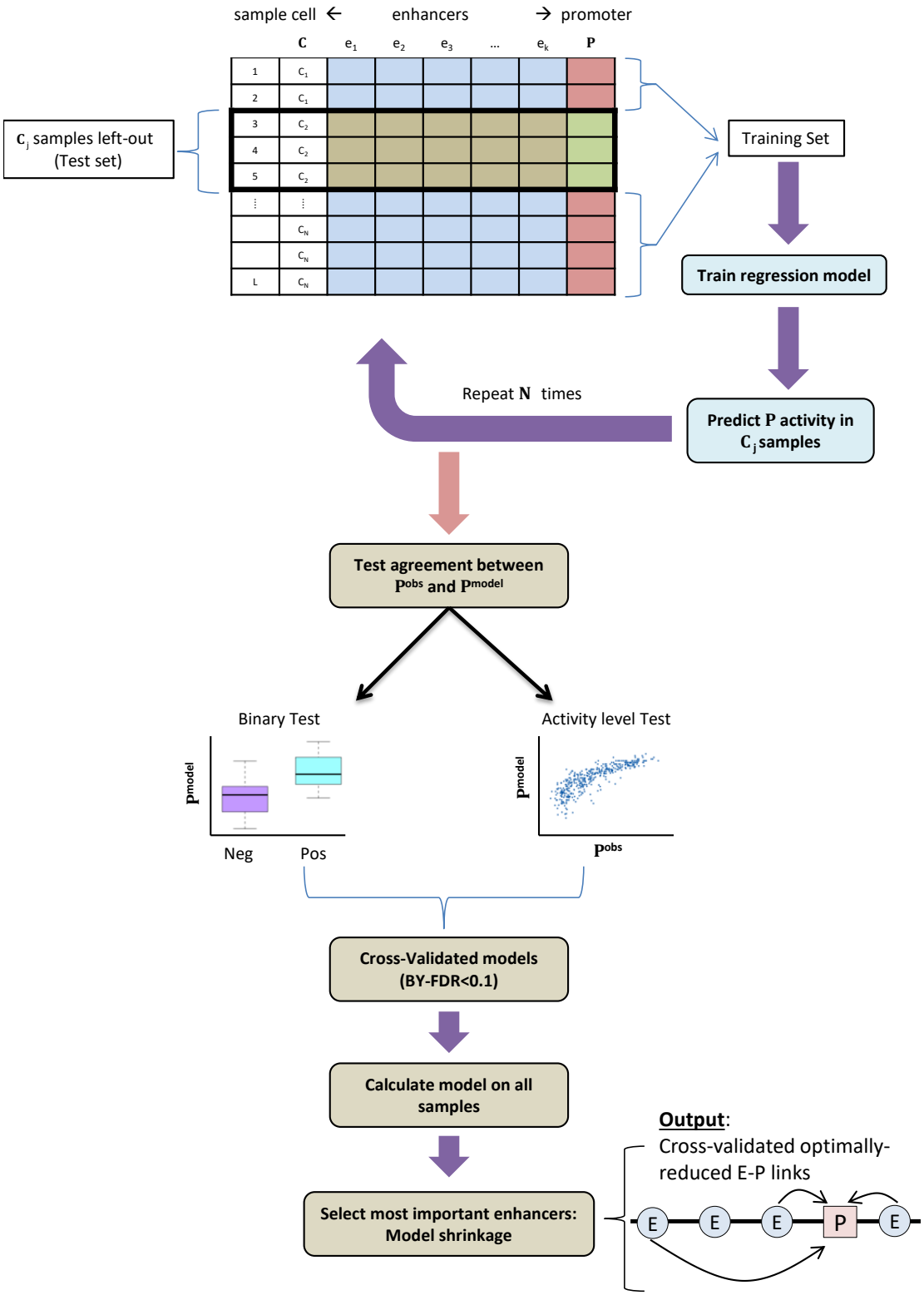


Figure 2.1. FOCS statistical procedure for inference of EP links. In a dataset with samples from N different cell types, FOCS starts by performing N cycles of leave-cell-type out cross-validation (LCTO CV). In cycle j, the set of samples from cell-type C_j is left out as a test set, and a regression model is trained, based on the remaining samples, to estimate the level of the promoter P (the independent variable) from the levels of its k closest enhancers (the dependent variables). The model is then used to predict promoter activity in the test set samples. After the N cycles, FOCS tests the agreement between the predicted (P^{model}) and observed (P^{obs})

promoter activities using two non-parametric tests. In the *binary test*, samples are divided into positive ($P^{obs} \geq 1$RPKM) and negative ($P^{obs} < 1$RPKM) sets, and the ability of the inferred models to separate between the sets is examined using Wilcoxon rank-sum test. In the *activity level test*, the consistency between predicted and observed activities in the positive set of samples is tested using Spearman correlation. P-values are corrected using the BY-FDR procedure, and promoters that passed the validation tests (FDR≤0.1) are considered validated, and full regression models, this time based on all samples, are calculated for them. In the last step, FOCS shrinks each promoter model using elastic net to select its most important enhancers.

FOCS results for ENCODE DHS epigenomic data

Applying FOCS to the ENCODE DHS dataset, we only considered promoters and enhancers that were active (that is, with signal > 1.0 RPKM) in at least 30 out of the 208 samples (This preprocessing step filtered out from the analysis 828 genes whose expression was most cell-type specific). Overall, this dataset contained 92,909 and 408,802 active promoters and enhancers, respectively (**Methods**). We first evaluated the performance of the three alternative regression methods in terms of the number of validated models each of them yielded. We found that the OLS method consistently produced more validated models that passed both the binary and activity level tests (**Fig. 2.2A-B; Supplemental Table S2.1**). Using OLS, out of the 92,909 analyzed promoters, 52,658 had models that passed both tests (q-value≤0.1), while for 7,007 promoters models passed none of these two tests (**Fig. 2.2C**). As expected, promoters with models that passed only the activity level test were active in a very high number of samples while those with models that passed only the binary test were active in much lower number of samples (**Fig. 2.2D**) (see **Supplemental Fig. S2.1** for examples of promoters in different validation categories). To examine the effect of the leave-cell-type-out cross validation (CV) procedure we compared R^2 values obtained by OLS models generated without CV to the values obtained when CV was applied (**Fig. 2.2E**). The results indicate that without CV, many models are over-fitted to the training samples and have low predictive power on new ones. This problem is more severe in other datasets that we analyzed, as shown in subsequent section. **Fig. 2.2F** shows an example of promoter model with low predictive power on new samples, and demonstrates the high sensitivity of Pearson's correlation (or equivalently, of R^2) to outliers. Such promoter models do not pass our CV tests and are considered to have low confidence.

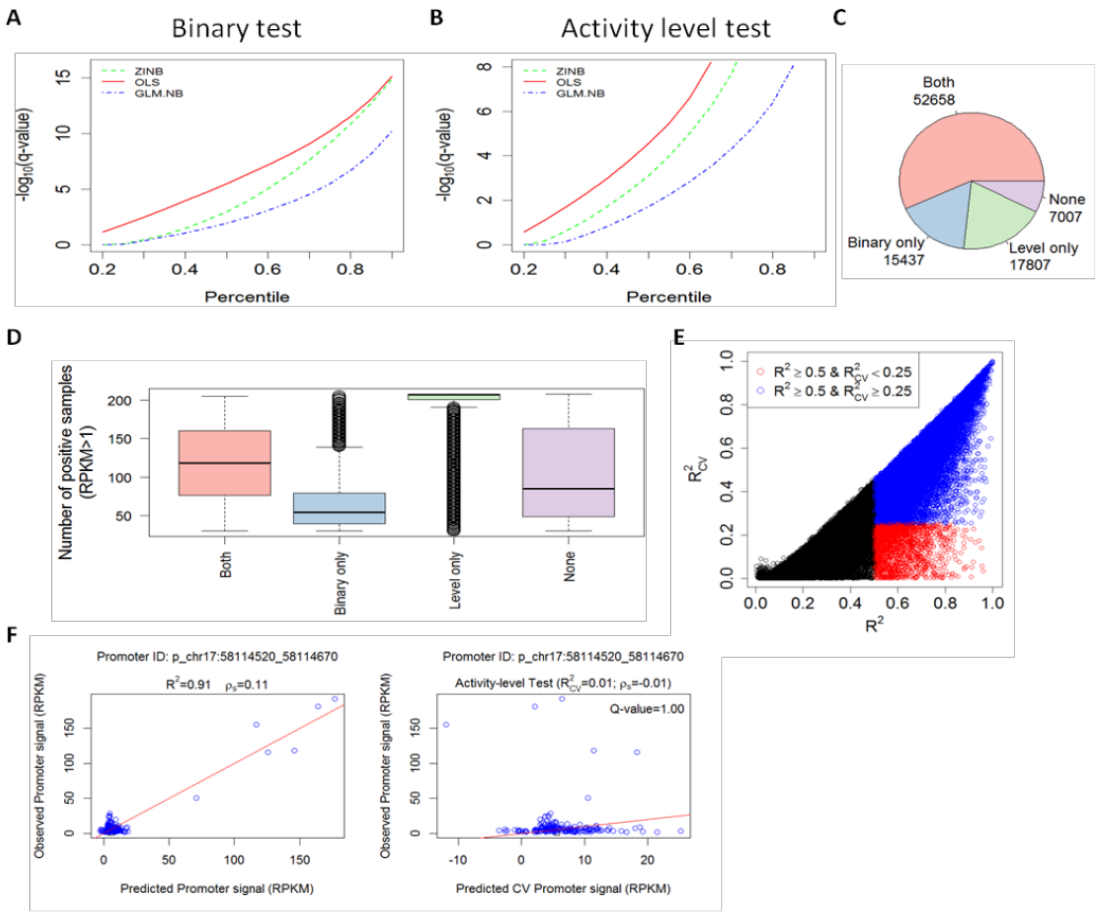


Figure 2.2. Performance of three alternative regression methods for inferring EP models. (a) Performance of ordinary least squares (OLS), generalized linear model with negative binomial distribution (*GLM.NB*) and zero-inflated negative binomial (*ZINB*) regression using the binary test. Point (x,y) on a plot indicates that a fraction x of the models had $-\log_{10}$[q-values] < y computed by Wilcoxon rank sum test. OLS yields a higher fraction of validated models at any q-value cutoff. (b) Same as (a) but using the activity level validation test, with p-values computed by the Spearman correlation test. Here too, OLS yields a higher fraction of validated models than the other methods. (c) Number of promoters whose OLS models passed (at q<0.1) each of the tests (or none). (d) The distribution of the number of positive samples (samples in which the promoter is active, i.e., has RPKM ≥ 1) for promoters in each category. (e) Comparison between the R^2 values with/without cross-validation (CV). Each dot is a promoter model. Blue dots denote models with $R^2 \geq 0.5$ and $R^2_{CV} \geq 0.25$. Red dots denote models with and $R^2 > 0.5$ and $R^2_{CV} < 0.25$ corresponding to over-fitted models with low predictive power on novel samples. (f) A promoter whose model as computed without CV gets very high R^2 (left plot) but when CV is applied a low R^2_{CV} is obtained (right plot). This example demonstrates the sensitivity of R^2 (and Pearson correlation) to outliers. ρ_S: Spearman correlation, Q-value: FDR corrected P-value.

The configuration of promoter regulation by enhancers

Next, we sought to characterize the configuration of promoter regulation by its enhancers, in terms of the number of regulating enhancers and their relative contribution. For each promoter that passed the validation tests, we now calculated a final model, this time considering all samples (**Fig. 2.1**), and estimated the relative contribution of each of its k

enhancers to this full model. As in [23], per model, we measured the proportional contribution of each enhancer by calculating the ratio r^2/R^2 where r is the pairwise Pearson correlation between the enhancer and promoter activity patterns and R^2 is the coefficient of determination of the entire promoter's model. In the analysis of the ENCODE DHS data, we included in this step the 70,465 promoters that passed the activity level test (or both tests). In agreement with previous observation [23], the closest enhancers have significantly higher contribution than the distal ones (**Fig. 2.3A**). However, the proportional contribution quickly reaches a plateau, indicating that above a certain threshold, distance to promoter is no longer an important factor, and enhancers #6-#10 (ordered according to their distance from the promoter) contribute similarly to promoter activity (**Fig. 2.3A**). Second, we examined the distribution of R^2 values of these statistically validated models. 54% of the models (37,716 out of 70,465) had $R^2 \geq 0.5$ (**Fig. 2.3B**). 61% of the 52,658 models that passed both tests had $R^2 \geq 0.5$, compared to 32% of the 17,807 models that passed only the activity level test (In contrast, only 13% of 15,437 models that passed only the binary test had $R^2 \geq 0.5$). We note that models that passed the CV tests but have low R^2 do contain confident and predictive information on EP links, though the low R^2 suggests that there are additional missing regulatory elements that play important roles in the regulation of the target promoter.

A promoter's model produced by OLS regression contains all k variables (i.e., enhancers), where each variable is assigned a significance level (p-value) reflecting its statistical strength. Next, to focus on the most informative EP interactions, FOCS seeks the strongest enhancers in each model. To this end, FOCS derives, per promoter, an optimally reduced model by applying model shrinkage (**Methods**). Lasso-based shrinkage was previously used for such task [23]. Here, we chose elastic-net (*enet*) approach, which combines Lasso and Ridge regularizations, since in cases of highly correlated variables (i.e., the enhancers), Lasso tends to select a single variable while Ridge gives them more equal coefficients (**Methods**). In this analysis too, we included the 70,465 models that passed the activity level test. **Fig. 2.3C** shows the distribution of the number of enhancers that were included in the enet-reduced models. On average, each promoter was linked to 2.4 enhancers. Inclusion frequency decreased with EP distance: the most proximal enhancer was included in 63% of the models while the 10^{th} enhancer was included in only 16% of them (**Fig. 2.3D**). Here too, the graph reaches a plateau and enhancers #6-#10 show very similar inclusion frequencies. **Supplemental Figures S2.2A-B** show the distribution of the actual EP distance for the enhancers considered by FOCS and **Supplemental Figure S2.2C** shows the inclusion frequency as a function of this distance. Regulatory elements located less than 5kb from their target promoter have markedly higher inclusion frequency. To estimate false positive rate among enhancers included in our final enet-reduced models, we randomly selected 10k promoter models from the 70,465 models that

passed the CV step, and added to each one of them an additional 11th enhancer randomly selected from a different chromosome. We then applied enet on these 10k models. Notably, the random enhancer was retained in only seven out of 10k models, which is significantly lower than inclusion frequency we observed for any EP distance bin (**Supplemental Fig. S2.2C**), indicating a low false positive rate also among the long distance EP links inferred by FOCS.

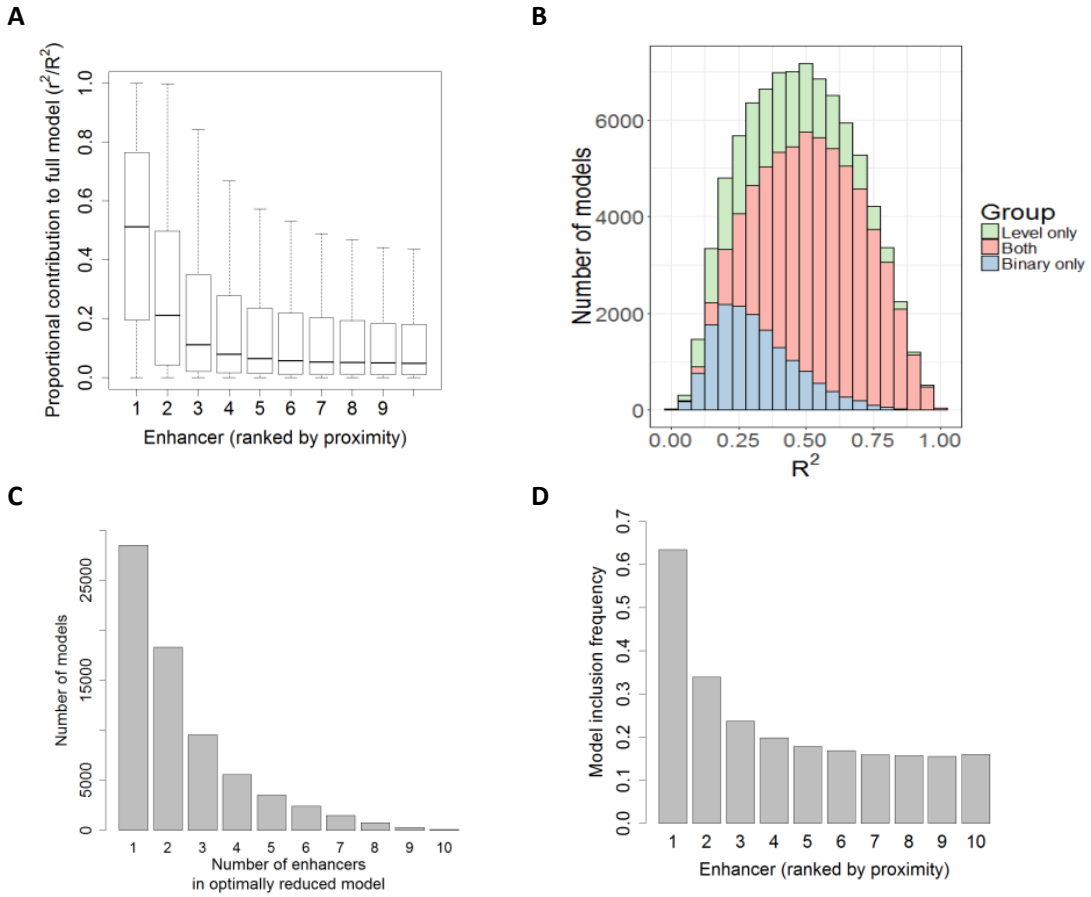


Figure 2.3. Configuration of promoter regulation by enhancers. (a) The proportional contribution of the 10 most proximal enhancers (within ±500kb of the target promoter) to models predicting promoter activity. The X axis indicates the order of the enhancers by their relative distance from the promoter, with 1 being the closest. (b) R^2 values of the models that passed one or both CV tests. (c) Distribution of the number of enhancers included in the validated, optimally reduced models (i.e. after elastic net shrinkage). Most shrunken models contain 1-3 enhancers. (d) Inclusion frequency of enhancers in the shrunken models as a function of their relative proximity to the target promoter.

Comparison of FOCS and extant methods performance using external validation resources

After optimally reducing the promoter models FOCS predicted in the ENCODE DHS dataset a total of 167,988 EP links covering 70,465 promoters and 92,603 distinct enhancers (http://acgt.cs.tau.ac.il/focs/data/encode_interactions.txt). Next, we compared the performance of FOCS and three alternative methods for EP mapping: (1) *Pairwise:* pairwise Pearson

correlation > 0.7 between EP pairs located within ±500 kbp, and accounting for multiple testing using BH (FDR $<10^{-5}$) (this was the main method used in [23], and also in [43] without multiple testing correction) (2) *OLS+LASSO*: Models are derived by OLS analysis using *all* samples without CV, selected based on $R^2 \geq 0.5$ and reduced using LASSO shrinkage (**Methods**) (this method was also applied in [23]). (3) *OLS+enet*: Same as (2) but with enet shrinkage in place of LASSO. **Table 1** summarizes the number of EP links obtained by each method. FOCS yielded ~75% more models than the other methods.

Table 1. Number of inferred promoter models obtained by four alternative methods on the ENCODE DHS dataset

Method type	#promoter models	#EP links	#Unique enhancers
Pairwise ($r \geq 0.7$)+ *FDR*	39,372	139,170	53,950
OLS-LASSO ($R^2 \geq 0.5$)*	39,368	122,064	74,104
OLS-enet ($R^2 \geq 0.5$)*	39,407	150,158	85,926
FOCS	70,465	167,988	92,603
(*) The number of OLS models ($\mathbf{R^2 \geq 0.5}$) was 39,892 before LASSO / enet shrinkage. These methods eliminate models in which no enhancer passed the shrinkage.			

To evaluate the validity of EP mappings predicted by each method, we used three external omics resources: physical EP interactions derived from RNAPII ChIA-PET data, physical EP interactions derived from YY1 HiChIP experiments, and functional EP links indicated by eQTL analysis (**Methods**). For physical EP interactions derived from RNAPII ChIA-PET we used data recorded in MCF7, HCT-116, K562 and HelaS3 cell lines (a total of 922,997 interactions). Physical EP interactions inferred from HiChIP for YY1 (recently suggested to act as a general structural regulator of EP links) were downloaded from [105] (911,190 interactions, measured in HCT-116, Jurkat and K562 cell lines). While 3C-based methods are generally not well equipped to identify DNA loops below 25Kb, we intersected our results with the best available loop calls for these data ranges. eQTL data was downloaded from the GTEx project (2,283,827 unique significant eQTL-gene pairs) [106]. We defined a 1 kbp interval for each promoter and enhancer and calculated the fraction of EP links that was supported by either ChIA-PET, HiChIP or eQTL data (**Methods**). Notably, FOCS not only yielded many more EP links (15,000-40,000 more), but also outperformed the alternative methods in terms of the fraction of predictions supported by either RNAPII ChIA-PET (**Fig. 2.4A**), YY1 HiChIP (**Fig. 2.4B**) or eQTL data (**Fig. 2.4C**). **Figure 2.5** shows two FOCS-derived promoter models that are supported by ChIA-PET and eQTLs. Note that for the promoter model of CD4 (**Fig. 2.5B**) the R^2_{CV} value was low (~0.1) while the Spearman correlation (ρ_s) was 0.53 after CV. This demonstrates that FOCS can capture promoter models that exhibit non-linear relationship between the promoter and enhancer activities.

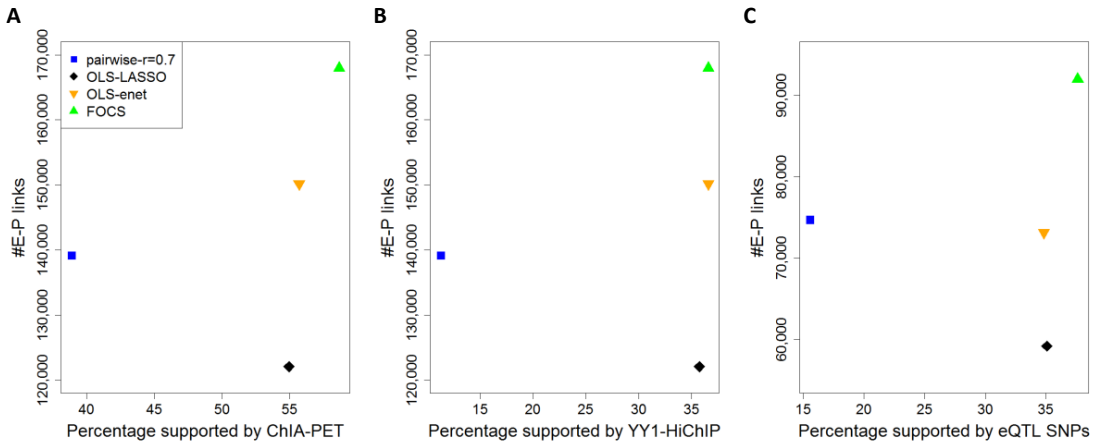


Figure 2.4. Comparison of the performance of different methods for predicting EP links using ChIA-PET and eQTL data as external validation. Y-axis shows the total number of predicted E-P links. X-axis shows the percentage supported by the external source: (A) Pol-II ChIA-PET. (B) YY1 HiChIP and (C) GTEX eQTLs. In (C) the y-axis shows the total number of predicted E-P links where the promoter is annotated with a known gene. FOCS (green triangle) makes more predictions and also manifests highest support rate by all methods: RNPII ChIA-PET (59%), YY1 HiChIP (37%) and eQTL (38%). In all methods, empirical p-value by random permutation test was < 0.01 (**Methods**).

FOCS performance on additional large-scale datasets

Having demonstrated FOCS proficiency in predicting EP links on the ENCODE DHS data, we next wished to expand the scope of our EP mapping. We therefore applied FOCS to three additional large-scale genomic datasets: (1) DHS profiles measured by the *Roadmap Epigenomics* project, consisting of 350 samples from 73 different cell types and tissues; and (2) FANTOM5 CAGE data that measured expression profiles in 1,827 samples from 600 human cell lines and primary cells. The analysis of FANTOM5 data uses eRNA and TSS expression levels for estimating the activity of enhancers and promoters, respectively (**Methods**). (3) A GRO-seq compendium that we compiled. Building on eRNAs as quantitative markers of enhancer activity and the effectiveness of the GRO-seq technique in detecting eRNA expression [21], we compiled a large compendium of eRNA and gene expression profiles from publicly available GRO-seq datasets, spanning a total of 245 samples measured on 23 different human cell lines (**Methods**).

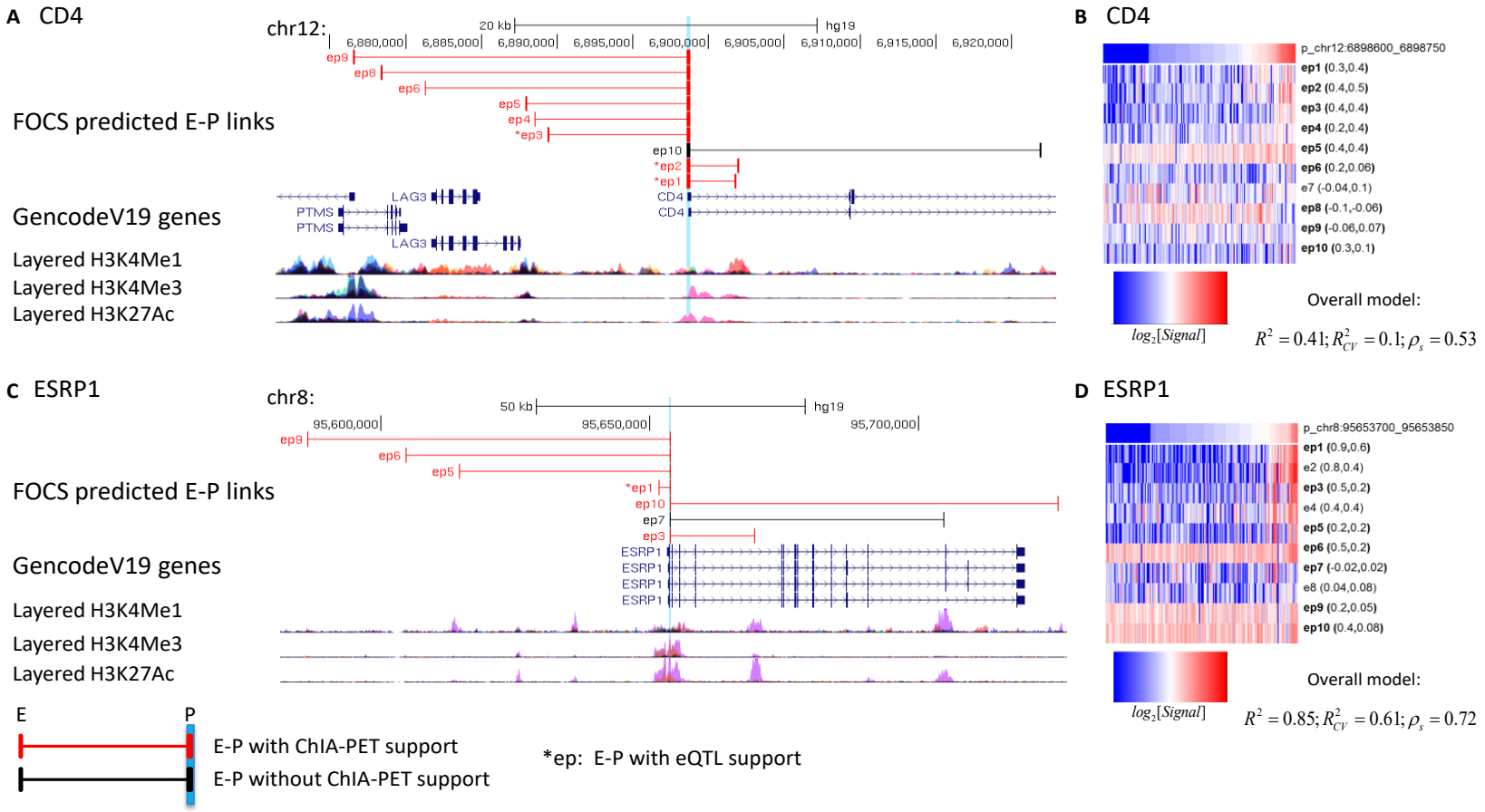


Figure 2.5. Examples of FOCS predicted EP links supported by ChIA-PET/eQTL data. (A-B) CD4. (C-D) ESRP1. TSS location is highlighted in light blue. (B,D) Heatmaps ($\log_2$[RPKM Signal]) for the activity patterns of CD4/ESRP1 promoters and their 10 nearest enhancers. Enhancers included in the shrunken model are denoted by '**ep**' and those that are not are denoted by 'e'. For each enhancer, its Pearson and Spearman correlations with the promoter are reported (left and right values in the parentheses). For each model, the R^2, R^2_{CV}, and the Spearman correlation after CV (ρ_s) are listed.

We applied to these datasets the same procedure that we applied above to the ENCODE data. In the analysis of these datasets, OLS yielded more validated models than the other regression methods on the Roadmap Epigenomics and GRO-seq datasets (as was the case in the ENCODE DHS data (**Fig. 2.2A-B**)), while GLM.NB and ZINB produced more models on FANTOM5 (**Supplemental Fig. S2.3A-C and Table S2.1**). The performance of GLM.NB and ZINB on the FANTOM5 dataset is probably due to the high fraction of zeros entries in the count matrix of this dataset (~54%) compared to ENCODE, Roadmap, and GRO-seq data matrices (8%, 4%, and 19%, respectively). As OLS performed better on most datasets, all the results reported below are based on OLS. The number of promoter models that passed each validation test in each dataset is provided in **Supplemental Fig. S2.4A-C**. The effect of CV is presented in **Supplemental Fig. S2.5A-C**. In these datasets too, many of the models with high coefficient of determination ($R^2 \geq 0.5$) when trained on all samples, had low predictive power on novel samples ($R^2_{CV} < 0.25$) (Empirical FDR 16%, 20%, and 22% in Roadmap, FANTOM5, and GRO-seq, respectively; **Supplemental Fig. S2.5**), demonstrating the utility of CV in alleviating overfitting and thus reducing false positive models.

We next examined the relative contribution of each of the 10 participating enhancers to the validated models, and in these datasets too, the most proximal enhancers had the highest role, but more distal ones had very similar contribution (**Supplemental Fig. S2.6A**). In terms of explained fraction of the observed variability in promoter activity, 41% and 84% of the models that passed both tests in the Roadmap Epigenomics and GRO-seq datasets, respectively, had $R^2 \geq 0.5$, but only 11% of the validated models reached this performance in the FANTOM5 dataset (**Supplemental Fig. S2.6B**), probably due to its exceptionally sparse data matrix. Last, FOCS applied enet model shrinkage to the models that passed the validation tests (The number of validated models and EP links derived by FOCS on each dataset is summarized in **Supplemental Table S2.2**). In the optimally-reduced models, each promoter was linked, on average, to 3.2, 2.8 and 3.6 enhancers, in the Roadmap, FANTOM5 and GRO-seq datasets, respectively (**Supplemental Fig. S2.7A**), and inclusion frequency decreased with EP distance (**Supplemental Fig. S2.7B and Fig. S2.8**). Finally, benchmarking against RNAPII ChIA-PET, YY1 HiChIP and eQTL data, for most comparisons, FOCS outperformed the alternative methods for EP mapping, by yielding many more EP predictions at similar external validation rates (**Supplemental Fig. S2.9 and Table S2.3**). Collectively, we provide a rich resource of predicted EP mapping that covers 16,349 known genes, 113,653 promoters, 181,236 enhancers, and 302,050 cross-validated EP links.

2.2. Methods

ENCODE DHS data preprocessing

DHS peak locations of enhancers and promoters were taken from a master list of 2,890,742 unique, non-overlapping DHS segments [43]: ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/combined_peaks/multi-tissue.master.ntypes.simple.hg19.bed

We extracted from the master list the set of known (n=68,762) and novel (n=44,853) promoter-DHS peaks taken from: ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/promoter_predictions

The remaining (n=2,777,127) non-promoter-DHS peaks in the master list were considered as putative regulatory elements, collectively referred here as enhancer elements. To create enhancer/promoter signal matrices, we used the BAM files of 208 UW DNase-seq samples (106 cell types) from GSE29692 GEO dataset [43,107,108]. The number of reads mapped within each DHS peak was counted using BEDTools utilities [109]. To reduce our FOCS running time we focused only on promoters/enhancers with signal ≥1RPKM in at least 30 samples, resulting in 92,909 promoters and 408,802 putative enhancers.

We defined for each promoter the set of k=10 candidate enhancers located within a window of 1Mb (±500Kb upstream/downstream from the promoter's center position). We mapped promoters to annotated genes using GencodeV10 TSS annotations (ftp://genome.crg.es/pub/Encode/data_analysis/TSS/Gencodev10_TSS_May2012.gff.gz). 54,650 promoters (out of 92,909) were linked to annotated TSSs.

Roadmap epigenomic DHS data preprocessing

DHS peak positions for 474,004 putative enhancer and 33,086 promoter non-overlapping DHS segments [110] were taken from:

- https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-intersect_release/DNase/p10/prom/25/state_calls.RData
- https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-intersect_release/DNase/p10/enh/25/state_calls.RData

To create enhancer/promoter signal matrices, we used the aligned reads (BED files) of 350 UW DNase-seq samples (73 cell types) from GSE18927 GEO dataset [107,108,111–113]. The

number of reads mapped within each DHS peak was counted using the BEDTools utilities [109]. We focused only on promoters/enhancers with signal ≥1RPKM in at least one sample, resulting in 32,629 promoters and 470,549 putative enhancers.

We defined for each promoter the set of k=10 candidate enhancers located within a window of ±500Kb. We mapped promoters to annotated genes using GencodeV10 TSS annotations (ftp://genome.crg.es/pub/Encode/data_analysis/TSS/Gencodev10_TSS_May2012.gff.gz) [114]. 17,941 (out of 32,629) promoters were linked to annotated TSSs.

FANTOM5 data preprocessing

Promoter (CAGE tags peak phase 1 and 2) and enhancer (human permissive enhancers phase 1 and 2; n=65,423) expression matrices (counts and normalized) covering 1,827 samples (600 cell types) were downloaded from FANTOM5 DB (http://fantom.gsc.riken.jp/). As in FANTOM5 paper [23] we focused on promoters with expression ≥1 TPM (Tags Per Million) in at least one sample, resulting in 56,290 promoters annotated with 26,489 RefSeq TSSs within ±500 bp. We defined for each promoter the set of k=10 candidate enhancers located within a window of ±250Kb from the promoter's TSS. The choice of smaller window here was done for consistency with the FANTOM5 choices.

GRO-seq data preprocessing

We downloaded raw sequence data of 245 GRO-seq samples from the Gene Expression Omnibus (GEO) database (**Additional file 3: Table S5**). See **Supplemental Methods** for further processing details. We defined for each gene the set of k=10 candidate enhancers located within a window of ±500Kb from its TSS.

FOCS Model Implementation

The input to FOCS is two activity matrices, one for enhancers (M_e) and the other for promoters (M_p), measured across the same samples. Activity is measured by DHS signal in ENCODE and Roadmap data, and by expression level in FANTOM5 and GRO-seq data. Samples were labeled with a cell-type label out of C cell-types. The output of FOCS is predicted E-P links.

First, FOCS builds for each promoter an OLS regression model based on the k enhancers whose center positions are closest to the promoter's center position (in ENCODE, Roadmap, and FANTOM5) or TSS (in GRO-seq). Formally, let y_p be the promoter p normalized activity pattern (measured in CPM - counts per million; y_p is a row from M_p) and let X_p be the normalized activity matrix of the corresponding k enhancers (CPM; k rows from

M_e). We build an OLS linear regression model $y_p = X_p\beta_p + \varepsilon_p$, where ε_p is a vector that denotes the errors of the model and β_p is the $(k+1)\ x\ 1$ vector of coefficients (including the intercept) to be estimated.

Second, FOCS performs leave-cell-type-out cross validation (LCTO CV) by training the promoter model based on samples from $C-1$ cell types and testing the predicted promoter activity of the samples from the left out cell type. This step is repeated C times. The result is a vector of predicted activity values y_p^{model} for all samples.

FOCS tests the predicted activity values using two validation tests: (1) The *binary test*. This test examines whether y_p^{model} discriminates between the samples in which p was active (observed activity $y_p \geq 1$ RPKM) and the samples in which p was inactive ($y_p < 1$ RPKM). (2) The *activity level test*. This test calculates, for the active samples, the significance of the Spearman correlation between y_p^{model} and y_p. Spearman correlation compares the ranks of the original and predicted activities. We obtain two vectors of p-values, one for each test, of length n (the number of promoter models).

Third, to correct for multiple testing, FOCS applies on each p-value vector the Benjamini - Yekutieli (BY) FDR procedure [104]. Promoter models with q-value≤ 0.1 in either both tests or in the activity level test were included in further analyses. In GRO-seq analysis, we also included models that passed only the binary test (m=2,580) since 57% of them had $R^2 \geq 0.5$ (**Supplemental Fig. S2.6B**). For promoters that passed these CV tests final models are trained again using all samples.

FOCS next selects informative enhancers for each final promoter model. The enhancer selection step is described in the **Supplemental Methods**.

Alternative regression methods

We compared the performance of OLS method with GLM.NB and ZINB regression methods. We repeated the FOCS steps but in the first step, instead of OLS we applied the GLM.NB or the ZINB methods (see **Supplemental Methods** for details).

FANTOM5 E-P linking using OLS regression was followed by Lasso shrinkage (defined as OLS-LASSO) as described in [23] (see **Supplemental Methods** for details).

GO enrichment analysis

GO enrichments were calculated using topGO R package [115] (algorithm="classic", statistic="fisher", minimum GO set size=10). We split the genes into target and background sets using their enhancer bin sets. Genes belonging to bins with 1-3/1-4/4-10/5-10 enhancers were considered as target set and compared to all genes from all bins as background set. Correction for multiple testing was performed using BH procedure [67].

External validation of predicted EP links

We used three external data resources for validating FOCS E-P link predictions: (1) RNAPII ChIA–PET interactions, (2) YY1-HiChIP interactions, and (3) eQTL SNPs.

We downloaded 922,997 ChIA-PET interactions (assayed with RNAPII, on four cell lines: MCF7, HCT-116, K562 and HelaS3) from the chromatin–chromatin spatial interaction (CCSI) database [116] (GEO accession numbers of the original ChIA-PET samples are provided in **Additional file 3: Table S6**). We used the liftOver tool (from Kent utils package provided by UCSC) to transform the genomic coordinates of the interactions from hg38 to hg19. HiChIP interactions mediated by YY1 TF (cell types: HCT116, Jurkat, and K562) were taken from [105] (GEO accession id: GSE99521). As done in [105], we retained 911,190 YY1-HiChIP high confident interactions (Origami probability>0.9). For eQTL SNPs, we used the significant SNP-gene pairs from GTEx analysis V6 and V6p builds. 2,283,827 unique eQTL SNPs covering 44 different tissues were downloaded from GTEx portal [106].

We used 1Kbp intervals (±500 bp upstream/downstream) for the promoters (relative to the center position in ENCODE/Roadmap/FNATOM5 or to the TSS position in GRO-seq) and the enhancers (±500 bp from the enhancer center). An E-P pair is considered supported by a particular capture interaction if both the promoter and enhancer intervals overlap different anchors of an interaction. An E-P pair is considered supported by eQTL SNP if the SNP is located within the enhancer's interval and is associated with the expression of the promoter's gene. For each predicted E-P pair we checked if the promoter and enhancer intervals are supported by capture interactions and eQTL data. We then measured the fraction of E-P pairs supported by these data resources. See **Supplemental Methods** for the significance calculation of the empirical P-value.

Statistical tests, visualization and tools used

All computational analyses and visualizations were done in the R statistical language environment [117]. We used the two-sided Wilcoxon rank-sum test implemented in wilcox.test() function to compute the significance of the binary test. We used the cor.test() function to compute the significance of the Spearman correlation in the activity level test. Spearman/Pearson correlations were computed using the cor() function. To correct for multiple testing we used the p.adjust() function (method='BY'). We used 'GenomicRanges' package [118] for finding overlaps between genomic positions. We used 'rtracklayer' [119] and 'GenomicInteractions' [120] packages to import/export genomic positions. Counting reads in genomic positions was calculated using BEDTools [109]. OLS models were created using lm() function in 'stat' package [117]. GLM.NB models were created using glm.nb() function in 'MASS' package [121]. ZINB models were created using zeroinfl() function in 'pscl' package [122]. Graphs were made using graphics[117], ggplot2 [123], gplots [124], and the UCSC genome browser (https://genome.ucsc.edu/).

3. Predicting cell-type specific enhancer-promoter maps

The FOCS algorithm described in the previous chapter predicts global EP links based on correlated activity patterns across many samples covering hundreds of different cell types. However, the predicted EP links are global and may not reflect links that are specific to a few cell types. A key challenge is to identify which of these predicted EP links are actually functional and in which specific cell types. To this end, in the study described in this chapter, we developed CT-FOCS (cell-type-FOCS), a linear mixed effect model (LMM) that estimates the cell type activity of an EP link based on multiple samples available for each cell type. We applied CT-FOCS on the FANTOM5 CAGE, Roadmap Epigenomics and ENCODE DHS datasets, and predicted a total of 229,518 cell type-specific EP links (termed as ct-links) across 651 cell types.

We compared CT-FOCS with extant methods in terms of concordance with experimentally derived chromatin interactions (from 3C-based genomic assays) and cell type gene expression specificity of linked genes. The direct way to validate predicted ct-links against experimental loops is to check whether the enhancer and promoter overlap the two anchors of the same loop. However, chromatin can be organized in intricate nested structures, reflected by overlapping anchors of different chromatin loops that should be considered when using 3D data for validation of predicted EP interactions. Moreover, predicted ct-links covering a linear distance of less than 20 kb, an area where ChIA-PET loops tend not to perform well as shown by previous studies [125], might not receive direct validation from that test. To this end, we devised a "two-step connected loop set" (TLSs) approach to broaden the set of anchors that are considered proximal for validating ct-links. A limitation of this approach is that it assumes a form of transitivity, which does not necessarily have to hold in a cell population. We show that transitivity is not common on real data.

Lastly, we asked whether predicted ct-links drive cell type-specific gene expression. To this end, we measured the specificity of 402 known TFs within the enhancers and promoters of the inferred links. We show that ct-links predicted by CT-FOCS drive highly cell type-specific TFs and are superior to extant methods, thus demonstrating CT-FOCS's capability to infer biologically relevant cell type-specific gene regulation.

The study was published in *Nucleic Acids Research* in 2022 [50]. Some large supplementary files are available from the journal's website through the links provided in the thesis.

3.1. Results

The CT-FOCS algorithm

We developed a novel method called CT-FOCS (*C*ell *T*ype FOCS) for inferring cell type specific EP links (ct-links). The method utilizes a single type of omics data (e.g., CAGE or DHS) from large-scale datasets.

The input to CT-FOCS is enhancer and promoter activity profiles for a set of cell types. The output is the set of ct-links called for each cell type. Note that the enhancers or promoters involved in ct-links can be broadly active separately. In contrast to methods that seek global correlations between the activity profiles of enhancers and promoters, the aspect emphasized and detected by CT-FOCS is the specificity of the link between the two elements: that is, links reported by CT-FOCS highlight the few cell types in which the enhancer and promoter are predicted to functionally interact.

CT-FOCS builds on FOCS [101], which discovers global EP links showing correlated enhancer and promoter activity patterns across many samples. FOCS performs linear regression on the levels of the 10 enhancers that are closest to the target promoter, followed by two non-parametric statistical tests for producing initial promoter models, and regularization to retrieve the most informative enhancers per promoter model. CT-FOCS starts with the full (non-regularized) FOCS promoter model (**Methods)**, and uses a linear mixed effect model (LMM), utilizing groups of replicates available for each cell type to adjust a distinct regression curve per cell-type group in one promoter model (**Figure 3.1; Methods**). We call a ct-link in a certain cell type if it meets the following criteria: (1) both the enhancer (E) and the promoter (P) show markedly positive activity levels in that cell type compared to other cell types, and (2) both P and E have significantly high random effect coefficients, reflecting an advantage of the LMM over the global FOCS model (**Methods**). The second criterion increases our confidence that the high activity detected by the first is specific to this cell type.

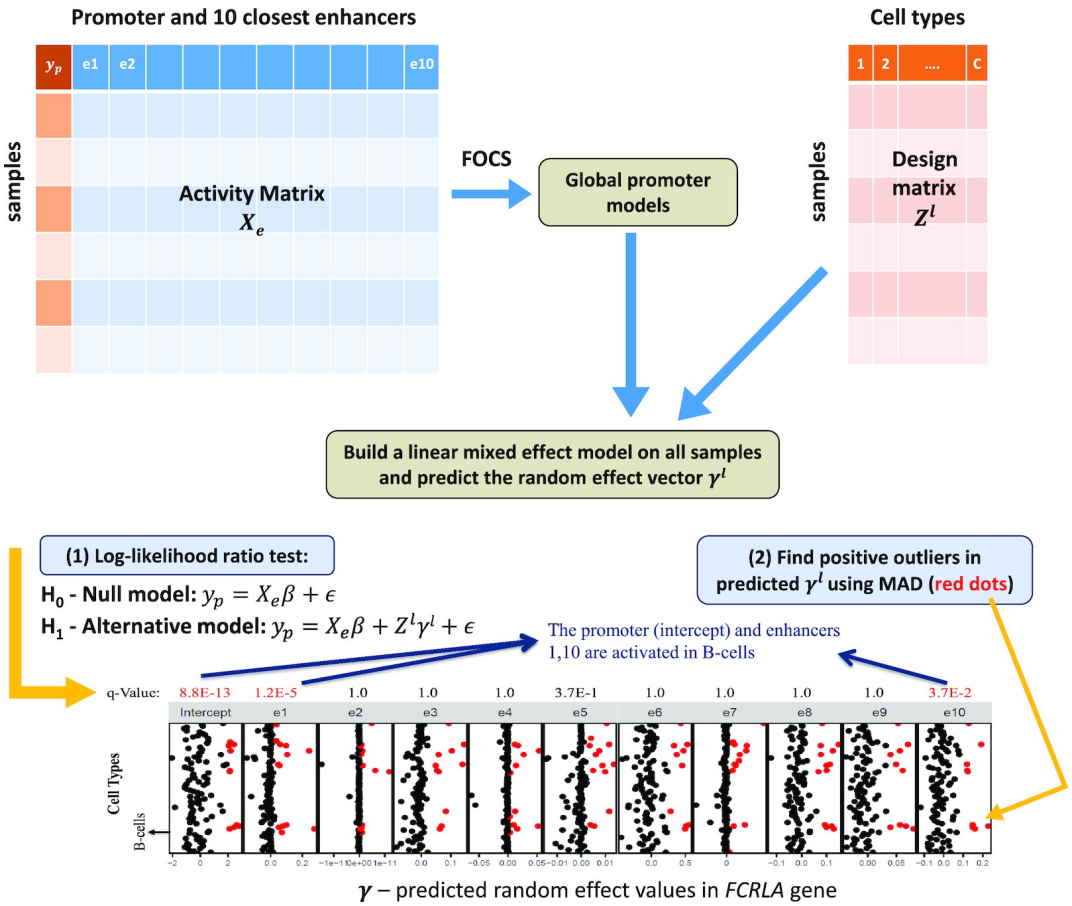


Figure 3.1. Outline of the CT-FOCS algorithm. Let y_p denote the observed activity of promoter p, and X_e be the activity matrix of the $k = 10$ closest enhancers to p. If $l \in \{1, \ldots, k+1\}$ is one of the variables (enhancer or promoter, i.e. the intercept), then $Z^l[i,j]$ equals to $X_e[i,l]$ if sample i belongs to cell type j and 0 otherwise (see **Methods**). First, a robust global promoter model is inferred by applying the leave-cell-type-out cross validation step in FOCS (see Hait et al. 2018 for details). Second, a linear mixed effects model (LMM) is built on all samples using y_p, X_e, and Z^l. The LMM includes the component $Z^l\gamma^l$ where γ^l is a vector of the predicted random effect values for each variable (i.e., enhancer or promoter) per cell type. Then, the algorithm performs two tests for every l: (1) log-likelihood ratio test (LRT) to compare the simple linear regression and the LMM model. The test is carried out eleven times (testing the 10 enhancers and the intercept). The p-values for these LRTs are adjusted for multiple testing (q-values). (2) The γ^l values produced by the LMM are standardized using the Median Absolute Deviation (MAD) technique and positive outliers (red dots) are identified. A cell type-specific EP link (ct-link) is called if: (1) both enhancer and promoter (i.e., the intercept) have q-value <0.1 (marked in red), and (2) the enhancer and the promoter are found as positive outliers in the same cell type. In the *FCRLA* gene given as an example, the promoter p and enhancers e_1, e_{10} are significant and are commonly found as positive outliers in B-cells. Therefore, E_1p and $E_{10}p$ are called by CT-FOCS as B-cell-specific EP links.

To demonstrate the difference between the linear and LMM predictions, **Supplementary Figure S3.1** shows, for the same promoter (P), two links involving distinct enhancers (E1 and E2), one predicted by CT-FOCS (E1P) and the other (E2P) by FOCS. The link between E1 and P is active only in neurons, while the link between E2 and P is active over a wider range of cell types of distinct lineages (amniotic membrane cells, whole blood cells, fibroblasts, endothelial cells and preadipocytes).

Note that choosing links by setting a threshold only on the $logEP$ value would produce many false-positive calls, as the signals in promoters tend to be higher than those in enhancers [23] (see the examples in **Supplemental Fig S3.1A** and **Supplemental Fig S3.1B**).

We applied CT-FOCS on FANTOM5 cap analysis of gene expression (CAGE) profiles, which include 808 samples from 225 cell lines, 157 primary cells, and 90 tissues [23] (**Methods**). CAGE quantifies the activity of both enhancers and promoters, and overall this dataset covers 42,656 enhancers and 24,048 promoters (mapped to 20,597 Ensembl protein-coding genes). For some analyses, we also applied CT-FOCS to ENCODE's DNase Hypersensitive Site (DHS) profiles [43,126], which cover 106 cell types, each with typically 2 replicates. This dataset includes measurements for 36,056 promoters (mapped to 13,464 Ensembl protein-coding genes) and 658,231 putative enhancers (**Methods**). Unlike the FANTOM5 dataset, which builds on the expression of enhancer-RNAs (eRNAs) as a robust readout for enhancer activity, open genomic regions identified by DHS do not necessarily mark functionally active enhancers and promoters. Thus, EP maps inferred using the ENCODE dataset may be less reliable, and we focus our analyses mainly on the FANTOM5 dataset.

Overall, CT-FOCS identified 195,232 ct-links in FANTOM5 dataset **(Table 3.1)**, with an average of 414 ct-links per cell type (median 594, **Table 3.1; Supplementary Figure S3.2A**). These results are in line with the low number of ct-links observed experimentally by the abovementioned studies, including for NPC and neurons [127,128], and further indicate that the EP links specific to a cell type constitute only a small portion of the EP links that are active in it. The EP links called by CT-FOCS were on average shared across 2.5 cell types (**Supplementary Figure S3.2B**). CT-FOCS predicted both proximal and distal interactions, with an average EP distance of ~160kb (median ~110kb; **Supplementary Figure S3.2C**). The complete set of predicted ct-links for each cell type is available at http://acgt.cs.tau.ac.il/ct-focs.

Since EP links are expected to function mostly within topologically associated domains (TADs) [129,130], we next tested if ct-links detected by CT-FOCS are enriched for intra-TAD genomic intervals. As TADS are largely cell-type invariant [131], we used for these tests the 9,274 TADs reported by Rao et al. in GM12878 [131]. Indeed, comparison with randomly matched EP links demonstrated that predicted ct-links tend to lie within TADs (**Supplementary Figure S3.3**).

Inferred ct-links correlate with cell type-specific gene expression

To evaluate the specificity of the CT-FOCS predictions, we compared the activity of the set of ct-links inferred for a particular cell type with their activity in all other cell types. We defined the activity of an EP link in a cell type as the logarithm of the product of the enhancer and promoter activities in that cell type. We used these measures to compute the cell-type specificity for the set of ct-links detected in each cell type, using a score akin to [132] (**Methods**). As an example, CT-FOCS called 340 ct-links on the GM12878 lymphoblastoid cell line. We scored the cell-type specificity of these 340 ct-links for each cell type. Reassuringly, GM12878 was the top scoring cell type, and other high scoring cell types were enriched for related lymphocyte cells (other B-cells and T-cells; **Figure 3.2A, C**). GM12878 was also ranked first in cell type-specificity scores calculated separately for the promoters and enhancers of these 340 ct-links (**Supplementary Figure S3.4**).

Next, we examined how the effect of ct-links is reflected by cell-type specific expression of the linked genes (**Methods**). The 340 ct-links called by CT-FOCS in GM12878 involve 197 genes. We examined their expression profiles over 112 cell types using an independent gene expression (GE) dataset [44]. In this analysis, we now scored each of the 112 cell types for the specificity in the expression of these 197 genes. Notably, here too, the lymphocyte group (B- and T-cells) showed the highest expression levels (**Figure 3.2B**) with GM12878 ranking first by GE specificity (**Figure 3.2D**). Overall, these results show that for GM12878, the ct-links predicted by CT-FOCS based on CAGE data are correlated with lymphocyte-specific GE programs. **Supplementary Figure S3.5** shows similar results for neurons cells.

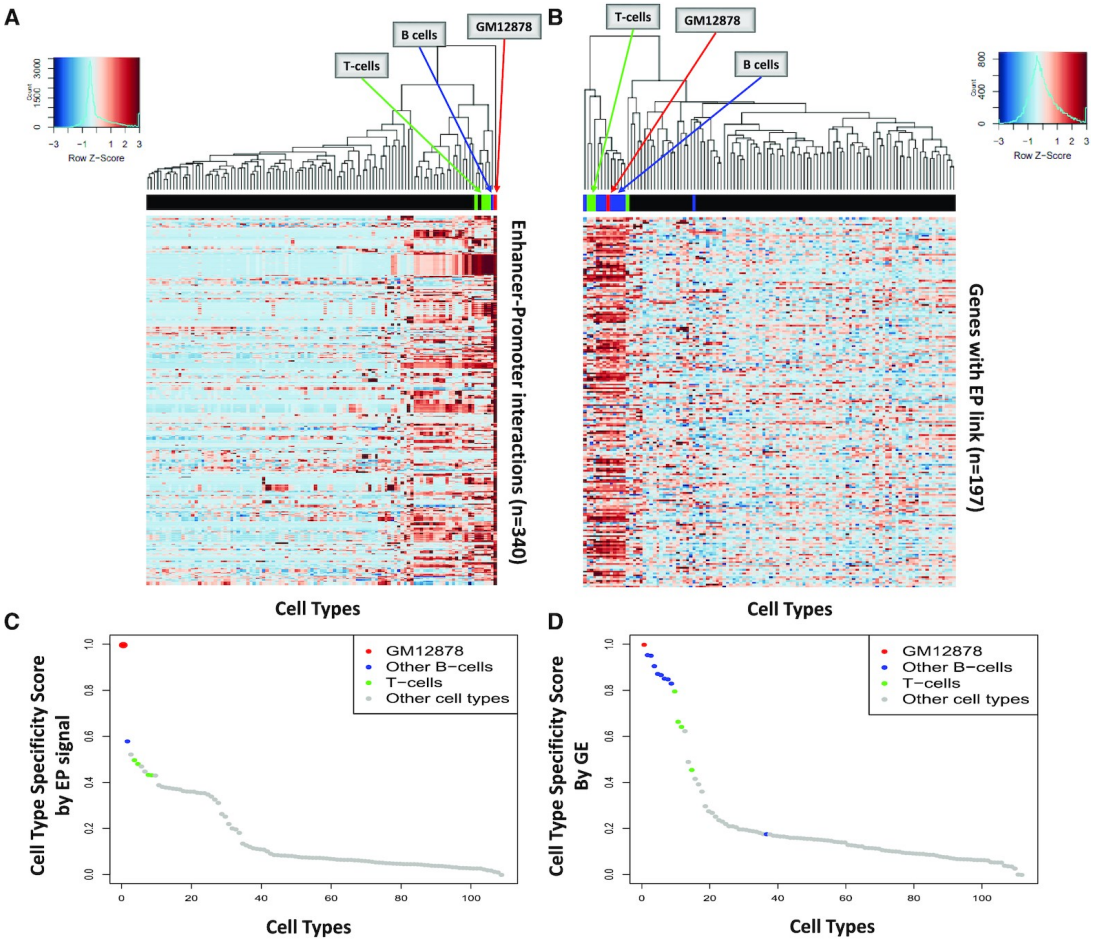


Figure 3.2. Specificity of ct-links predicted for GM12878 cell line. (A) Heatmap of EP signals for 340 ct-links predicted on GM12878 cells. Rows – EP links, columns – cell types, color – z-score of EP signal. Cell types related to lymphocytes (B/T-cells) are highlighted in color. **(B)** Heatmap of gene expression (GE) for 197 genes involved in the predicted ct-links. Rows – genes, columns – cell types, color – z-score of GE. **(C)** Cell type specificity scores based on the EP signals. **(D)** Cell type specificity scores based on expression for the gene set in B (Methods). In A and C, 109 cell types with at least 3 replicates are included in the analysis; in B and D, 112 cell types with ENCODE GE data are included [44].

Comparison of CT-FOCS to other methods

We compared CT-FOCS predictions on the FANTOM5 dataset with those made by four alternative methods: (1) JEME [133], which predicts EP links that are active in a particular cell type but are not necessarily cell type-specific. (2) A naive variant of FOCS, which takes the shrunken promoter models from FOCS, and predicts ct-links by detecting cell types in which the promoter and any of the model's enhancers show exceptionally high activity, based on the median absolute deviation (MAD) index. We call this variant MAD-FOCS (**Methods**). (3-4) To overcome large differences among methods in the numbers of predicted links, we created

subsets of the solutions of JEME and MAD-FOCS by filtering of their reported links to produce sets of links of the same size as the ones detected by CT-FOCS (**Methods**). We call these subsets cell-type-JEME (CT-JEME) and cell-type-MAD-FOCS (CT-MAD-FOCS), respectively.

Supplementary Figure S3.2 shows basic properties of the solutions provided by the five methods. EP links predicted by JEME and MAD-FOCS were, on average, shared across 11 and 12 cell types (median=3 and 13 respectively; **Supplementary Figure S3.2B**). In contrast, the CT-FOCS, CT-MAD-FOCS and CT-JEME EP links were, on average, shared across <4 cell types (Median=2, 2 and 1, respectively), demonstrating that they identified EP links that are more specific. The same number of predicted links allows fair comparison between CT-FOCS, CT-MAD-FOCS and CT-JEME.

Next, we calculated cell-type specificity scores for the EP links called by CT-FOCS, CT-MAD-FOCS and CT-JEME on the 276 FANTOM5 cell types. For each cell type, we used the ct-links called on it to calculate its specificity score on all cell types, and ranked the cell types by their scores. We expect the given cell type to score the top. In this analysis, CT-MAD-FOCS and CT-FOCS performed similarly, and significantly better than CT-JEME (**Supplementary Figure S3.6A**). In terms of GE of the genes associated with the EP links, examining the four cell types (GM12878, K562, HepG2 and MCF-7) that were present in both FANTOM5 and the independent GE dataset of Sheffield et al. [43], CT-FOCS was the only method that ranked 1st all the four cell types (**Supplementary Figure S3.6B**). Overall, these three methods seem to capture ct-links with highly specific EP and GE signals.

Next we ranked the cell types according to cell-type specificity scores obtained when considering separately the signals of the linked enhancers and promoters. Using ct-link enhancers signals, the median rank of the 'root' cell type (the cell type in which the link was found) was 1st by all methods, possibly because enhancers tend to be cell type specific. However, when using ct-link promoter signals, the median rank of the root cell type obtained by CT-JEME was only 23rd, while reassuringly, it was 1st for CT-FOCS and CT-MAD-FOCS. The low ranks of CT-JEME's linked promoters can explain why its predicted ct-links ranked lower compared to CT-FOCS and CT-MAD-FOCS.

Last, we compared the CT-FOCS predictions on ENCODE's DHS dataset with those obtained by six other methods: (1-2) CT-MAD-FOCS and MAD-FOCS; (3) TargetFinder [134], which predicts EP links based on features in enhancer, promoter and the window between them using GradientBoosting trees; (4) ABC score model [135,136], which inferred cell type-specific

functional EP links in 131 human biosamples; and (5-6) Subsets of TargetFinder and ABC model solutions having, for each cell type, a similar number of predictions as CT-FOCS (**Methods**). We call these subsets CT-TargetFinder and CT-ABC, respectively. Note that while our evaluation of the different methods using the FANTOM5 data was done on 276 cell types (that had at least 50 predicted EP links in all methods), the evaluation using the ENCODE dataset is done only on 5-10 cell types (see **Methods**). Overall, considering the specificity scores of the ct-links calculated based on DHS signals, CT-FOCS, CT-MAD-FOCS and ABC ranked the root cell type first for most cell types, better than the other three methods. On the basis of GE specificity, CT-FOCS, ABC and CT-ABC ranked the root cell type first for most cell types, performing better than the other three methods (**Supplementary Table 1B**).

Introducing 'two-step connected loop sets' in 3C assays to improve the evaluation of ct-links

We validated the ct-links predicted on GM12878 using empirical loops that were detected in this cell type by both *POLR2A* ChIA-PET and promoter-capture (PC) Hi-C [137,138]. The direct way to validate a predicted ct-link is to check whether the E and P regions overlap the two anchors of the same loop. However, as loops indicate 3D proximity of their anchors, overlapping anchors of different loops indicate proximity of their other anchors as well [139,140]. Furthermore, predicted ct-links that span a linear distance of < 20kb, a range where ChIA-PET loops perform poorly [141], may not be directly supported by that assay. Thus, for the validation of ct-links, we broadened the set of anchors that are considered to be proximal as follows: We define the 'two-step connected loop set' (**TLS**) of a loop as the set of anchors of all loops that overlap with at least one of its anchors (**Figure 3.3A**). We consider a predicted ct-link as validated if its enhancer and promoter regions overlap different anchors from the same TLS (**Figure 3.3B;** see **Supplementary Figure S3.7** for an additional example**; Methods**). To increase our confidence that TLSs indeed represent genuine chromatin interactions, we checked for each TLS if there is a loop from the same assay that is not part of the TLS but has both anchors overlapping TLS anchors (for example, in **Figure 3.3A** - loop E and the TLS of loop y). In the *POLR2A* ChIA-PET (from GM12878) and YY1 HiChIP (from K562), 54% and 64% of the TLSs were supported by such loops, respectively.

Out of the 340 ct-links inferred by CT-FOCS in GM12878, 10% were supported by ChIA-PET single loops, and 33% were supported by TLSs. Using loops from PCHi-C in GM12878, validation rates were 7.6% and 15%, respectively (Although these rates might seem low, in the next section we show that most methods predicting EP links have a low support from 3D conformation data). To test the significance of the observed validation rate, we generated

random sets of 340 intra-TAD links having the same linear distances between E and P regions as the ct-links predicted by CT-FOCS (**Methods**). In 1,000 random sets, TLSs supported, on average, 9.4% (32 out of 340) and at most 14% (46 out of 340) (**Supplementary Figure S3.8A**), and the number of predicted ct-links supported by ChIA-PET data was significant with P<0.001. Similar significance was achieved when validating the predicted ct-links directly against single loops (**Supplementary Figure S3.8C**). The same tests for PCHi-C loops gave an average overlap of matched random loops with PCHi-C TLSs of 8.5% (29 out of 340) and at most 12.4% (42 out of 340), with P=0.003 for TLS (**Supplementary Figure S3.8B**) and P=0.048 for single loops; **Supplementary Figure S3.8D**).

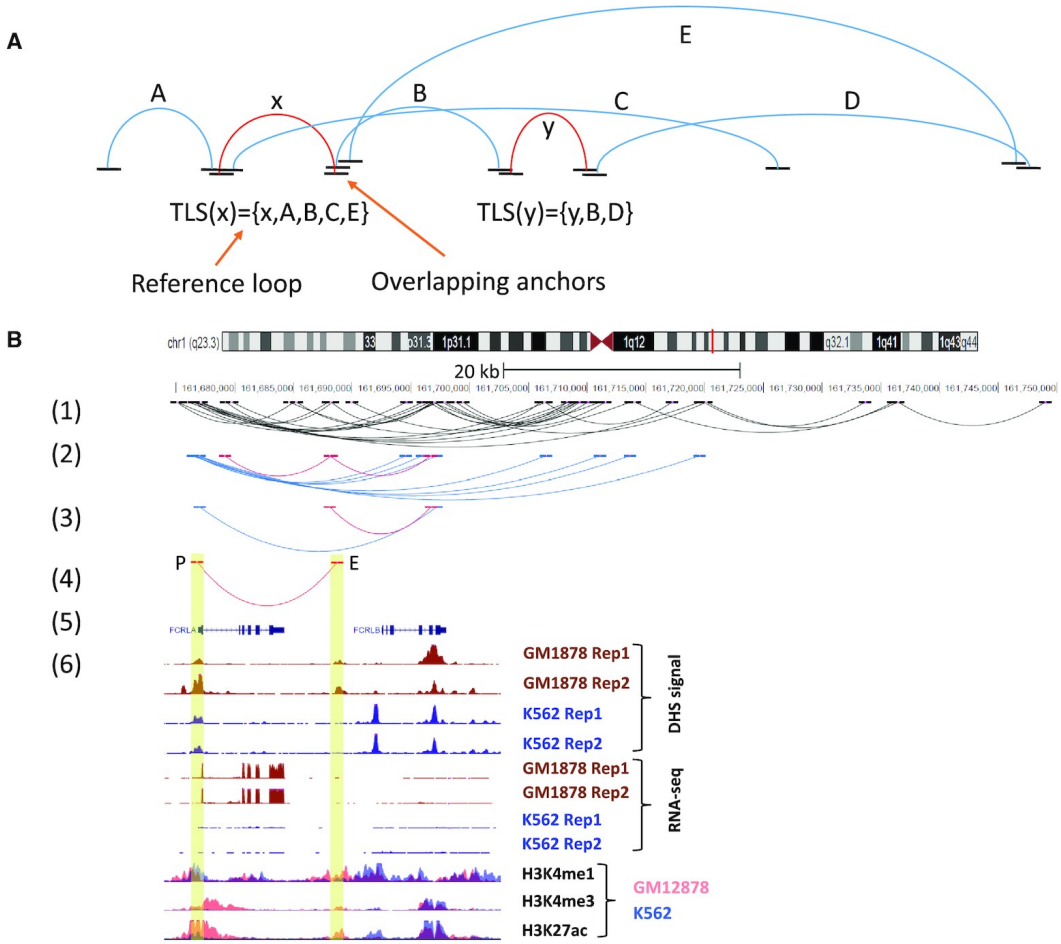


Figure 3.3. ChIA-PET TLSs support predicted ct-links. The two-step connected loop set (TLS) of a reference loop x is defined as the set of all loops that have an anchor overlapping one of x's anchors including loop x. (A) Examples of TLSs. Loop x's anchors overlap with at least one of the anchors of loops A, B, C, and E, and, therefore, the TLS of x is composed of loops x, A, B, C, E. Similarly, the TLS of y is composed of loops B, y, and D. Loop E overlaps anchors of both B and D but is not part of TLS(y) as it does not overlap y's anchors. (B) (1) A 70kb region of Chromosome 1 showing ChIA-PET loops detected in GM12878. (4) A ct-link predicted by CT-FOCS. (2) The same region showing only loops that have anchors overlapping the anchors of the ct-link. Pink: loops overlapping the enhancer; blue: loops overlapping the promoter. (3) A TLS that supports the predicted ct-link. The ct-link in

(4) is validated by the TLS, but not by any single ChIA-PET loop. (5) Gene annotations. (6) Gene expression (RNA-seq) and epigenetics signals (DHS-seq and selected histone modifications) for the region. Tracks are shown using UCSC Genome Browser for data from GM12878 and K562 cell lines. The data indicates that this link is active in GM12878 but not in K562.

Validating predicted links by 3D conformation data

We compared the links predicted by CT-FOCS, CT-JEME and CT-MAD-FOCS to experimentally measured 3D chromatin loops, defined as the positive set. We chose the CT versions of these algorithms, which make the same number of calls, in order to allow fair comparison. In GM12878, using POLR2A ChIA-PET, CT-JEME achieved the best precision (21%) followed by CT-MAD-FOCS (19%) and CT-FOCS (10%). In K562, using YY1 HiChIP, CT-FOCS achieved the best precision (17.5%) followed by CT-MAD-FOCS (14%) and CT-JEME (3.45%). The low precision shows that single loops do not support the majority of the links predicted by any method.

Repeating the comparison using TLSs instead of single loops resulted in 2-3 fold increase in precision compared to single loop validation in all methods. On GM12878 loops, precision was 54%, 50% and 30% in CT-JEME, CT-MAD-FOCS and CT-FOCS, respectively. On K562 loops, precision was 33%, 28% and 22% in CT-FOCS, CT-MAD-FOCS and CT-JEME, respectively. Again, the precision obtained by TLS validation for all methods was still low.

We repeated the same analysis on the ENCODE DHS dataset, comparing CT-FOCS to CT-TargetFinder and CT-ABC. Here, CT-FOCS performed markedly better in validation based on both single loops and TLSs. For example, on GM12878 with single-loop validation, CT-FOCS achieved 31% precision while CT-TargetFinder and CT-ABC model achieved 10% and 13%, respectively. With TLS validation, CT-FOCS had 66% precision while CT-TargetFinder and CT-ABC model achieved 30% and 47%, respectively. Similarly, on K562 with single loop validation, CT-FOCS had 54% precision, CT-ABC 30% and CT-TargetFinder 1.4 %. With TLS validation, CT-FOCS had 74% precision, CT-ABC 43% and CT-TargetFinder 3.7%.

Overall, ct-links predicted by all methods had relatively low support from 3D chromatin loops. CT-FOCS tended to achieve higher precision than the other tested methods.

Assessing cell type-specificity via 3D experimental loops

As an additional test, we checked to what extent ct-links called on different cell types are supported by TLS loops that are called from GM12878's POLR2A ChIA-PET data. If ct-links called by a certain prediction method on GM12878 are indeed highly specific, we expect

GM12878 to show the highest support rate in this analysis. To quantify this, we defined for each cell type, the logarithm of the ratio between the validation rate observed in GM12878 and the validation rate observed for that cell type. For most cell types we expect to obtain values>0. Indeed, CT-FOCS ct-links predicted for GM12878 showed significantly higher support rate compared to the ct-links that were predicted in most other cell types (median $\log_2$(ratio) ~1.7; **Figure 3.4A**). Moreover, the six cell types that showed higher validation rate than GM12878 (that is, had $\log_2$(ratio)<0; **Figure 3.4A: CT-FOCS boxplot**) were all biologically related to GM12878 (e.g., B cell line and Burkitt's lymphoma cell line). CT-MAD-FOCS and MAD-FOCS performance was significantly lower (median $\log_2$(ratio) ~1.1), followed by CT-JEME (~0.7) and JEME (~0.6). Note that in this analysis too, the comparisons between CT-FOCS, CT-MAD-FOCS and CT-JEME are more proper, since these methods have a similar number of predictions per cell type (and thus, comparable recall). The results for MAD-FOCS and JEME are added only for reference. The results were more significant in favor of CT-FOCS when considering only TLS anchors overlapping GM12878 H3K27ac peaks downloaded from ENCODE (**Supplementary Table 2A**). We obtained similar results when validating against ChIA-PET single loops (**Figure 3.4B**), and when using HiChIP from K562 (**Figure 3.4C**). When using PCHi-C, HiChIP and ChIA-PET for eight individual tissues**,** CT-FOCS performed best overall (**Figure 3.4D** and **Supplementary Table 2A**).

We repeated the analysis of CT-FOCS, CT-MAD-FOCS, CT-TargetFinder and CT-ABC, now using ct-link predictions derived from the ENCODE dataset (**Supplementary Table 2B**. Interestingly, CT-MAD-FOCS obtained the highest precision and TLS support on GM12878. On K562, all methods had rather low performance ($\log_2(ratio) \simeq 0$). Note, however, that the number of cell types compared was very low (5-10 cell types, compared to 276 for FANTOM5), so these results are anecdotal.

Overall, on FANTOM5 dataset, the particularity of the links of CT-FOCS was higher than those of CT-MAD-FOCS and CT-JEME.

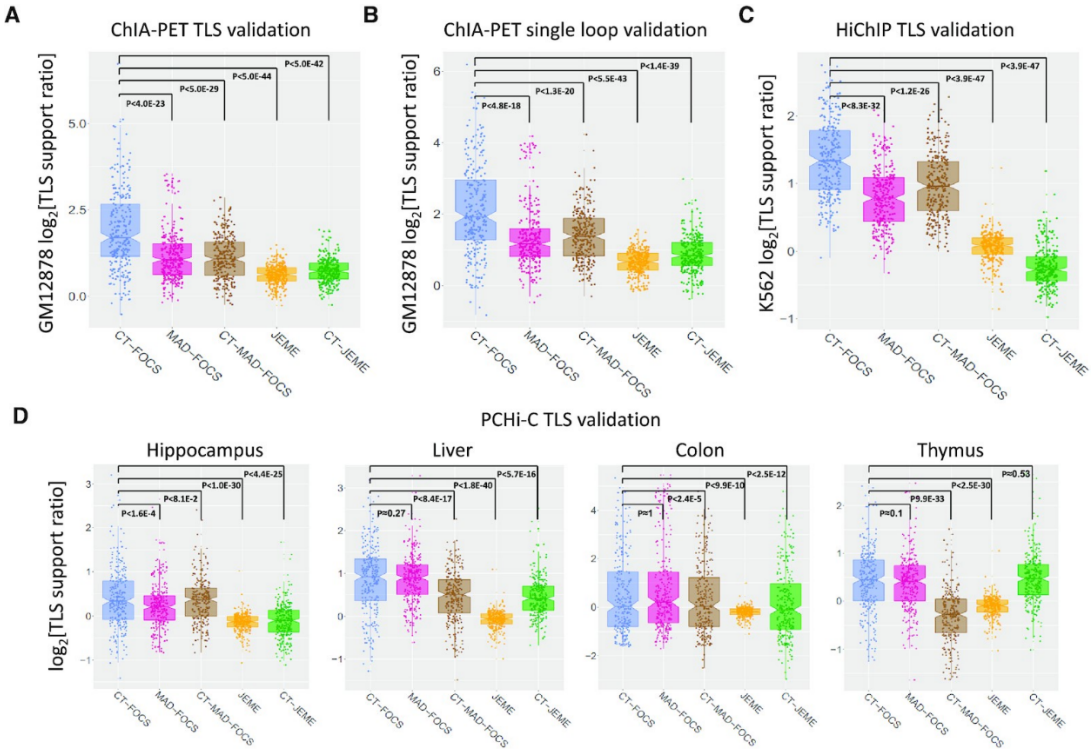


Figure 3.4. The particularity of each algorithm's predictions as measured by ChIA-PET, HiChIP, and PCHi-C assays. (A-B) Each algorithm was applied to each cell type, and the predicted links were benchmarked against GM12878 ChIA-PET loops and TLSs. Comparison included 276 FANTOM5 cell types that had at least 50 predicted EP links in CT-FOCS, MAD-FOCS, CT-MAD-FOCS, JEME and CT-JEME. The plots show, for the indicated cell type, the distribution of the ratios between the percentage of predicted EP links on GM12878 that had GM12878 ChIA-PET support and the percentage of predicted links in that cell type that had GM12878 ChIA-PET support (**Methods**). **(A)** ChIA-PET TLS support. **(B)** ChIA-PET single loop support. **(C)** The same analysis as in (A) for K562 cell line compared to TLSs derived from K562 HiChIP assay. **(D)** The same analysis as in (A) but here using TLSs derived from PCHi-C in four additional cell types and tissues. All comparisons are summarized in **Supplementary Table 2**. *p*-values are based on one sided Wilcoxon paired test.

Predicted ct-links drive cell type-specific gene regulation

We next asked whether the enhancers and promoters in the ct-links inferred by CT-FOCS demonstrate signals of cell type-specific transcriptional regulation, as shown previously for lineage-determining TFs [142] and in K562 [128]. To this end, we searched for occurrence of 402 known TF motifs (position weight matrices; PWMs) within the enhancers and promoters of the inferred links. To lessen false discoveries, we restricted our search to digital genomic footprints (DGFs; **Methods**), which are short genomic regions (~20 bp on average) identified by DHS that tend to be stably bound by TFs [143]. We used ~8.4M reported DGFs in the human genome, covering 41 diverse cell and tissue types derived from ENCODE DHS data

[112]. For each TF and cell type, we calculated the overrepresentation factor of the TF motif in the target set (enhancers or promoters of the inferred ct-links) compared to a matched control set harboring a similar nucleotide distribution (**Methods**).

We first applied this test to the ct-links predicted on GM12878 using the ENCODE DHS dataset. 13 overrepresented TFs were identified in promoters, and a different set of 13 TFs was identified in enhancers. These TFs showed on average higher overrepresentation in both enhancers and promoters compared to their occurrence in the ct-links inferred for other cell types (**Figure 3.5A-B**). In terms of the specificity score of the TF overrepresentation factors, GM12878 ranked first in both enhancers and promoters (**Figure 3.5C-D**).

Many of the TFs whose motifs were detected as overrepresented on GM12878 ct-links have known roles in regulation of B cell lineage commitment [144,145]. Among them are the EBF TF 1 (*EBF1*) and the interferon regulatory factor 4 (*IRF4*) (which had, respectively, the 2nd and 8th highest overrepresentation factors in GTM12878 ct-link promoters), and the paired box 5 (*PAX5*) and the interferon regulatory factor 8 (*IRF8*) (ranked 7th and 11th in enhancers, respectively). Furthermore, *EBF1*, *SPI1*, *BATF*, *RUNX3*, *IRF4*, and *PAX5*, detected by our analysis, were shown to cooperate with the *STAT5A-CEBPB-PML* complex, predicted to be involved in chromatin looping. Since these cofactors exhibit GM12878-specific expression (**Supplementary Figure S3.9)**, they define highly specific chromatin binding profile for the *STAT5A-CEBPB-PML* complex in GM12878, which does not appear in the related K562 cell line [146]. Note that while Zhang et al. 2016 [146] used ChIP-seq data from multiple TFs as well as Hi-C data to identify TF complexes involved in chromatin looping in GM12878 and K562 cell lines, our method requires data generated by only a single omics technique to pinpoint putative TF complexes that mediate EP chromatin looping for hundreds of cell types.

Next, we applied this TF motif overrepresentation analysis and specificity ranking on the ct-links inferred from ENCODE DHS data for 68 cell types that had at least 50 predicted EP links. The analysis identified an average of 12 overrepresented TF motifs in enhancers and 19 in promoters, per cell type (**Supplementary Table S3**). Calculating cell-type specificity scores based on the set of overrepresented TFs detected on the ct-link's enhancers in each cell type, ranked the studied cell type as the top one in 57 out of the 68 cell types. Similarly, using the set of overrepresented TFs detected on the ct-link's promoters, ranked the studied cell type as the top one in 58 out of 68 cell types.

Last, we applied this analysis on 276 FANTOM5 cell types that had at least 50 predicted EP links in all methods. CT-FOCS analysis identified an average of 16 TFs in enhancers and 25

in promoters per cell type (Supplementary Table S4). JEME identified 33 and 69, CT-JEME identified 17 and 35, MAD-FOCS identified 9 and 20, and CT-MAD-FOCS identified 9 and 5, respectively. CT-FOCS ranked the studied cell types first in ~57% and ~61% of the cases for enhancers and promoters, respectively, while the other methods ranked first ~1-37% in enhancers and 2-53% in promoters, with CT-MAD-FOCS showing the lowest numbers. Overall, CT-FOCS tended to find TFs that are more cell type-specific.

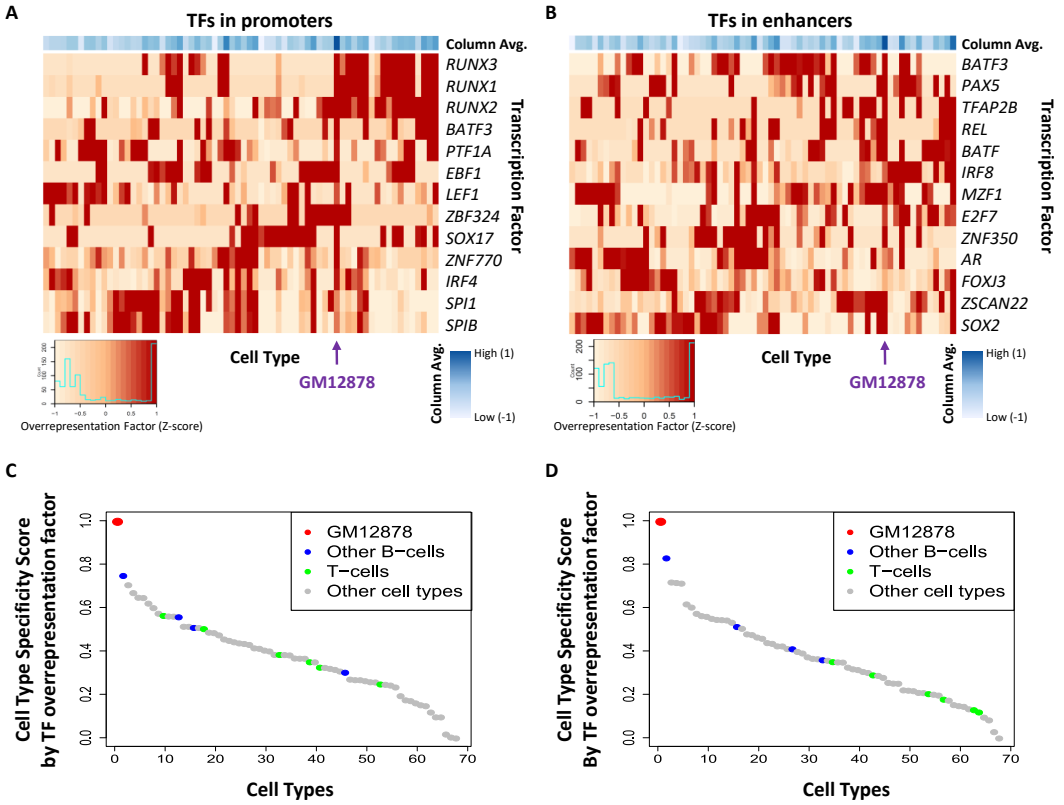


Figure 3.5. Overrepresented transcription factor motifs in enhancers and promoters of GM12878 ct-links. (A,B) Heatmaps of TF motif overrepresentation factor (after Z-score transformation) in promoters (A) and enhancers (B) of GM12878-specific EP links identified by CT-FOCS on ENCODE DHS data. TFs shown had q-value < 0.1 (Hyper Geometric test). **(C-D)** Cell type specificity score ranks based on GM12878-specific TF overrepresentation factors in promoters (C) and enhancers (D) compared to other cell types.

3.1. Methods

FANTOM5 and ENCODE data preprocessing

Details on data preprocessing are provided in the Supplementary Methods sections: 'FANTOM5 CAGE data preprocessing' and 'ENCODE DHS data preprocessing'.

CT-FOCS model Implementation

Our model for promoter p (**Figure 3.1**) includes its k closest enhancers. The activity of the promoter across the n samples is denoted by the n-long vector $y_{p,}$ and the activity level of the enhancers across the samples is summarized in the matrix X_e of dimensions $n \times (k+1)$, with the first column of ones for the intercept and the next k columns corresponding to the candidate enhancers. There are $C < n$ cell types and each sample is labeled with a cell type. $k = 10$ was used.

To find ct-links based on the global links identified by FOCS, CT-FOCS starts with the full (that is, non-regularized) promoter model. We use the non-regularized promoter model as regularization reduces the overall model variance needed for making inferences. In principle, one could apply ordinary least squares regression with the cell types as additional coefficients to estimate cell type specificity. However, such models will perform poorly when the sample size is not much larger than the number of coefficients (e.g., in FANTOM5 we have 808 samples and a total of 483 coefficients: 472 cell types + k=10 enhancers + intercept). By using LMM, we can treat the cell type group level as a random effect coefficient, splitting the samples (replicates) based on their cell type of origin, at the cost of assuming a random effect distribution.

The application of an appropriate mixed effects model to the data depends on the distribution of the promoter and enhancer activities. We observed that FANTOM5 data have normal-like distribution and ENCODE data have zero-inflated negative binomial (ZINB) distribution (**Supplementary Figure S3.10**). For FANTOM5, we applied regular linear mixed effect regression. For ENCODE, we applied generalized linear mixed effect regression (GLMM).

For each promoter, we defined a null model and $k+1$ alternative models, each corresponding to a single random effect (i.e., random slope for enhancer or random intercept for the promoter). We defined the null model as the simple linear regression $y_p = X_e\beta + \epsilon$, and each of the alternative models as the LMM model $y_p = X_e\beta + Z^l\gamma^l + \epsilon$, where $X_e\beta$ is the fixed effect, $Z^l\gamma^l$ is the random effect, and ϵ is a random error. $l \in \{1, \dots, k+1\}$ is one of the variables (enhancer or the intercept). γ^l is a C-long vector of random effects to be predicted. Z^l is a $n \times C$ design matrix that groups the samples by their cell types, namely:

$$Z^l[i,j] = \begin{cases} X_e[i,l] & sample\ i\ belongs\ to\ cell\ type\ j \\ 0 & otherwise \end{cases}$$

We applied a likelihood ratio test between the residuals of the $k+1$ alternative models and the null model, and got $k+1$ p-values. Such p-values were calculated for each of the $|P|$ promoters, and corrected together for multiple testing using FDR [104], with the number of tests performed $|P| \cdot (k+1)$.

Each predicted random effect vector $\gamma^l = \left(\gamma_1^l, \ldots, \gamma_C^l\right)$ of the alternative models was normalized using the median absolute deviation (MAD), i.e., $\gamma'^l_i = |\gamma_i^l - median(\gamma^l)|/mad(\gamma^l)$, where $mad(\gamma^l) = median(|\gamma^l - median(\gamma^l)|)$ is calculated over all cell types together. If $\gamma'^l_i > 2.5$ then enhancer l (or the promoter, if $l = 1$) was regarded as having an outlier activity in cell type i. We chose a moderately conservative MAD threshold, 2.5, as suggested in [147]. We chose to use the MAD statistic since the mean and the standard deviation are known to be sensitive to outliers [147].

Finally, we defined cell type-specific EP links (abbreviated *ct-links)* as those that had: (1) significant random effect intercept of the promoter (P), (2) significant random effect slope of the enhancer (E), both with q-value < 0.1, and (3) E and P random effect values were identified as outliers in the same cell type according to the MAD criterion.

MAD-FOCS model

MAD-FOCS takes the global EP links predicted by FOCS [101]. Then, for every global EP link, MAD-FOCS calculates the E and P median activity values across the multiple replicates per cell type. Last, it normalizes the median activities across cell types using the MAD method. EP links are identified as ct-links in a certain cell type if both E and P activities are positive outliers in that cell type using MAD cutoff > 2.5.

Filtered EP links sets

To validate the cell type-specificity of predicted EP links we use experimental 3D loops as a benchmark (see next section). The very small number of cell types assayed does not allow us to identify true cell type-specific loops and exclude those common to many cell types. Therefore, the benchmark does not provide a gold standard of positive and negative ct-links. (validations against all experimentally detected loops without considering the cell type-specificity of predicted EP links are available in **Supplementary Results** and **Supplementary Figures S3.11-3.12**). To allow a fair comparison between the performance of prediction methods that produce very different numbers of links, for each method and cell type, if CT-

FOCS gave n links, then we took the subset of n top scored links predicted by that method. We call these subsets CT-X where X is the method's name.

1. CT-JEME

JEME reports a classification score (between 0.3 to 1) for every EP link representing how active the EP link is in each cell type. We created a subset of the original JEME EP links called CT-JEME. For cell type j in FANTOM5 with n CT-FOCS ct-links, we chose the top n scoring EP links of JEME as the CT-JEME subset for that cell type. For cell types in which JEME had a lower number of EP-links than CT-FOCS, we included all JEME's EP links for that cell type in CT-JEME. **Supplementary Figure S3.2A** shows that the number of EP links per cell type is similar between CT-FOCS and CT-JEME. In addition, the average number of cell types sharing an EP link is 2.9 in CT-JEME compared to 11 in JEME (**Supplementary Figure S3.2B)**.

2. CT-MAD-FOCS

To allow a fair comparison between the predictions of CT-FOCS and MAD-FOCS, we created a subset of MAD-FOCS EP links called CT-MAD-FOCS, as described for CT-JEME above. We sorted the EP links by their $logEP$ signal.

3. CT-TargetFinder and CT-ABC

Data for ABC model was taken from ftp://ftp.broadinstitute.org/outgoing/lincRNA/ABC/AllPredictions.AvgHiC.ABC0.015.minus150.ForABCPaperV3.txt.gz. Among the 131 biosamples analyzed in ABC, 75 were taken from ENCODE and Roadmap epigenomics consortia [25,126] and 8 of them were also present in the CT-FOCS database and used for comparison (GM12878, HeLa-S3, K562, HCT-116, HepG2, A549 and H1-hESC). As for TargetFinder, we applied the program (https://github.com/shwhalen/targetfinder) on five cell types from ENCODE (GM12878, HeLa-S3, HUVEC, NHEK and K562) for which preprocessed multi omics data was available on the TargetFinder website, using as input candidate DHS sites representing enhancers and promoters from ENCODE DHS data. For each cell type in ENCODE with n CT-FOCS ct-links, we chose the top n scoring EP links of TargetFinder (by classification score) and of the ABC model (by ABC score) as the predicted ct-links for that cell type for the two models, and called these subsets CT-TargetFinder and CT-ABC, respectively. Statistics on the analyzed data are summarized in **Supplementary Table 1A**.

External validation of predicted EP links using ChIA-PET, HiChIP and PCHi-C loops

We used 3C loops to evaluate the performance of CT-FOCS and of other methods for EP linking. We downloaded ChIA-PET data of GM12878 cell line (GEO accession: GSE72816; ~100 bp resolution) assayed with *POLR2A* [137], HiChIP data of Jurkat, HCT-116, and K562 cell lines (GEO accession: GSE99519; 5 kb resolution) assayed with *YY1* [105], and PCHi-C data across 27 tissues (GEO accession: GSE86189; 5 kb resolution) [138]. Each loop identifies an interaction between two genomic intervals called its *anchors*. In ChIA-PET data, to focus on high confidence interactions, we filtered out loops with anchors' width >5kb or overlapping anchors. Loop anchors were resized to 1kb (5kb in HiChIP and PCHi-C) intervals around the anchor's center position. We filtered out loops crossing topologically associated domain (TAD) boundaries, as functional links are usually confined to TADs [148–151]. For this task, we downloaded 3,019 GM12878 TADs [152], which are largely conserved across cell types [131], and used them for filtering ChIA-PET and PCHi-C loops from all cell types.

To overcome the sparseness of the ChIA-PET loops, and the 8kb minimum distance between loop anchors [137,153], we combined loops into two-step loop sets (**TLSs**) as follows: for every reference loop, x, its TLS is defined as the set of anchors of all loops that overlap with at least one of x's anchors by at least 250 bp (**Figure 3.3A**). We used the igraph R package [154] for this analysis.

To evaluate if a ct-link is confirmed by the ChIA-PET data, we checked if both the enhancer and the promoter fall in the same TLS. Specifically, we defined 1kb genomic intervals (±500 bp upstream/downstream; 5kb genomic intervals: ±2.5kb upstream/downstream in HiChIP and PCHi-C) for the promoters (relative to the center position; relative to the TSS in FANTOM5 dataset) and the enhancers (relative to the enhancer's center position) as their genomic positions. Both inter- and intra- TAD predicted EP-links were included in the validation. An EP link was considered supported by a TLS if the genomic intervals of both its promoter and enhancer overlapped different anchors from the same TLS **(Figure 3.3B** and **Supplementary Figure S3.7)**.

We used randomization in order to test the significance of the total number of EP links supported by ChIA-PET single loops. We denoted that number by N_t. We performed the test as follows: (1) For each predicted EP link, we randomly matched a control EP link, taken from the set of all possible EP pairs that lie within 9,274 GM12878 TADs from Rao et al. [131], with similar linear distance between E and P center positions. We restricted the matching to the same chromosome in order to account for chromosome-specific epigenetic state [155]. The

matching was done using MatchIt R package (method='nearest', distance='logit', replace='FALSE') [156]. This way, the final set of matched control EP links had the same set of linear interaction distances as the original EP links. (2) We counted N_r, the number of control EP links that were supported by ChIA-PET single loops. We repeated this procedure for 1,000 times. The empirical *p*-value was $P = \frac{\#(N_r \geq N_t)}{1000}$, or $P<0.001$ if the numerator was zero. A similar empirical *p*-value was computed for the validation rate obtained by using single loops and TLSs.

We used the following formula to calculate the GM12878 ChIA-PET TLS support ratio:

$$ratio\left(\frac{GM12878}{CellType}\right) = \frac{\%GM12878\ specific\ EPs\ in\ GM12878\ TLS}{\%CellType\ specific\ EPs\ in\ GM12878\ TLS}$$

Calling cell-type specific active EP loops reported in a capture Hi-C study

We wished to identify cell-type specific EP links reported in capture Hi-C data [138]. We downloaded 906,721 promoter-other (PO) capture Hi-C loops generated across 27 tissues (GEO accession: GSE86189) [138]**.** These loops involve a known gene's promoter and a non-promoter region, which may be an enhancer. To define a set of strictly ct-specific loops, we retained PO loops that were detected in exactly one cell type. We set the PO anchors to 1kb intervals around their center positions. This analysis detected a median of 630 EP loops that were unique to a specific cell type.

To call promoter and enhancer regions, we downloaded 474,004 enhancer and 33,086 promoter regions predicted by a 15-state ChromHMM model on Roadmap epigenetic data across 127 tissues (https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-intersect_release/DNase/p10/enh/15/state_calls.RData; https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-intersect_release/DNase/p10/prom/15/state_calls.RData) [25]. We kept the enhancers of state Enh or EnhG (genic enhancers) in any of 127 Roadmap tissues. Similarly, we kept the promoters of state TssA (active TSS) or TssAFlnk (Flanking Active TSS). Then, we resized each region to a 1kb interval around its center position. We called the resulting sets active promoters and enhancers. A retained PO loop whose P and O anchors had at least 250 bp overlap with active ChromHMM promoter and enhancer, respectively, was considered as cell type-specific active EP loop.

Cell type specificity score

We quantified the intensity of an EP link in a given sample by $\log_2 a + \log_2 b$ where a and b are the enhancer and promoter activities in that sample. The *EP signal* of the link for a particular cell type is the average of the signal across the samples from that cell type. Define $x_c = (x_{c1}, \ldots, x_{cn})$ as the vector of signals in cell type c, where n is the total number of EP links discovered in cell type c, and define $d_{c,i}$ as the Euclidean distance between the vectors of cell types c and i, both with the same EP links from cell type c. Following the definition of [132], the *specificity score* of EP links predicted in cell type c is:

$$S_c = \frac{1}{\sum_{i \neq c} d_{c,i}} \sum_{i \neq c} d_{c,i} \sum_{k=1}^{n} \left(x_{c,k} - x_{i,k}\right)$$

Similarly, cell-type specificity can be computed for the expression values of the genes annotated with EP links, or on the overrepresentation factors of TFs found at enhancers and promoters.

Motif finding on ct-links

We examined the occurrence of transcription factor (TF) binding site motifs in sequences of ct-links' promoters and enhancers. Finding all TF motif occurrences (hits) in a large set of promoter and enhancer sequences, each hundreds of bases long, is prone to high false positive rate. We therefore limited the search for hits to digital genomic footprint (DGF) regions, very short segments that are more likely to contain genuine TF binding sites. We downloaded ~8.4M DGF sequences inferred from DNase-seq in ENCODE [112]. The mean DGF length was $L \approx 20$ bp, with a maximum length of 68 bp.

We intersected the DGFs with enhancer and promoter regions of predicted ct-links. We call the resulting set of sequences the *target set*. We looked for hits of 402 HOCOMOCO V11 [71] TF core motifs (taken from MEME suite database [66]; http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.18.tgz) in the target sets. Hits were found using FIMO [70] with 0-order Markov model as background created using fasta-get-markov command from MEME suite [69]. For each TF, matches with FIMO q-value<0.1 were considered hits. To evaluate the statistical significance of the findings we repeated the search on a control set from matched regions (one per target region) having similar distribution of single nucleotides and dinucleotides. Matching was done using MatchIt R package [156] (method='nearest', distance='mahalanobis'). For each TF we used a one sided Hyper-Geometric (HG) test to compare between the prevalence of its hits in the target and background (target+control) sets. Motifs having q-value < 0.1 were selected.

If a k-long TF motif had l_t hits on a target set containing m_t possible k-mers in total (in both strands) and the same motif had l_b hits in the background set containing m_b possible k-mers, then the *overrepresentation factor* of the TF is defined as $(l_t/m_t)/(l_b/m_b)$. To avoid division by zero we used the Laplace correction (adding +1 to all four terms). If l_t was zero then we set the overrepresentation factor as 1.

Statistical methods, visualization and tools

All computational analyses and visualizations were done using the R statistical language environment [157]. To correct for multiple testing we used the p.adjust() function (method='BY'). We used 'GenomicRanges' package [118] for finding overlaps between genomic intervals. We used 'rtracklayer' [119] and 'GenomicInteractions' [120] packages to import/export genomic positions. Linear mixed effect regression models were created using lme R function from nlme package [158]. Generalized linear mixed effect with zero inflated negative binomial models were created using glmmTMB R function from glmmTMB package [159]. Counting reads in genomic intervals was done using BEDTools [109]. Graphs were created using graphics [157], ggplot2 [123], gplots [124], ComplexHeatmap [160], and the UCSC Genome Browser (https://genome.ucsc.edu/).

4. Inferring transcriptional activation and repression activity maps in single-nucleotide resolution using deep-learning

Regulatory elements that control transcription such as enhancers and promoters have been studied extensively over the past two decades [2]. In contrast, silencers, which turn-off or reduce the transcription of their target genes, have received less attention, mainly because they are harder to verify experimentally.

Classification methods aiming at predicting cell type-specific functional enhancers and promoters have used sequence and epigenetic data. Positive enhancers and promoters used as training, validation and test sets were mainly labeled using epigenetic data. For example, functional enhancers are known to be marked by the H3K4me1 and H3K27ac histone modifications within their genomic context. In contrast, it is not clear which epigenetic marks define functional silencers.

Recently, an ATAC-STARR-Seq study experimentally identified many enhancer and silencer elements in GM12878 cell line [35], each with per-nucleotide contribution scores to activation or repression of the target gene's expression. Using these data, we aimed to develop a more robust method for enhancer and silencer classification. Feature importance techniques applied on the trained model were used to pinpoint the precise epigenetic combination defining functional silencers. We compared three published DL models; each implemented a different architecture from one or both CNN and RNN families. The first is a CNN-based model composed with five convolutional layers [161]. The second, deepTACT, is composed with one convolution layer and additional RNN layers to study relationships between adjacent positions in the input sequence [45]. The third, ResNet, is a combination of convolution layers and residual network blocks composed of RNN layers to avoid the vanishing gradient problem [86]. The last two were originally designed for a different classification task and we modified them for our task.

In this study we investigated the following question: will a deep learning (DL) model trained on DNA sequences labeled as enhancers and silencers using experimental identified elements be less or more accurate compared to the same DL model trained on REs labeled using epigenetic data?

In addition, previous methods that predicted silencers have used the DNA sequence only as an input. Our second question was whether incorporating epigenetic data alongside the sequence in the input will improve prediction performance.

Lastly, we hypothesized that the abovementioned ATAC-STARR-Seq study might miss true enhancers and silencers. To this end, we trained a DL model on experimentally identified enhancers and silencers and predicted 3,752 novel enhancers and 518 novel silencers on a set of genomic sequences without overlapping experimentally identified elements. Downstream analyses such as TF and GWAS enrichments within these novel enhancers and silencers were used to provide support for these novel REs being genuine ones.

4.1. Results

Training on experimentally identified regulatory elements improves predictive accuracy of silencers models

Due to lack of broad sets of experimentally identified silencers, the computational models for silencers developed by Huang and Ovcharenko [161] were trained on sets of putative silencers that were defined based on their epigenomic profile rather than on experimentally detected silencer elements. The recent ATAC-STARR-Seq study by Hansen et al. provides extensive sets of identified enhancer and silencer elements in the lymphoblastoid cell line GM12878 [35]. Therefore, first, we wished to compare the performance of silencer models trained on putative silencers that were defined based on epigenomic marks to the performance of models trained on experimentally identified silencers.

Following the epigenetic criteria used by Huang and Ovcharenko, we defined as the set of putative silencers in GM12878 all H3K27me3 peaks not overlapping either H3K27ac, H3K4me1 or H3K4me3 peaks in this cell line. In parallel, we defined a set of putative enhancers that are active in GM12878 as the regions of ATAC-seq peaks overlapping H3K27ac, but not H3K27me3 peaks in this cell line. We also defined a background set of regulatory elements that are non-functional in GM12878 as regions of ATAC-seq and H3K27me3 peaks randomly chosen from five other cell types that were not detected in GM12878. Overall, this epigenetic approach defined 41,548 enhancers, 24,554 silencers and 396,612 nonfunctional peaks. We applied the Convolutional Neural Network (CNN) method introduced by Huang and Ovcharenko on the GM12878 training set using 1kb sequence as the only feature, and evaluated how accurately it classified the experimentally identified elements detected by ATAC-STARR-Seq in this cell line (22,336 enhancers, 19,289 silencers and 175,088 nonfunctional ATAC-seq peaks; **Methods**). The CNN model achieved for enhancers 0.3 AUPRC, and for silencers 0.06 AUPRC (**Supplementary Fig. 4.1**).

Next, we applied the same CNN method, but now trained the model using the sequences identified experimentally as regulatory elements by ATAC-STARR-Seq (Hansen et al.) Chromosomes 1-5, 9-22 and X constituted the **training set**. Chromosome 6 was used as a **validation set** for tuning the model's hyper-parameters. The **test set** used for evaluation of the model's performance included chromosomes 7 and 8.

The predictive performance of enhancer models trained on the experimentally identified enhancers was 0.37 AUPRC, a bit higher than the performance obtained by the enhancer models trained on putative enhancers defined based on epigenomic marks (0.3 AUPRC). In contrast, for silencers, the performance of the models trained on experimentally identified silencers was 0.77 AUPRC, dramatically higher than that obtained by the silencer model trained on REs defined by epigenomic marks (0.06 AUPRC) (**Supplementary Fig. 4.1**). This result reflects the much better knowledge we currently have on epigenomic marks defining active enhancers compared to those that mark active silencers. Furthermore, as extensive sets of experimentally identified enhancers and silencers are available for only a limited number of cell lines, our result indicates that the availability of epigenomic profiles for canonical marks in various cell lines is sufficient for reasonable prediction of enhancers in these cells, but it does not allow accurate prediction of the landscape of active silencers.

Improved deep-learning model for prediction of enhancer and silencer elements

Next, we aimed to build a DL model for regulatory elements with improved accuracy. We reasoned that a DL model can utilize the quantitative output measured by STARR-Seq for the effect of the probed genomic intervals on transcriptional activity, rather than using discrete classes (Enhancer/Silencer/Non-functional categories) in the model learning phase. Therefore, we implemented a two-steps model as follows: Step 1 implements a regression model that predicts, in a single-nucleotide resolution, activation and repression effects in the trained cell type. Step 2 is a 3-class classification model built upon the trained regression model (**Fig. 4.1a**). The input to our model are 1kb sequences of ATAC-seq peaks together with epigenetic signals of DNA methylation, H3K27ac, and H3K4me1 in that interval (**Fig. 4.1b;** see next section for how we selected the epigenetic marks).

The regression model was built using activation and repression profiles measured for GM12878 ATAC-Seq peaks by STARR-Seq in 50-bp windows [35] (**Methods**). These windows were computationally merged to 21,125 silencers and 30,078 enhancers. We also generated an exploratory set composed of 70,937 GM12878 ATAC-seq peaks that did not overlap any silencer or enhancer identified by ATAC-STARR-Seq in this cell line. These peaks were

excluded from the training phase and used in downstream analyses. We tested three different DL architectures previously used in genomic analyses: deepTACT [45], CNN [161] and ResNet [86]. We also tested a simple linear regression as a baseline model. In each DL architecture, we replaced the last layer by a new dense layer that outputs 1,000 regression scores, one per position in the input sequence (**Fig. 4.1a**; **Methods**). Models were compared based on their classification performance in the second step.

In Step 2 we implemented a 3-category classification model by appending two dense layers to the regression network, to account for dependency between adjacent nucleotides' activation and repression levels. The first layer consists of 300 outputs, and the second, final layer, has three outputs, corresponding to the classes to be predicted: enhancer, silencer and nonfunctional. The predicted class is the one receiving the highest probability.

For the classification task, input 1kb sequences were labeled using the following scheme: (1) we scored each sequence by summing over the activation and repression levels at every nucleotide, (2) we divided the sequences into two sets: those with positive and negative sums, (3) in the positive set, the top 25^{th} percentile were labeled as enhancers, (4) in the negative set, sequences at the bottom 25^{th} percentile were labeled as a silencer, (5) all other sequences were labeled as nonfunctional. Overall, the 85% and 76% of the silencers and enhancers called by the original ATAC-STARR-Seq matched the labels they got by this scheme.

We again used enhancers and silencers from chromosomes 1-5, 9-22 and X for the **training set**. Elements from chromosome 6 were used as a **validation set**, and the **test set** included he elements from chromosomes 7 and 8. The three DL architectures had similar performance (**Fig. 4.1c**), and all performed better than the simple linear regression model. All DL models performed quite well in predicting silencers (AUPRC 0.81-0.86), and much better than the sequence based model of Huang and Ovcharenko [161] (AUPRC 0.77; **Supplementary Fig. 4.1).** Performance of the DL models in predicting enhancers were much lower (AUPRC 0.51-0.55). This might be attributed to the fact that the observed activation levels of enhancers are not clearly distinguishable from the nonfunctional levels (**Fig. 4.2**). Overall, the deepTACT model performed best in predicting both enhancers and silencers. Thus, we used this model in downstream analyses.

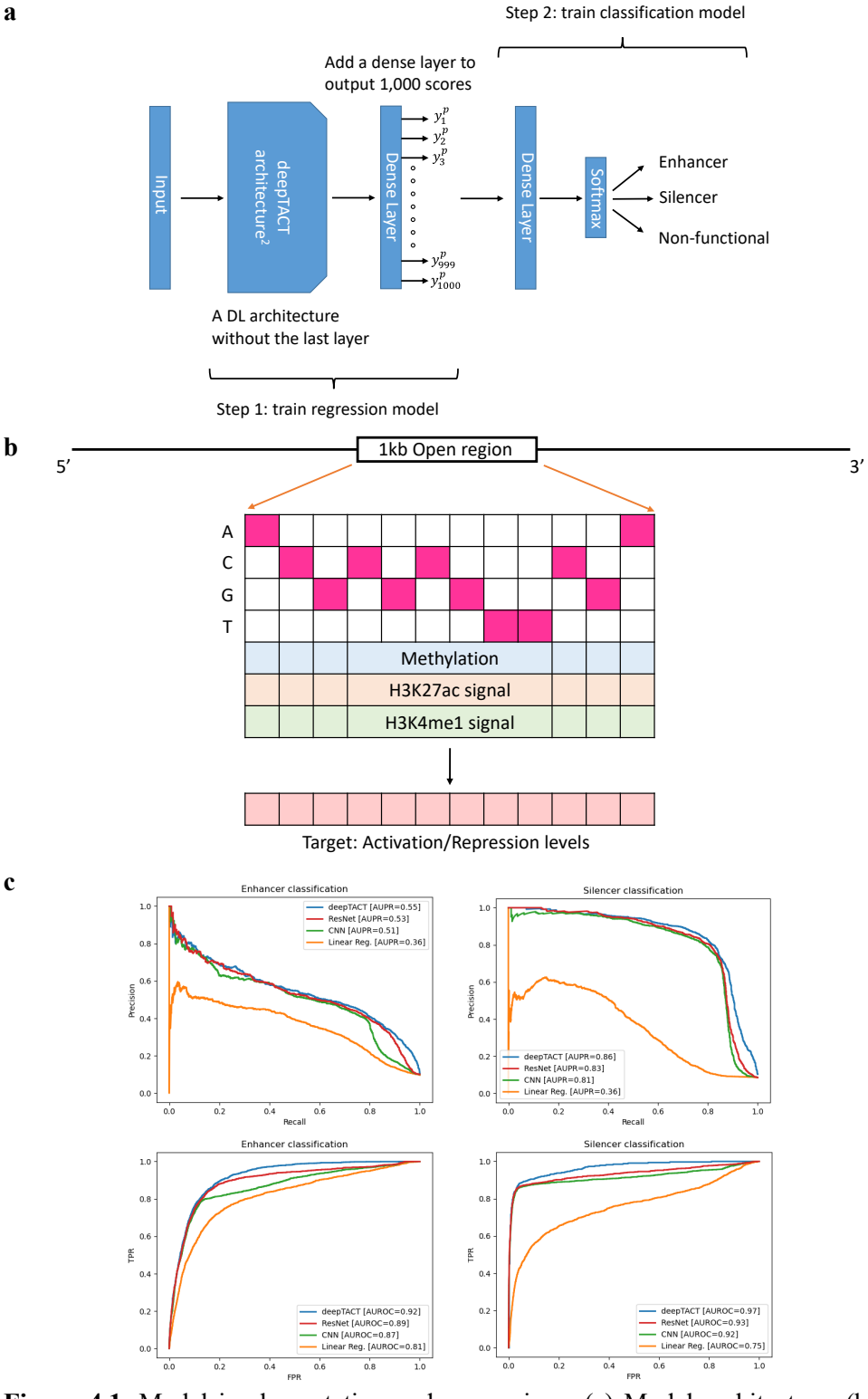


Figure 4.1. Model implementation and comparison. (a) Model architecture. (b) Schematic figure of the input and output structure. (c) Performance of the models (**Methods**).

Epigenetic markers improve prediction performance

The silencers prediction models developed by Huang and Ovcharenko used only sequence information as input. Our DL model utilizes also epigenetic data. Therefore, next, we examined whether the addition of epigenetic information improves the prediction performance. To this end, we trained the deepTACT model on sequences alone or on sequences together with combinations of additional epigenetic markers. Indeed, our result shows that adding the epigenetic data, and specifically H3K27ac and H3K4me1 signals, improved the prediction performance of our model, with more prominent improvement obtained for enhancers (AUPRC improves from 0.29 to 0.54 for enhancers and from 0.76 to 0.85 for silences) (**Supplementary Table 4.1**).

When plotting the average signal across sequences of predicted classes, we found that our model captures epigenetic signals that were not part of the input training data and are relevant to the activity of enhancers and silencers (**Fig. 4.2**). For example, high signal for the transcriptional co-activator P300 (EP300), a histone acetyltransferase known to bind active enhancers, was obtained within predicted enhancers but was markedly depleted within silencers. On the other hand, in flanking nucleosomes of predicted silencers we observed high signals for the enhancer of zeste homolog 2 (EZH2), which is part of the PRC2 complex, and for H3K27me3. In addition, predicted silencers seem to be more methylated compared to the other two classes. EZH2 can also serve as an activator [162], which could explain the high signals it obtained at the center of the predicted enhancers in the test set (**Fig. 4.2**).

Overall, silencers predicted by our model tend to be more methylated and more strongly marked by H3K27me3 than enhancers (**Fig. 4.2**). On the other hand, as expected, predicted enhancers tend to be marked by H3K27ac and H3K4me1 and bound by EP300.

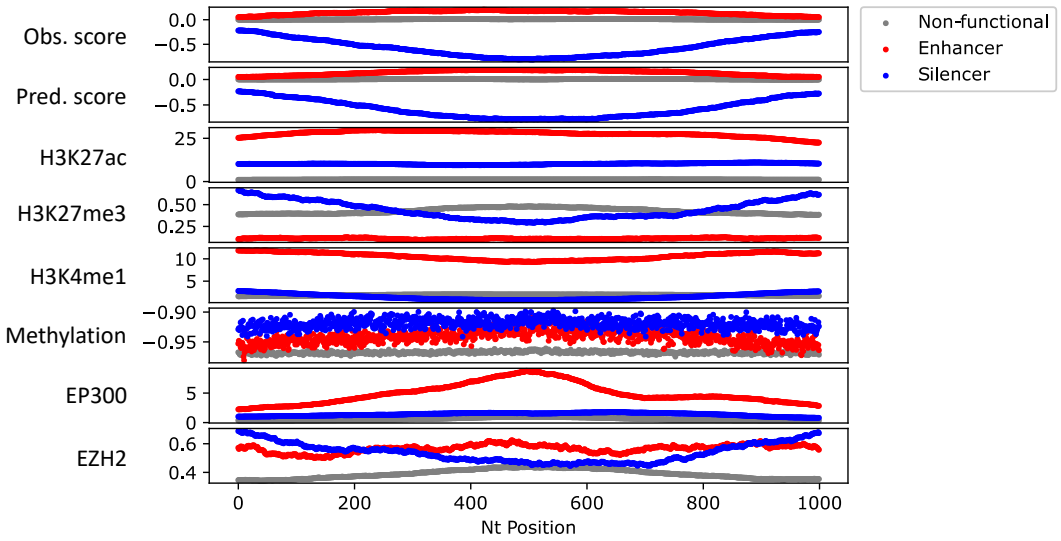
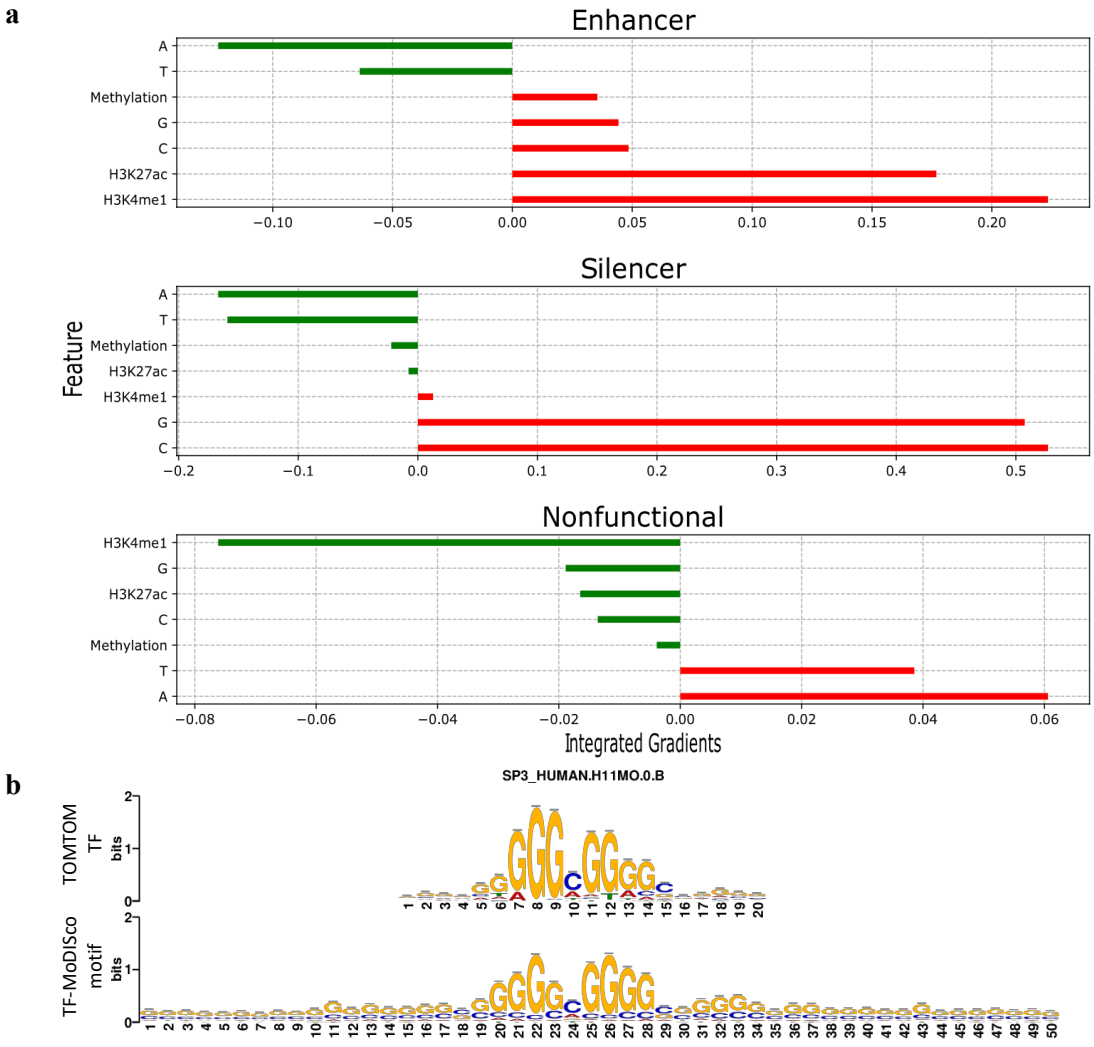


Figure 4.2. Summary of epigenetic markers in the test set. Top to bottom: observed scores (as measured by STARR-Seq), predicted scores (output of Step1 – the regression model), H3K27ac, H3K27me3, H3K27me1, Methylation, EP300 and EZH2. Predicted enhancers, silencers and nonfunctional are marked by red, blue and grey colors, respectively. In each predicted class and each track, the average signal per position in the 1kb sequences is shown. In b, the grey curve overlaps the blue curve for H3K27ac and the red curve for the EZH2.

Next, we set to determine which features contributed the most to the classification. For this task, we used the integrated gradients (IG) approach [89] (**Methods**), which calculates feature importance scores per input sample given their labels. The sign of these scores indicate a positive or negative correlation between the feature signal and the classification score. The magnitude of these scores indicates the contribution of the feature to the classification score. We applied this approach to input sequences in the test set given their labels. We found that enhancer classification scores were most positively correlated with H3K4me1 and H3K27ac levels followed by the DNA bases C and G, and DNA methylation features (**Fig. 4.3a**). The contribution of both H3K27ac and methylation is in agreement with previous findings of their bivalent role in enhancers [163]. In addition, methylation is associated with GC-rich regions, and, as expected enhancers tend to be GC-rich. Interestingly, the G and C features were the only major contributors to silencer classification, with little contribution from the epigenetic marks. This could be attributed to the fact that silencers tend to be closer to TSSs compared to enhancers (mean distance: 18,782 bp vs. 52,324 bp; **Supplementary Fig. 4.2**). Regions closer to TSSs, e.g., promoters, are highly GC-rich [164]. Both enhancer and silencer classification scores were negatively correlated with A and T features. In contrary to enhancer and silencer classifications, classification scores for nonfunctional elements were most positively correlated

with A and T features. Chromatin in AT-rich regions is more compacted than in GC-rich regions [165], which could explain why nonfunctional regions are AT-rich. Nonfunctional classification scores were strongly negatively correlated with H3K4me1 levels, as this marker is mostly associated with enhancer regions.

Next, we used the feature importance scores to find enriched motifs in the sequences using TF-MoDISco [166] (**Methods**). We identified one motif within silencers in the test set. This motif matched the binding motif of SP2 and SP3 TFs (using TomTom [167]) (**Fig. 4.3b**), which bind GC-rich elements. A richer set of 8 motifs was found within enhancers in the test set. Among the motifs, one matched Myocyte enhancer factor (MEF) TFs (**Fig. 4.3c**), and others matched known B-cell TFs such as: PRDM6, BCL11A, and IRF3 (Supplementary Table 4.2).



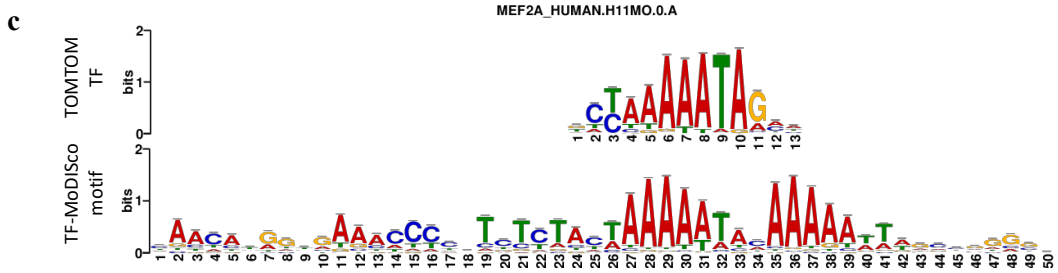


Figure 4.3. Feature importance scores computed for each class on the test set. (a) We used the integrated gradients approach to assign feature importance scores to the sequences per class: enhancer (top), silencer (middle) and nonfunctional (bottom). Positive or negative importance scores reflect a positive or negative correlation between the feature and the classification score, respectively. The magnitude of these scores measures the contribution of the feature to the classification score. (b) The top enriched motif in silencers as computed by TF-MoDISco and the corresponding known TF matched by TomTom. (c) Same as (b) for enhancers.

deepTACT predicts novel enhancers and silencers in GM12878

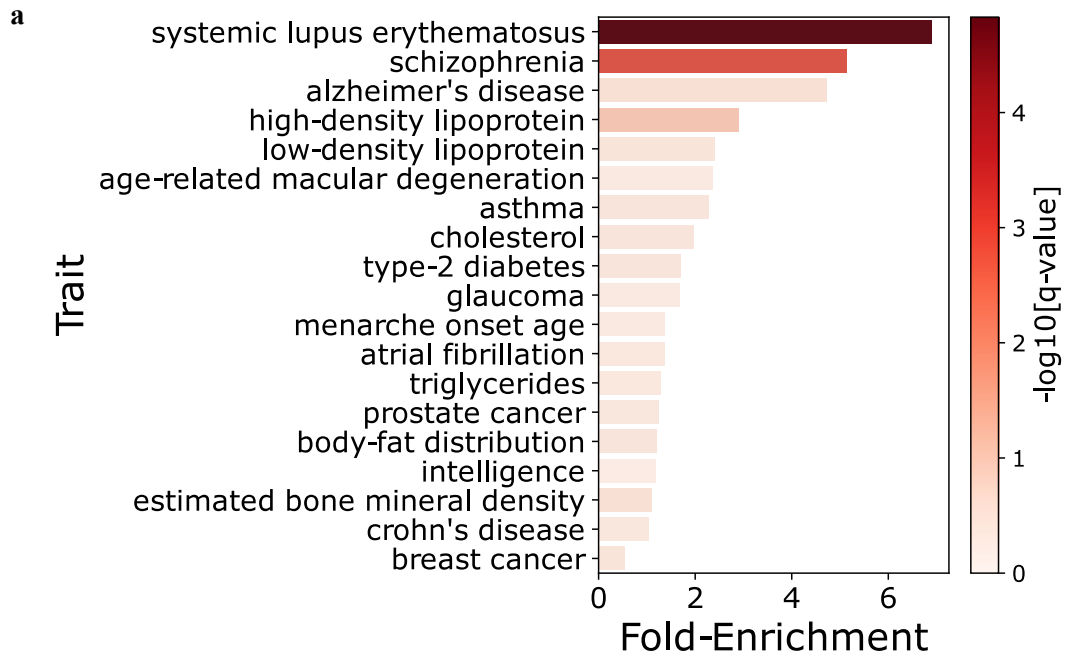
We applied the trained deepTACT model on the ATAC-seq peaks in the exploratory set (containing the set of 70,937 ATAC-seq peaks in GM12878 that were not detected by the ATAC-STARR-Seq assay as having an effect on transcription) in order to find novel enhancers and silencers in GM12878 which were missed by the ATAC-STARR-seq experiment (**Methods**). The model predicted 3,752 novel enhancers and 518 novel silencers. The epigenetic marks on these predicted elements are similar to those obtained on the experimentally identified enhancers and silencers (**Fig. 2; Supplementary Figure 4.3**).

To provide further support for the functionality of these novel predictions, we examined their enrichment for eQTLs and GWAS variants. We used eQTL data from Lymphoblastoid cell lines downloaded from the GEUVADIS database (http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/GEUVADIS/ge/; **Methods**). Using logistic regression and accounting for the potential confounding effect of distance to nearest TSS (**Methods**), we found that the set of novel enhancers predicted by our model is significantly enriched for eQTLs (P<3.1E-24; compared to ATAC-seq peaks not predicted as enhancers/silencers). We observed no eQTL enrichment in the set of predicted silencers, possibly due to their low number (n=518). On the other hand, the sets of experimentally identified enhancers and silencers were both enriched for eQTL signal (P<8.0E-27 and P<6.7E-45, respectively).

Next, we used GWAS summary statistics for 50 diseases and traits from Groenewoud et al. [168]. In each one, we kept the SNPs with p-value < 10^{-7}. When performing enrichment analysis

of the SNPs in each predicted class, we found that the set of experimentally identified enhancers was enriched for systemic lupus erythematosus (SLE) risk SNPs (Q<5.2E-5; **Supplementary Fig. 4.4a**), an autoimmune disease involving B-cells, as well as for schizophrenia (SCZ) risk SNPs (Q<5.4E-7), in line with a study that implicated increased levels of B-cell cytokines and autoantibodies in SCZ [169]. The silencers were also enriched for some diseases albeit at lower statistical significance (**Supplementary Fig. 4.4b**). Reassuringly, the set of novel enhancers predicted by our model was also enriched for SLE (**Fig. 4.4a**; Q<1.5E-5) and schizophrenia risk SNPs (Q<1.2E-3). No enrichment for GWAS risk SNPs was found within the set of predicted silencers.

Among the SLE risk SNPs in the novel enhancers is rs8052690, located within an enhancer that interacts, according to C-HiC analysis, with the promoter of the IRF8 gene [170] (**Fig. 4.4b**). As an another example, the SLE risk SNP rs13240595, which has ~2.5-fold enhancing effect as measured using MPRA [171], is located within a novel enhancer, which is predicted (by FOCS [101] and GeneHancer [170] enhancer-promoter maps) to interact with the promoter of the TNPO3 gene. TNPO3 was previously shown to be associated with SLE (**Supplementary Fig. 4.5**) [172].



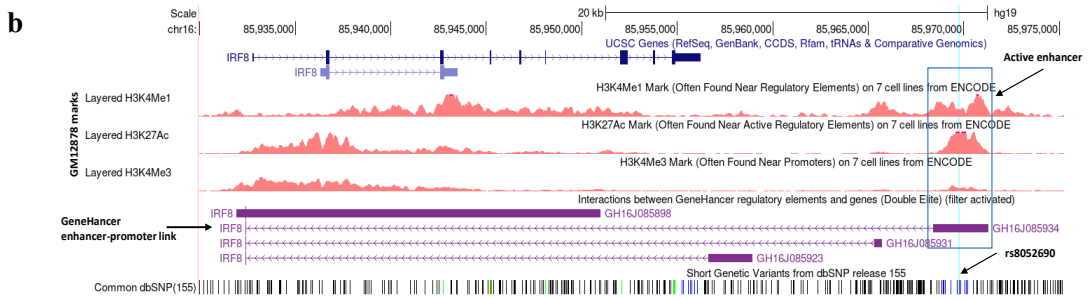


Figure 4.4. Enrichment of GWAS risk SNPs within predicted enhancers. (a) Enrichment for GWAS SNPs. Traits with at least one risk SNP overlapping an element in the exploratory set are shown. q-values are FDR-corrected Hypergeometic test p-values. (b) UCSC genome browser tracks of SLE risk SNP, rs8052690 (marked in arrow), falling within a predicted active enhancer that physically interacts with the promoter of IRF8.

Predicted novel enhancers and silencers are enriched for binding motifs of known transcriptional activators and repressors

To further support the functionality of the novel enhancers and silencers predicted by our model for GM12878 in the exploratory set, we performed motif enrichment analyses (**Methods**). Using very stringent cutoffs of q-value=1E-40 and 1.5 fold-enrichment, 54 motifs were found within the novel enhancers, including some well-established B-cell TFs: PAX5, IRF8, BCL11A and SPIB (**Supplementary Fig. 4.6a**). 42 (78%) of these motifs were also found among the 93 enriched TFs detected in the set of experimentally identified enhancers. Within the novel predicted silencers, we detected four enriched TFs (**Supplementary Fig. 4.6b**). Among them, ZBTB17 and PATZ1 were implicated as transcriptional repressors [173]. These four enriched TFs were also found among the 146 enriched TF motifs detected in the set of experimentally identified silencers.

In addition to motif analysis, we also examined enrichment for physical TF binding sites in GM12878. To this goal, we downloaded all 154 available GM12878 ChIP-seq experiments from ENCODE project and analyzed their enrichment within the predicted and experimentally identified sets of enhancers and silencers. For the novel silencers, using stringent cutoffs of q-value=1E-20 and at least 10 fold-enrichment, we found four enriched proteins (**Fig. 4.5a**), SUZ12, HDAC6, EZH2 and NRF1. Notably, SUZ12 and EZH2 are members of the PRC2 complex, which represses transcription [174]. HDAC6 is a histone deacetylate and marks epigenetic repression [175]. The experimentally identified silencers were enriched for binding of 35 proteins (**Supplementary Fig. 4.7a**)

The predicted enhancers were enriched for 26 proteins (**Fig. 4.5b),** including MAX and MYC, which when in complex act as activators in B-cells [176], and IRF3, which is known to be

involved in B-cell functions [177]. 18 out of 26 enriched proteins (~69%) were also enriched within the experimentally identified enhancers (**Supplementary Fig. 4.7b**).

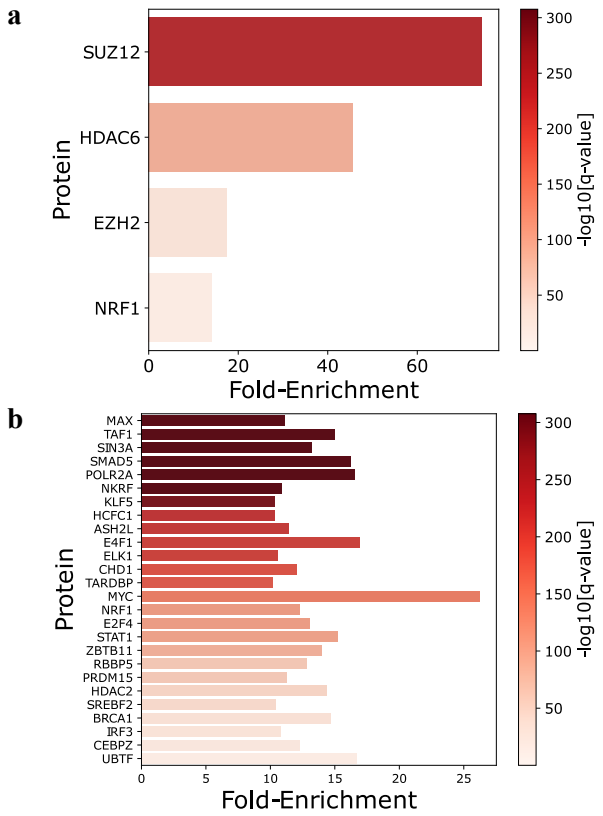


Figure 4.5. ChIP-seq enrichment analysis in predicted enhancers and silencers detected in the exploratory set. (a) Silencers. (b) Enhancers.

4.2. Methods

GM12878 data preparation

101,896 GM12878 ATAC-STARR-seq peaks were obtained from [35] (GEO dataset GSE181317) and resized to 1kb around their central positions. Experimentally identified silencer (n=21,125) and enhancer (n=30,078) regions and their repression or activation signals, as measured by STARR-Seq in GM12878, were also taken from the same dataset. Transcriptional repression and activation signals were measured at resolution of 50 bp. ATAC-seq, H3K4me1, H3K27ac, H3K27me3 and WGBS DNA methylation signal datasets in GM12878 were downloaded from the ENCODE project (https://www.encodeproject.org/). 44,494,433 CpG sites with at least 4 mapped reads were kept. The methylation level in each CpG site is the fraction of methylated reads that cover it. CpGs with insufficient coverage were given a methylation level of -1.

The input data to our model is a 1000x7 matrix. For each of the 1000 bases, the first four features are one-hot encoding of the DNA sequence of the ATAC-STARR-seq peak, followed by nucleotide-resolution signals for DNA methylation, H3K27ac and H3K4me1. We normalized the features to mean 0 and std 1. The 1k target vector is a per-position value with a positive activation signal for enhancers, negative repression signal for silencers, and 0 otherwise.

The model was trained on data from chromosomes 1-5, 9-23. Data from chromosome 6 were used for validation of the model while tuning the hyper-parameter (the number of training epochs). Data from chromosomes 7 and 8 were held out as a test set to assess the model's performance.

For model training and testing, positive cases were ATAC-seq peaks overlapping experimentally identified enhancers or silencers. Following the approach of Huang and Ovcharenko [161], we used as negative cases ATAC-seq peaks that were detected in other five cell types, but not in GM12878. For each positive peak, six negative ones were randomly sampled from the same chromosome. Overall, our dataset contained 216,713 cases: 30,959 positive peaks and 185,754 negatives. GM12878 ATAC-seq that were not detected by the ATAC-STARR-Seq assay as having an effect on transcription were left out from the phase of model training and testing, and were used as an exploratory set in downstream analyses.

Model implementation

Our model implementation is divided into two steps: In step 1, we implemented a deepTACT model as follows: (A) we used model architecture and hyper-parameters similar to those implemented in Li et al [45]. (B) The last dense layer outputs 1,000 scores, one for each position in the input sequence, aiming to predict the activation or repression scores measured by STARR-Seq for this genomic interval. Intermediate batch normalization and Dropout layers were used to prevent overfitting. Model training was performed with the mean squared error (MSE) loss function using the 'rmsprop' optimizer. We found the number of epochs required for training the model using the MSE on the validation set. In step 2, the 1000-scores output of the last dense layer of the model in step 1 is fed into a dense layer of 300 outputs scores followed by a dense layer that outputs three scores – for predicting Enhancer, Silencer or Non-functional elements - with the softmax activation function.

Inferring enhancer and silencer intervals

Given a sequence, x, and its epigenetic signals, Step 1 of our model outputs for each position j a transcriptional activity score. The score can be positive, indicating that position j is involved

in transcriptional activation (in GM12878 cell line), negative, indicating that position j is involved in transcriptional repression, or 0 (i.e., suggesting position j is non-functional). To summarize the output, we applied a 50bp sliding window with step size of 10 on the 1,000 scores the model outputs. We define a window as an Enhancer if all scores within that window are above a certain threshold (t_e). Similarly, we define a window as a Silencer if all scores within that window are below a certain threshold (t_s). We merged overlapping windows that had the same label. We selected the t_e and t_s thresholds as those yielding the maximum $F1$ score on the test set. For enhancers, the F1 score was computed by considering as positives the true activating positions in the test set, and considering as negatives - all other positions in the test set. Predicted activating positions that matched true activating positions were considered as true positives whereas unmatched predicted activating positions were considered as false positives. The same principle was applied for silencers.

The novel enhancer and silencer windows predicted (for GM12878) in the exploratory set are provided in **Supplementary Table 3**.

Alternative models

We implemented three alternative models: (1) a simple linear regression implemented as a single dense layer in a DL model, (2) the CNN model of Huang and Ovcharenko [161], and (3) the ResNet-based model from Luo et al [86].

Comparison of models trained on either experimentally identified or on epigenetically called enhancers and silencers

We took the CNN architecture as implemented by Huang and Ovcharenko [161] and used it to compare models trained either on (A) regulatory elements called based on epigenomic markers as done by Huang and Ovcharenko: (1) silencers: H3K27me3 ChIP-seq peaks not overlapping either H3K27ac, H3k4me1 or H3k4me4 ChIP-seq peaks, (2) enhancers: ATAC-seq peaks overlapping H3K27ac ChIP-seq peaks, and (3) nonfunctional: ATAC-seq peaks from five other cell types not detected in GM12878; or (B) regulatory elements experimentally identified by the ATAC-STARR-Seq assay (as described above). We measured the performance of the two models in terms of AUPRC of detection experimentally identified elements. In both approaches, only sequences (without any epigenetic signal) were provided to the CNN model as input (as done in Huang and Ovcharenko).

Feature importance scores using integrated gradients

To determine which features contribute the most to correct classification we used the integrated gradients (IG) approach [89]. The main idea behind this approach is to find the contribution of input features to the prediction by calculating the integral of the model's output gradients over a straight path between a chosen 'proper baseline' input and the actual input. To do so, 50 points are sampled along the path and the output gradients are calculated for each point. Accumulating the gradients from all points defines the integrated gradients, which are used as the feature importance score. We chose a proper baseline input as follows:

$$baseline_{i,j} = \begin{cases} 0 & j \neq Methylation \\ -1 & j = Methylation \end{cases}$$

Where $1 \leq i \leq 1000$ and j is the feature type: A, C, G, T, Methylation, H3K27ac or H3K4me1. This baseline corresponds to a sequence with 0 (or NA) signal for all seven features. After computing the integrated gradients per position i and feature j, we summed them up across all positions to represent the integrated gradients of feature j. The feature importance score of each feature is the average of the integrated gradients across all inputs per class.

Identification of motifs within importance scores

We used TF-MoDISco to find recurrent motifs within subsequences with highly positive or negative importance scores [166]. The tool first finds subsequences of high importance scores, aligns and clusters them, and then finds a set of recurring motif patterns. We computed the IG of the enhancer and silencer sequences in the test set. To account only for changes in the nucleotide composition we kept the epigenetic features fixed along the path between the baseline and the input. The null IG distribution used in TF-MoDISco was generated by dinucleotide shuffling the original sequences and computing their IG.

Motif finding

We applied the simple enrichment analysis (SEA) tool from the MEME suite (https://meme-suite.org/meme/tools/sea) on the inferred sequences with Human HOCOMOCO v11 PWMs [178]. A Markov model of order of 1 was chosen to model the background distribution.

ChIP-seq enrichment

We downloaded all 158 ENCODE ChIP-seq narrowpeak bed files of GM12878 cell type. For each peak file, we represented each peak by the single nucleotide position that had maximum ChIP-seq signal. Then, we computed the number of these positions that overlapped with 1kb predicted enhancers, silencers and nonfunctional sequences from the exploratory dataset. We computed the enrichment fold-change as follows:

$$FC(protein) = \frac{\frac{\#overlaps\ in\ target\ set}{|target\ set|}}{\frac{\#overlaps\ in\ bg\ set}{|bg\ set|}}$$

Where the target set is either the enhancers or silencers. The bg set included all sequences in the exploratory set. We used the Hypergeometric test (python 'scipy.stats.hypergeom') to evaluate the significance of the enrichment. Benjamini-Hochberg multiple testing correction was used to correct the p-values [67].

eQTL and GWAS risk SNPs enrichment

GWAS summary statistics of 50 traits were downloaded from the GWAS catalog (https://www.ebi.ac.uk/gwas/) and preprocessed as described in Groenewoud et al. [168]. For each trait, we retained associated variants with p-value < 1E-7. Then, similar to ChIP-seq enrichment above, we computed the overlap of the risk SNPs with the predicted enhancers and silencers from the exploratory dataset and computed the significance of the enrichment using HG test.

As for eQTL enrichment analysis, we downloaded the lymphoblastoid cell line (LCL) GEUVADIS eQTL dataset (http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/GEUVADIS/ge/) and computed the overlap of the eQTLs with the predicted enhancers, silencers and nonfunctional sequences in the exploratory dataset. To find whether eQTLs tend to overlap enhancers more than the nonfunctional sequences we implemented a logistic regression test: $\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ where: Y_i denotes whether sequence i has an eQTL or not, X_1^i denotes whether sequence i is an enhancer or a nonfunctional sequence, and X_2^i is the distance from region i mid-position to the nearest gene TSS. We added the distances to the nearest gene as this distance may confound the association between eQTLs and genomic intervals. We used the python statsmodels.sd.Logit function to implement logistic regression and to infer significance of the coefficients. If β_1 is positive and significant then we concluded that the eQTLs are significantly enriched within the set of enhancers. Similar analysis was done for silencers versus the nonfunctional sequences.

5. Discussion

In this thesis we described our contributions to inference of regulatory elements and their linked target genes using thousands of samples covering hundreds of cell types and tissues. We introduced two algorithms for discovery of enhancer-promoter (EP) links. Development of the first algorithm, FOCS, started during my MSc and was continued in my PhD. This tool detects global EP links with significant correlation between enhancer and promoter activity patterns across many samples. In the first stage of my PhD, I significantly expanded FOCS's scope and validation by using 10-fold more experimental data to train on. The second algorithm, CT-FOCS, detects which of the global EP links predicted by FOCS are specific to a few cell types. In both cases, we compared the algorithms to the state of the art and showed a very significant improvement. In addition, we showed how the predicted EP links can be used in downstream applications such as enrichments for GWAS signals and TF motifs.

In the last part of my PhD I turned to an understudied class of regulatory elements, silencers, which reduce the transcription of their target genes. Using DL models and a recently published experimentally identified set of enhancers and silencers in GM12878 cell line, we answered various questions on how to properly predict functional silencers and what defines them epigenetically.

All algorithms and analyses in this thesis were implemented (in R or Python) and are freely available for the community. All analyzed datasets, including the FOCS and CT-FOCS databases, are made available by us as well.

Below we discuss each study separately and suggest future research directions.

5.1. Global and cell type-specific enhancer-promoter inference

5.1.1 The FOCS algorithm

In the last two decades a major challenge has been to detect the target genes of each enhancer. This task is crucial as this allows researchers to study diseases that are a result of irregular gene expression levels caused by perturbations of their enhancers. Such perturbation could be a result of genetic variants altering the enhancer sequences, thus altering their effect on their target genes. While several experimental techniques for identifying physical EP links have emerged (e.g., HiChIP and ChIA-PET described in **Chapter 1**), these techniques were applied only on a few cell types and tissues. In contrast, NGS techniques for identifying open chromatin regions such as enhancers and promoters (.e.g., DNase-seq and ATAC-seq described in **Chapter 1**) were applied to thousands of samples covering hundreds of cell types and tissues.

The datasets that summarize these experiments are publicly available and can be leveraged to identify EP links computationally on a very large scale. Using such datasets, the FOCS algorithm, described in **Chapter 2**, aims for predicting global EP links. These links show significant and high correlation between the enhancer and promoter activity patterns across many samples.

The FOCS link inference is achieved by training for each promoter a regression model where we predict the promoter activity based on the activity of its 10 closest enhancers (**Figure 2.1**). To avoid over-fitting of the regression models to the training samples, FOCS performs a leave-one-cell-type-out cross validation scheme where given C cell types, the regression model is trained on samples belonging to C-1 cell types. The trained model is then used to predict the promoter activity in samples belonging to the left out cell type. This procedure is repeated C times, one for each cell type, resulting with a predicted promoter activity vector. This vector is compared to the observed promoter activity vector by performing two non-parametric statistical tests for testing: (1) how well the model discriminates between samples in which the promoter is active and samples in which it is inactive, and (2) how well the model preserves the original activity ranks in samples whose promoter is active. Using the Spearman correlation test in the second nonparametric test in FOCS can also account for promoter models where the relationship between the enhancer and promoter activity patterns is not linear (as speculated in the majority of FANTOM5 and Roadmap models showing $R^2 < 0.5$; **Supplementary Fig. 2.6B**). Models that passed these tests were regarded as statistically cross-validated. These models then underwent model shrinkage (using elastic net) to select the subset of enhancers that are most informative to the promoter's model. The selected informative enhancers comprise the final global EP maps.

We applied FOCS on four different genomic data sources: one resource of 246 samples generated during my MSc studies, and three additional resources comprising a total of 2,384 samples analyzed during my first year PhD studies. We derived an extensive resource of statistically cross-validated EP links. Our EP mapping resource illuminates different facets of transcriptional regulation. First, we found that ~26% of the enhancers in FOCS EP links were mapped to a promoter that is not the closest one (**Supplemental Fig. 2.10**). This is in contrast to the common naïve approach that maps enhancers to their nearest promoter. Second, we found that 70% of the linked enhancers are located within intronic regions (**Supplemental Table S2.2**). Third, we found that promoters were linked to ~3 enhancers on average, with many of them linked to a single dominant enhancer or to a large number of enhancers (8-10).

Next, we set to explore the relationship between the inferred EP links and housekeeping genes taken from [179]. Housekeeping genes are ubiquitously expressed across different cell

types, and, thus are likely to have a simple regulation logic. Indeed, these genes had a lower number of linked enhancers compared to all other genes (P<0.001 in all data types; **Supplementary Fig. 2.11**). We also calculated the Shannon entropy of each gene promoter activity across cell types (where higher entropy indicates larger activity breadth) to explore a possible relationship the activity breadth and the complexity of transcriptional regulation. We observed a strong negative relationship where promoters with more restricted activity profiles (i.e., lower entropy) are associated with a larger number of linked enhancers (**Supplementary Fig. 2.12**).

Furthermore, our observations indicate that although a significant proportion of enhancers (~90%) in FOCS's models had positive Pearson and Spearman correlation coefficients with their target promoter activity patterns, some of these models also included cases of negative correlation, suggesting that the same regulatory elements sometimes also function as silencers (see two examples in **Supplementary Fig. 2.13**).

Lastly, in this study we implicitly assumed that transcription rate at promoters is positively related with promoter DHS signal (ENCODE and Roadmap Epigenomics datasets). We examined the DHS-expression correlations in 17 cell lines for which both DHS and RNA-seq data were in hand in the ENCODE project. In all cases, we observed high Spearman but low Pearson correlations (**Supplementary Fig. 2.14**), indicating a strong monotonic nonlinear relationship.

We compared the performance of FOCS and three alternative methods for predicting EP links. (1) Pairwise: pairwise Pearson correlation > 0.7 between EP pairs under FDR < 10^{-5}. (2) OLS+LASSO: models derived by ordinary least squares (OLS) using all samples with cross validation (CV) followed by LASSO shrinkage. (3) OLS+enet: same as (2) but with elastic-net shrinkage in place of LASSO. FOCS derived 75% more promoter models than the other methods (**Table 2.1**). In addition, for most comparisons, FOCS outperformed the other methods in terms of fraction of predicted EP links supported by ChIA-PET and HiChIP interactions, and eQTL data (**Fig. 2.4**; **Supplementary Fig. 2.9**).

The FOCS algorithm has several limitations. First, while the leave-cell-type-out CV scheme is conservative and ensures that the inferred models have predictive power in diverse cellular contexts, it will not infer models for genes whose expression is strictly cell type-specific. However, when analyzing many diverse cell types that contain also related cell types, we expect a low chance of missing gene models that are cell type-specific. Second, the FOCS gene models are limited to the ten closest enhancers, as done in previous analysis [23]. Such limitation might miss true linked enhancers located further away from the genes. On the other hand, as most EP links are confined to chromosomal territories called topologically associated

domains (TADs) of 185 kb median length [131], limiting the gene models to the ten closest enhancers might reduce the number of false-positive EP links.

The FOCS broad compendium of predicted EP links can greatly aid the functional interpretation of genetic variants that are associated with disease susceptibility since the majority of the variants (~90%) identified by GWAS are within noncoding regions [107]. Similarly, the compendium can also assist in interpreting somatic mutations (SM) in cancer genomes. The rapid accumulation of whole-genome sequencing (WGS) of tumor samples has already led to swift identification of SM hotspots in regulatory regions [180,181]. In addition, our compendium can also be integrated into bioinformatics tools for identifying DNA motifs in regulatory elements that putatively regulate co-expressed gene clusters.

Overall, we found that FOCS predicts ~1.5-fold more EP links ($n = 302{,}050$) compared to the standard pairwise method. FOCS EP links show a higher support rate by external validation resources compared to the pairwise method, demonstrating an improved prediction power and control of false positive rate. FOCS employs two non-parametric tests for model robustness. These tests enable us to correct for multiple promoter models and to use them in cases where the relationship between the EP activity patterns is not linear. This is in contrast to the Pearson correlation used in the pairwise method that assumes a linear relationship.

5.1.2 The CT-FOCS algorithm

In **Chapter 3** we described CT-FOCS, an algorithm for detecting cell type-specific EP links. It builds on the FOCS algorithm described above, designed for predicting global EP links that are not necessarily cell type-specific. CT-FOCS uses the FOCS EP links and performs additional analyses to infer which of these links are active in only very few cell types among the hundreds of cell types, by utilizing linear mixed effect models (LMMs). We applied CT-FOCS on CAGE and DHS profiles from FANTOM5, ENCODE, and Roadmap Epigenomics consortia [23–25]. The resulting compendium consists of 229,518 cell type-specific EP links (ct-links) covering 651 cell types.

The CT-FOCS algorithm uses LMMs to account for two effects. The first one is the joint contribution of multiple enhancers to the promoter activity, as previously shown for predicting GE more accurately than the simple pairwise enhancer-gene correlations [133]. The second one is the contribution of distinct cell type groups of samples to the promoter activity. By taking into account the influence of each cell type group, we were able to predict promoter activity independently for each cell type group. This way, the calculated regression coefficients are not the same for all samples but are adjusted for the cell type. Using the difference between cell types in the predicted regression coefficients we were able to infer ct-links.

The leave-cell-type-out CV limitation described in the previous section, where FOCS cannot infer models that are strictly cell type specific, i.e., functional in exactly one cell type, is also true for CT-FOCS, as it is built upon FOCS predictions. However, we found that cases in which an enhancer is active in only one cell type are very rare (see **Supplementary Results** in section 5.2 – 'Loops involving enhancers active in a single cell type' and **Supplementary Table S5**). In addition, the CT-FOCS ct-links demonstrated high cell type specificity: the ct-links were shared, on average, by approximately three cell types (**Supplementary Fig. 3.2B**) with ~44% of them called in a single cell type. Thus, the predicted ct-links correspond to cases in which an enhancer shows activity in several highly related cell types, but, its impact on the target promoter activity is limited to very few of them.

Another limitation of CT-FOCS is the requirement that there will be at least two replicates per cell type in order to allow prediction of their random effect variance in the LMM model. Cell types with a single replicate are also included in our models, as they contribute to estimating the fixed effect coefficients. In FANTOM5 dataset, 179 out of 472 cell types had at least two replicates. When we applied CT-FOCS only on these 179 cell types, the TLS ratio performance improved (see **Supplementary Results** in section 5.2; **Supplementary Fig. 3.13**). Thus, we recommended to use these predictions in case their cell type of interest is among the 179 cell types analyzed.

We compared CT-FOCS with two alternative prediction methods, CT-MAD-FOCS and CT-JEME (**Section 3.1 'Methods'**), based on a cell type specificity score computed either on the EP signals or on the target GE levels (see **Supplementary Fig. 3.6** on FANTOM5 dataset; **Supplementary Table 1B** on ENCODE dataset; **Section 3.1 'Methods'**). On the FANTOM5 dataset, CT-FOCS achieved slightly better cell type specificity ranks compared to other methods on both EP signals and target GE (**Supplementary Fig. 3.6**). In addition, we introduced the notion of two step connected loop set (TLS) support ratio for benchmarking predicted ct-links against chromatin interaction datasets (**Figure 3.4**; **Supplementary Table 2**; **Section 3.1 'Methods'**). By using this measure, we found that the cell type particularity of the CT-FOCS ct-links was significantly higher than those of the other methods in five to six out of eight examined cell types for which chromatin interaction data was available.

CT-FOCS holds the potential to address downstream genomic analyses. It can be used to improve our understanding of the cell type-specific transcriptional cascades that determine cell fate decisions. For example, identification of known and novel cell type-specific TFs that mediate ct-links can expand our current biological knowledge on cell type-specific gene regulatory programs (as shown in **Fig. 3.5** and **Supplementary Fig. 3.9**). Integration of protein-protein interactions (PPIs) with TF identification in ct-links may help identify cell type-specific

PPI modules [182]. Such PPI modules may contain new proteins (e.g., cofactors and proteins that are part of the mediator complex) that establish the cell type-specific 3D chromatin structure.

Overall, CT-FOCS identified ct-links for 651 cell types inferred from ENCODE, Roadmap Epigenomics, and FANTOM5 data. On average, ~354 ct-links were discovered per cell type. The inferred ct-links showed substantially higher cell type-specificity compared to previous methods. The inferred ct-links correlate with cell type-specific gene expression and regulation. We validated predicted ct-links with experimental 3D chromatin interactions by using the notion of connected loops.

5.2. Silencer inference

In **Chapter 4** we described our study for the inference of an understudied class of regulatory elements, silencers, which decrease the transcription of their target genes. In contrast to enhancers, silencers are harder to validate experimentally. There is still no consensus on how to identify silencers. For example, two recent studies applied different genomic screening techniques and identified 2,664 and 3,001 silencers in K562 cell line [34,183]. Strikingly, there is no overlap between these two sets. Furthermore, these candidate regions may contain sub-regions that are interchangeably activating and repressing [35], making their detection even harder. Thus, a robust characterization and annotation of functional silencers is a major genomic challenge.

Our first goal in this study was to test if a DL model trained on putative silencers labeled using epigenomic data can accurately detect experimentally identified silencers. To this end, we compared two class labeling approaches: the epigenetic approach, in which putative enhancers and silencers are labeled using epigenetic data, and the experimentally identified approach, in which enhancers and silencers are labeled using elements experimentally identified by ATAC-STARR-Seq assay [35]. We trained a CNN model proposed by Huang and Ovcharenko [161] on each dataset and tested the performance of the models on experimentally identified test set. While both trained CNN models performed similarly on predicting true enhancers (0.3 and 0.37 AUPRC for the models trained using the epigenetic and the experimental approaches, respectively), the silencer prediction performance of the model trained on the experimental dataset was dramatically higher than that obtained by the model trained on the epigenetic dataset (0.77 vs. 0.06 AUPRC, respectively; **Supplementary Fig. 4.1**). These results reflect the much better knowledge that we currently have on epigenomic marks defining active enhancers compared to those defining active silencers.

Our second goal was to build a computational model that predicts activation and repression transcriptional activities at single nucleotide resolution within regulatory elements.

To this end, we used the ATAC-STARR-Seq experimental results to train a regression-based DL model combined with a classification model to classify sequences into enhancer, silencer or nonfunctional elements. We compared several published DL architectures and found that the deepTACT model [45] performed best in terms of AUROC and AUPRC (**Fig. 4.1c**). Predicted silencers harbored high levels of the H3K27me3 repressive mark, whereas predicted enhancers harbored high levels of H3K27ac and H3K4me1 activation marks (**Fig. 4.2**; **Supplementary Fig. 4.3**).

We applied the trained deepTACT model on an exploratory dataset, which included ATAC-seq peaks in GM12878 that were not detected by the ATAC-STARR-Seq assay as having an effect on transcription. Within this set, the model identified 3,752 novel putative enhancers and 518 novel putative silencers, which were possibly missed by the experiment. Reassuringly, 18 of the predicted novel enhancers overlapped 42 Lupus risk SNPs, including rs13240595 Lupus risk-SNP, which was shown to have 2.5-fold enhancing effect by MPRA analysis [171]. We showed that predicted enhancer sequences contain significantly more eQTLs than predicted nonfunctional sequences. ChIP-seq enrichment analysis within predicted novel silencers identified that they are enriched for the binding sites of four major transcriptional regulators: SUZ12, HDAC6, EZH2 and NRF1. SUZ12 and EZH2 form the PRC2 repressive complex known to bind silencers (**Fig. 4.5a**). Predicted enhancers, on the other hand, were enriched for many proteins, the majority of which are known to induce transcription (**Fig. 4.5b**).

Our study is limited by the fact that it was performed on a single cell type for which genome-wide experimentally identified enhancers and silencers are available. Additional validation would necessitate experiments in more cell types. A major future challenge is to transfer the model trained on GM12878 cell type to other cell types in which activation and repression levels are not in hand (see **Future Research** section below for a proof of concept).

Overall, in this study we presented how a regression-based DL model can predict per-nucleotide activation and repression activities within candidate sequences. Using this model we predicted many additional enhancers and silencers possibly missed by the ATAC-STARR-seq experiment, and expanded the current biological knowledge of what defines functional silencers. In addition, we found that computational models trained on enhancers and silencer sequences labeled using epigenetic data generally perform poorly in predicting silencers. Leveraging data from experimentally identified enhancers and silencers substantially improved silencer prediction accuracy.

5.3. Future Research

The studies in this thesis can be used as a basis for additional research directions. Above, we already outlined some improvements such as downstream analyses that can be applied on the data produced by our models. Below we list additional research directions.

5.3.1. Enhancer-Promoter inference

Application to single cell ATAC-seq data

The bulk NGS datasets processed and analyzed in FOCS and CT-FOCS algorithms contains average signals from a population of cells. These datasets do not provide measurements at the single cell level. In the last decade, single cell (sc) NGS technologies have become an exciting area of research as they provide a higher resolution of cellular differences between individual cells [184]. Single-cell RNA or epigenetics techniques can reveal cell-to-cell variabilities and shed light on how closely related cells differ from each other [185]. We expect that as the technology prices decrease, single cell open chromatin datasets, e.g., scATAC-seq [186], which allow the identification of enhancer and promoter regions at the single cell level, will become more prominent and allow more accurate identification of cell type-specific EP links.

Data from a single-cell experiment is usually presented as a count matrix where rows are individual cells and columns are genomic features. For example, in a scATAC-seq experiment, the count matrix contains number of reads per individual cell per open chromatin regions. In a 10x Genomics scATAC-seq experiment it is possible to target up to 10,000 individual cells. These cells, dissociated from a tissue, are slightly different from each other, and predicting global and cell type-specific EP links for them, in order to study their lineage specification, is still an open biological challenge.

Pliner et al. developed the Cicero algorithm to predict global EP links from a scATAC-seq data [187]. Cicero calculates pairwise correlations between all pairs of sites within 500 kb. However, we have seen in our FOCS study that pairwise correlations are prone to high false-positive rate, due to outliers and the combinatorial nature of transcriptional regulation in which a promoter is regulated by multiple enhancers. Such situations are not addressed by the pairwise approach. As an alternative, one could apply the FOCS algorithm on scATAC-seq data to derive global EP links and then compare the predication performance against Cicero.

Another future direction of research is to study how EP link signals change between the cells in a scATAC-seq experiment. Using these signals, one can reconstruct trajectories of the cells to reveal their lineage specification that are affected by cell-specific 3D structure changes, e.g., by using the STREAM method [188]. To this end, applying the CT-FOCS

algorithm on scATAC-seq data can predict cell type-specific EP links. By using the random effect component from our CT-FOCS LMM models, one might be able to infer cell type-specific EP signals to be used in trajectory reconstruction.

5.3.2. Silencer inference

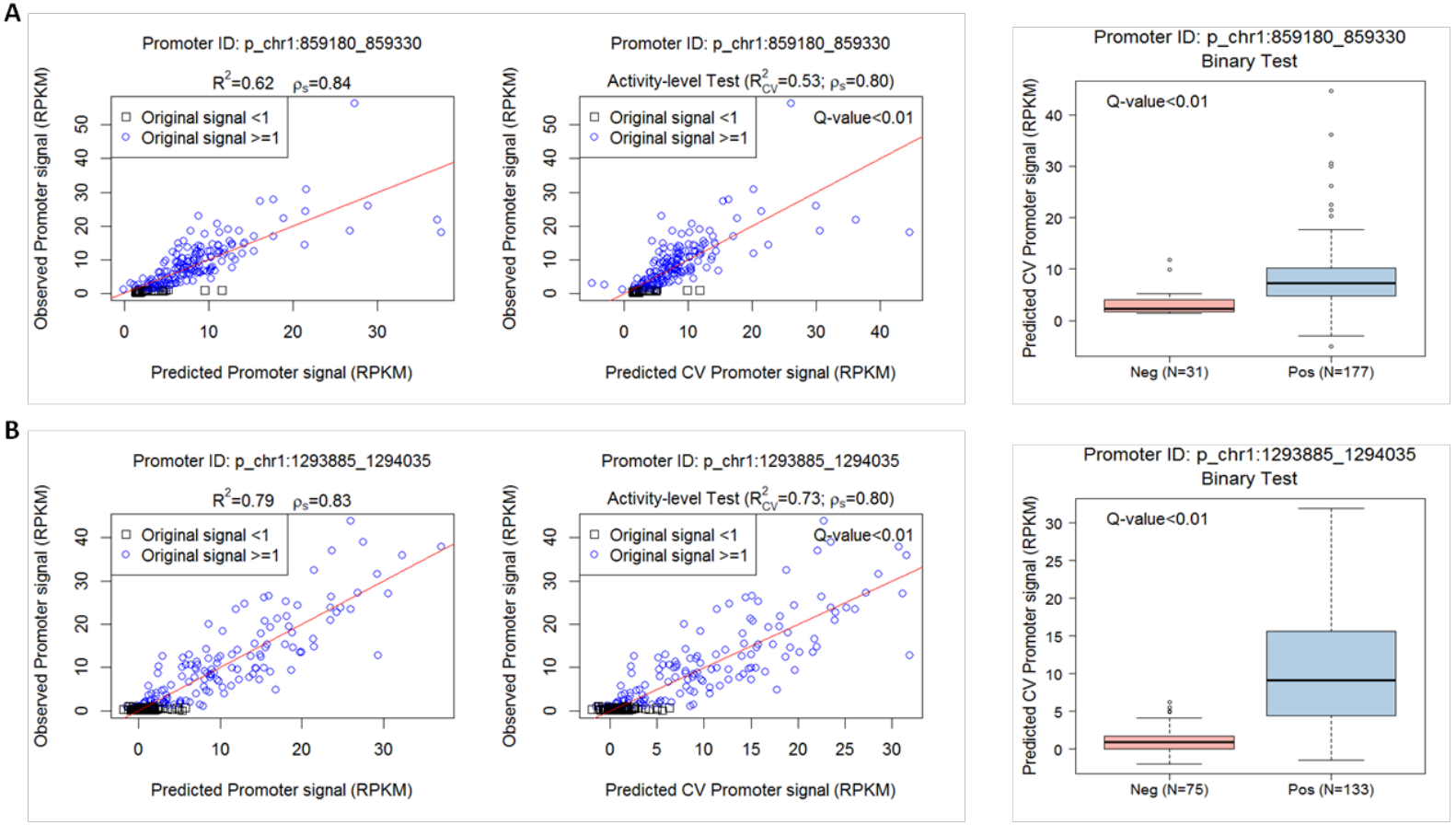
In our third study we used a DL method built on data of thousands of experimentally identified enhancers and silencers in the GM12878 cell line to predict novel REs. However, this study is limited as it cannot infer REs for other cell types where experimentally identified REs are not in hand.

Transfer learning refers to a machine learning approach where knowledge acquired from one task is repurposed to enhance the performance on a closely related task [189]. Finding a way to transfer a model learnt on GM12878 cell line to other cell lines is a desirable future research direction. As proof of concept, we retrained only the last two dense layers in the classification step of the deepTACT model on HepG2 and K562 cancer cell lines (**Fig. 4.1a**). We constructed training and test sets for these cell lines using (a) enhancers detected by STARR-seq experiments done by ENCODE in these cell lines, and (b) ATAC-seq regions overlapping H3K27me3 ChIP-seq peaks as putative silencers. Our results on the test sets from HepG2 and K562 achieved high AUROC and moderate AUPR scores for enhancer and silencer classifications (**Supplementary Fig. 4.8**). This analysis indicates a great promise (and a nontrivial challenge) in the application of transfer learning techniques for predicting REs in many cell types.

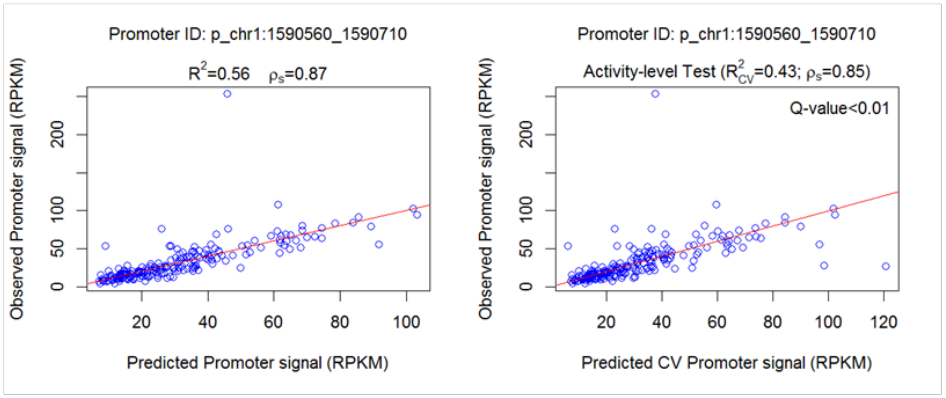
6. Appendix

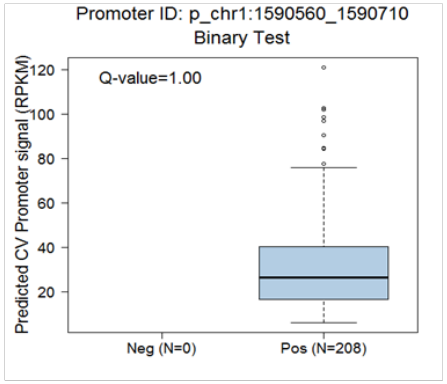
6.1. Supplement 1: Predicting global enhancer-promoter maps

Supplemental Figures

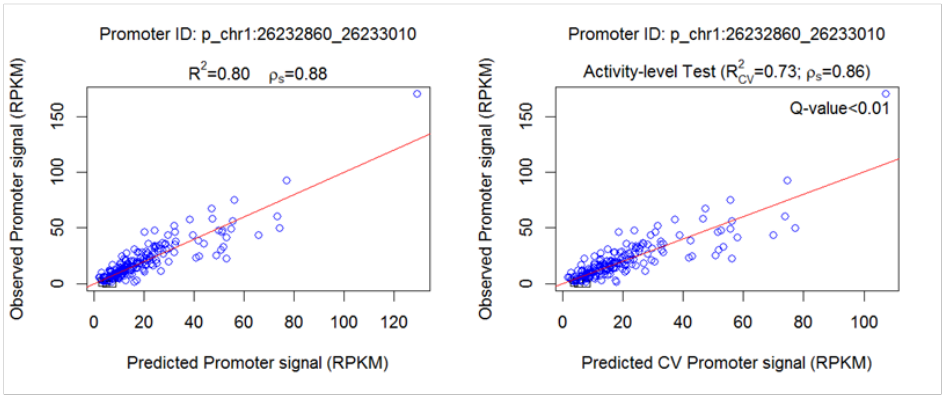


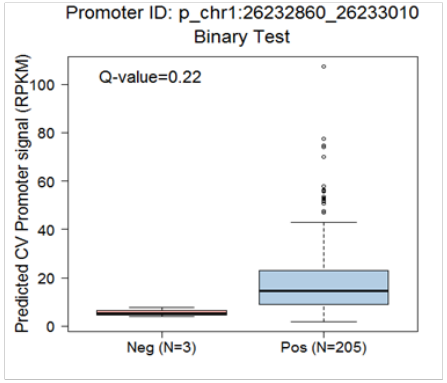
C


Promoter ID: p_chr1:1590560_1590710
R^2=0.56 ρ_s=0.87
Observed Promoter signal (RPKM)
Predicted Promoter signal (RPKM)
Promoter ID: p_chr1:1590560_1590710
Activity-level Test (R^2_{CV}=0.43; ρ_s=0.85)
Q-value<0.01
Observed Promoter signal (RPKM)
Predicted CV Promoter signal (RPKM)


Promoter ID: p_chr1:1590560_1590710
Binary Test
Q-value=1.00
Predicted CV Promoter signal (RPKM)
Neg (N=0)
Pos (N=208)

D


Promoter ID: p_chr1:26232860_26233010
R^2=0.80 ρ_s=0.88
Observed Promoter signal (RPKM)
Predicted Promoter signal (RPKM)
Promoter ID: p_chr1:26232860_26233010
Activity-level Test (R^2_{CV}=0.73; ρ_s=0.86)
Q-value<0.01
Observed Promoter signal (RPKM)
Predicted CV Promoter signal (RPKM)


Promoter ID: p_chr1:26232860_26233010
Binary Test
Q-value=0.22
Predicted CV Promoter signal (RPKM)
Neg (N=3)
Pos (N=205)

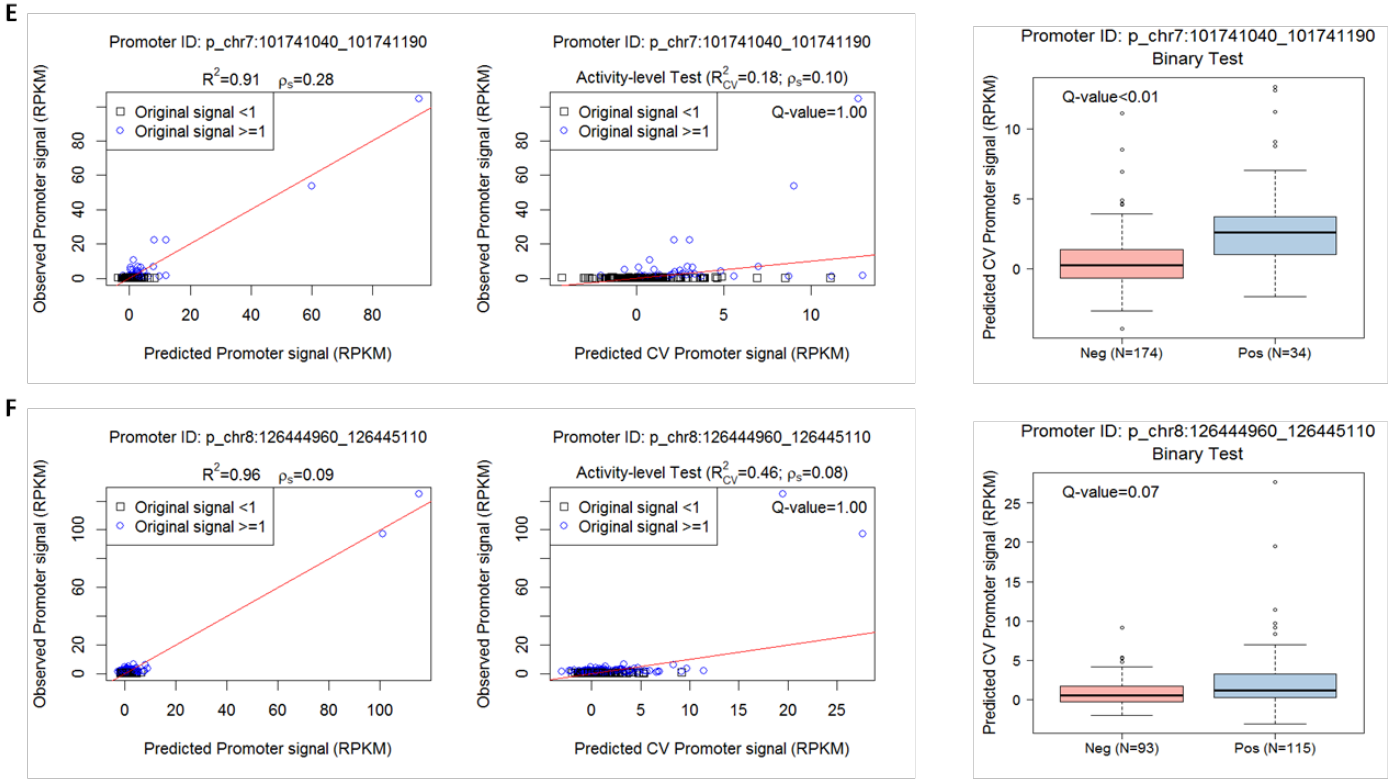


Figure S2.1. Examples of cross-validated promoter models. Examples of promoter models that passed one or both cross-validation tests: (A-B) passed both binary and level tests (C-D) passed only the activity level test and (E-F) passed only the binary test. For each promoter, the left panel shows the correlation between observed and predicted promoter activities using OLS without cross-validation; the middle panel shows the results of the activity level validation test. Namely, the correlation between observed activities and activities that were predicted on left-out samples (LCTO CV procedure). In this test, correlation is calculated only over positive samples. The right panel shows the results of the binary test. Note in E and F left panel, the sensitivity of R^2 (and, equally, of Pearson correlation) to outliers.

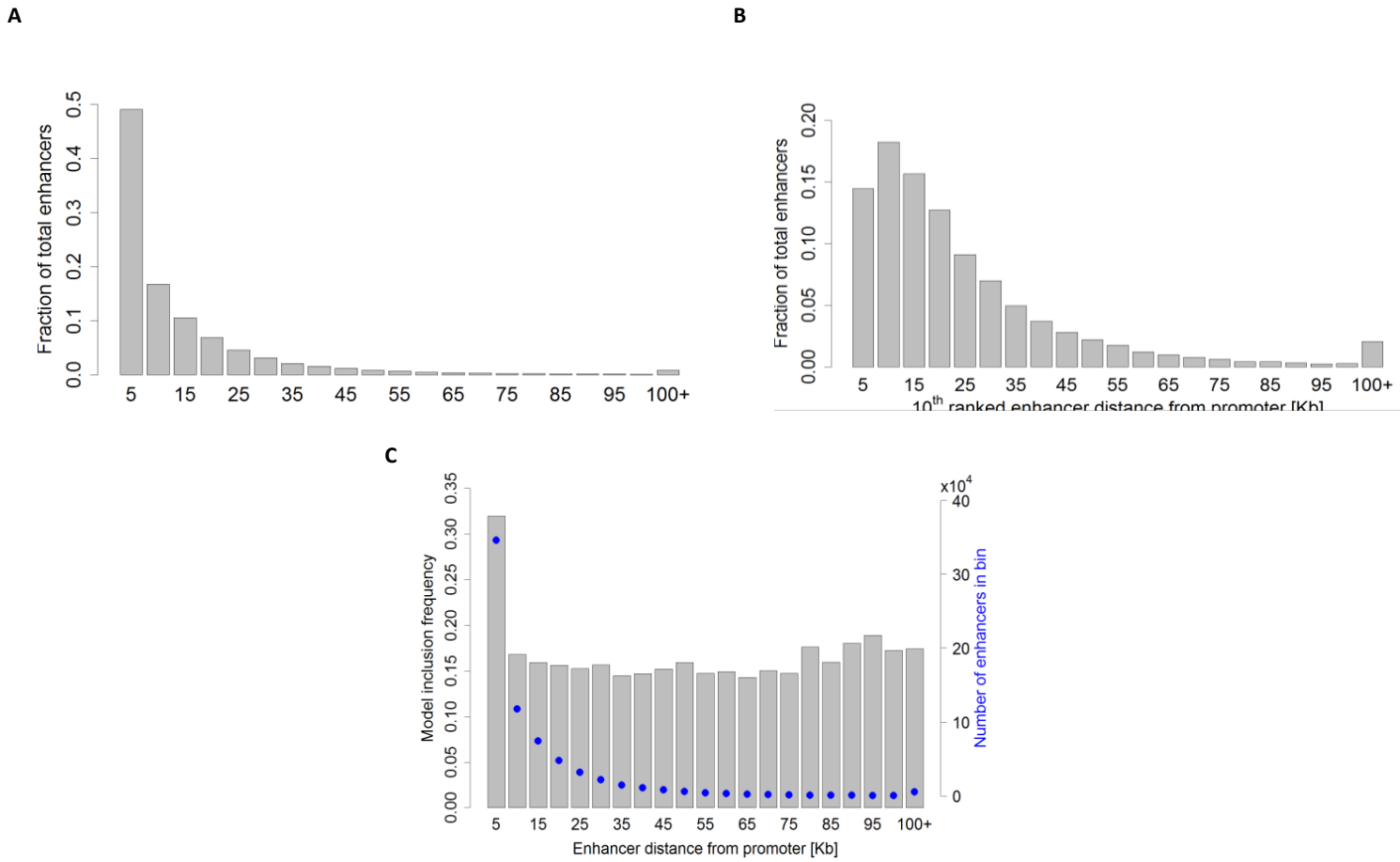


Figure S2.2. EP distance distribution. EP distance distribution for: (A). All 10 enhancers in the models that passed cross validation. (B). The 10^{th} enhancer (ranked by distance to promoter) in the models that passed cross validation. (C). Enhancer inclusion frequency in the optimally reduced models. Blue dots denote the total number of enhancers (right y-axis) in each distance bin before the shrinkage step.

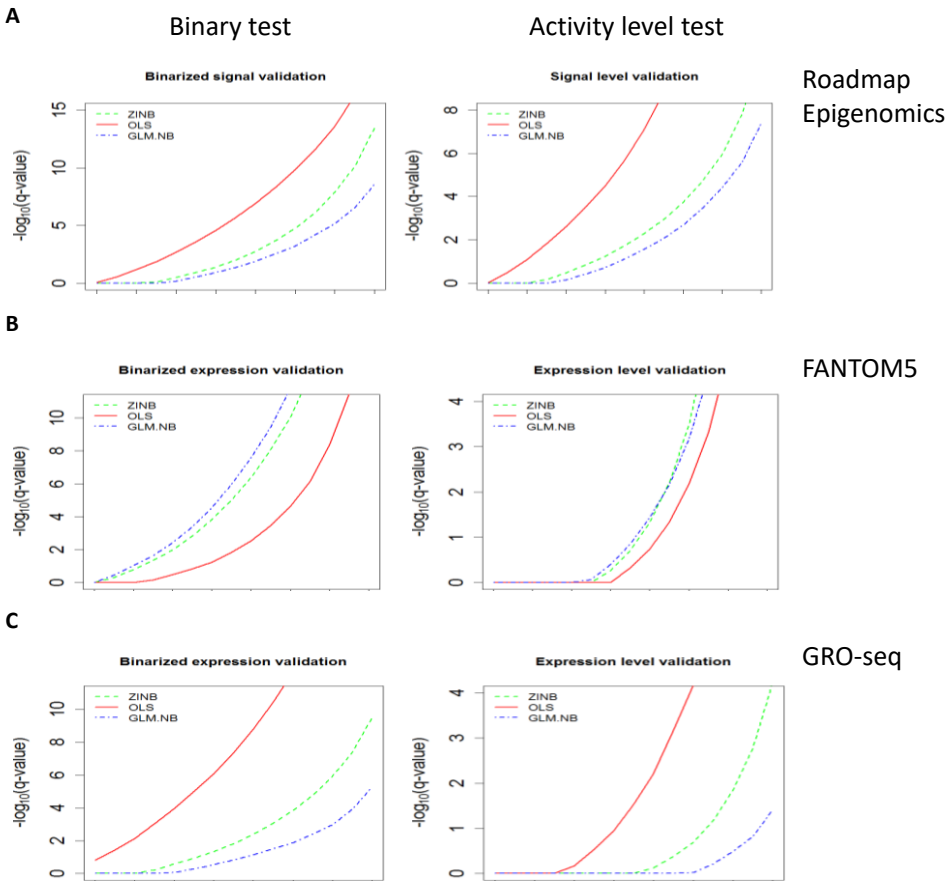


Figure S2.3. Performance of three alternative regression methods for inferring EP models. Same as Figure 2A-B, but here analysis was applied to Roadmap Epigenomics (A), FANTOM5 (B) and the GRO-seq (C) datasets. Results of the binary (left panel) and activity level (right panel) validation tests are shown. OLS performed better on the Roadmap Epigenomics and GRO-seq datasets (in addition to the ENCODE data (Fig. 2A-B)), while GLM.NB and ZINB performed better on the FANTOM5 dataset.

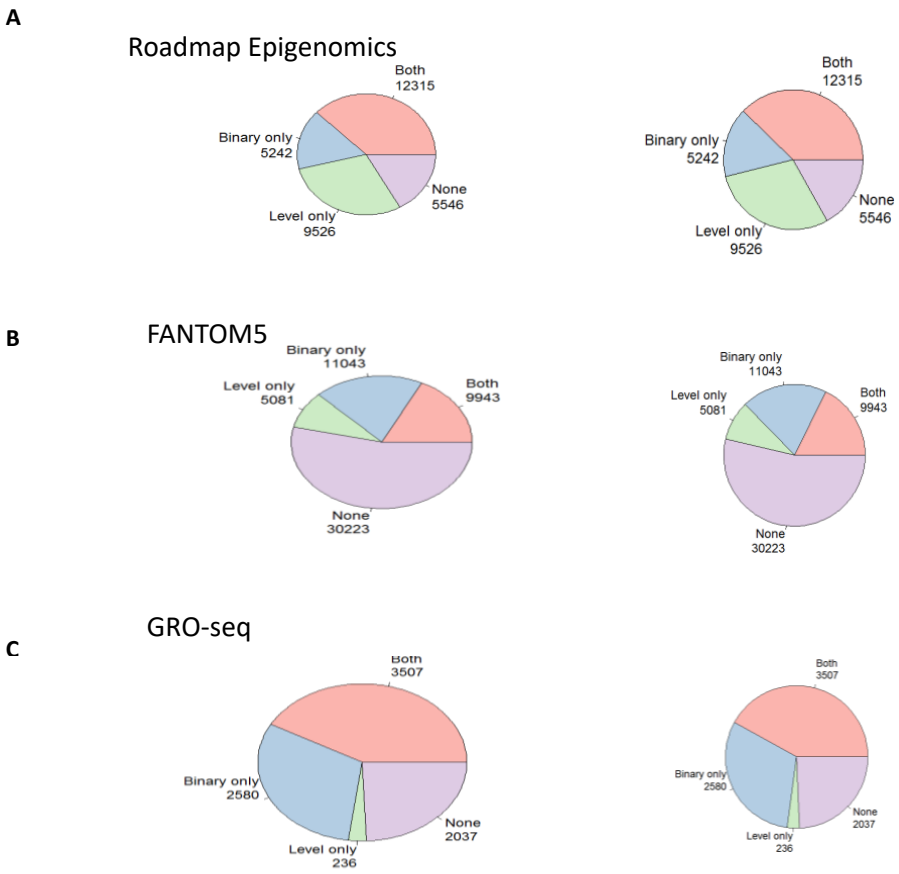


Figure S2.4. Number of validated promoter models. Number of promoters whose OLS models passed (at q-value<0.1) each of the validation tests (right panel) and the distribution of the number of positive samples in each category. (A). Roadmap Epigenomics; (B) FANTOM5 and (C) GRO-seq datasets.

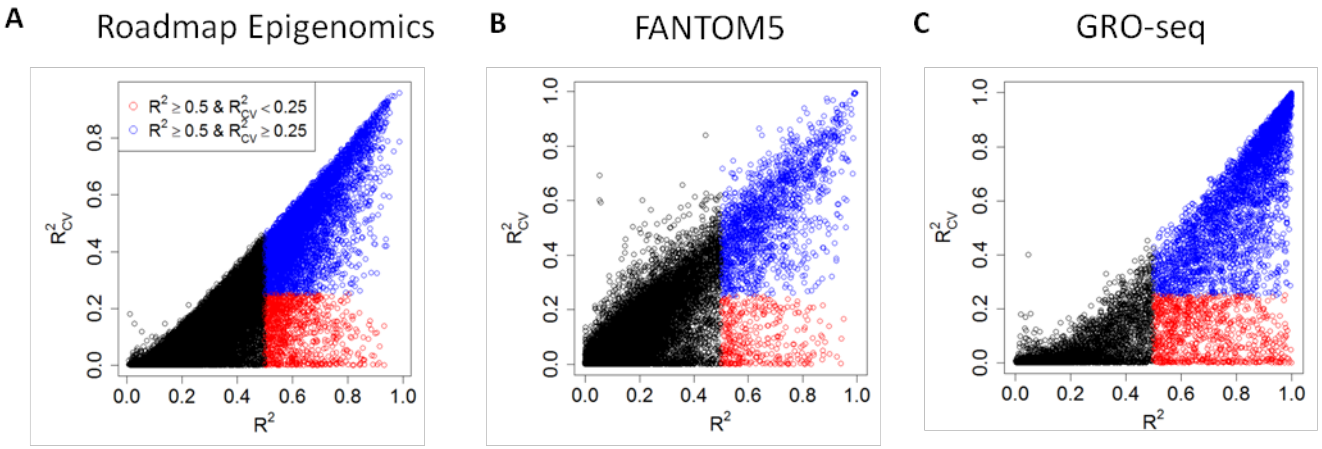


Figure S2.5. **Comparison between the R^2 values with and without cross-validation (CV)**. (A). Roadmap Epigenomics; (B) FANTOM5 and (C) GRO-seq datasets. Each dot is a promoter model. Blue dots denote models with $R^2 \geq 0.5$ and $R^2_{CV} \geq 0.25$. Red dots denote models with and $R^2 > 0.5$ and $R^2_{CV} < 0.25$. The high rate of red dots (Roadmap (16%), FANTOM5 (20%) and GRO-seq (22%)) indicates that training the models on all samples suffer from overfitting.

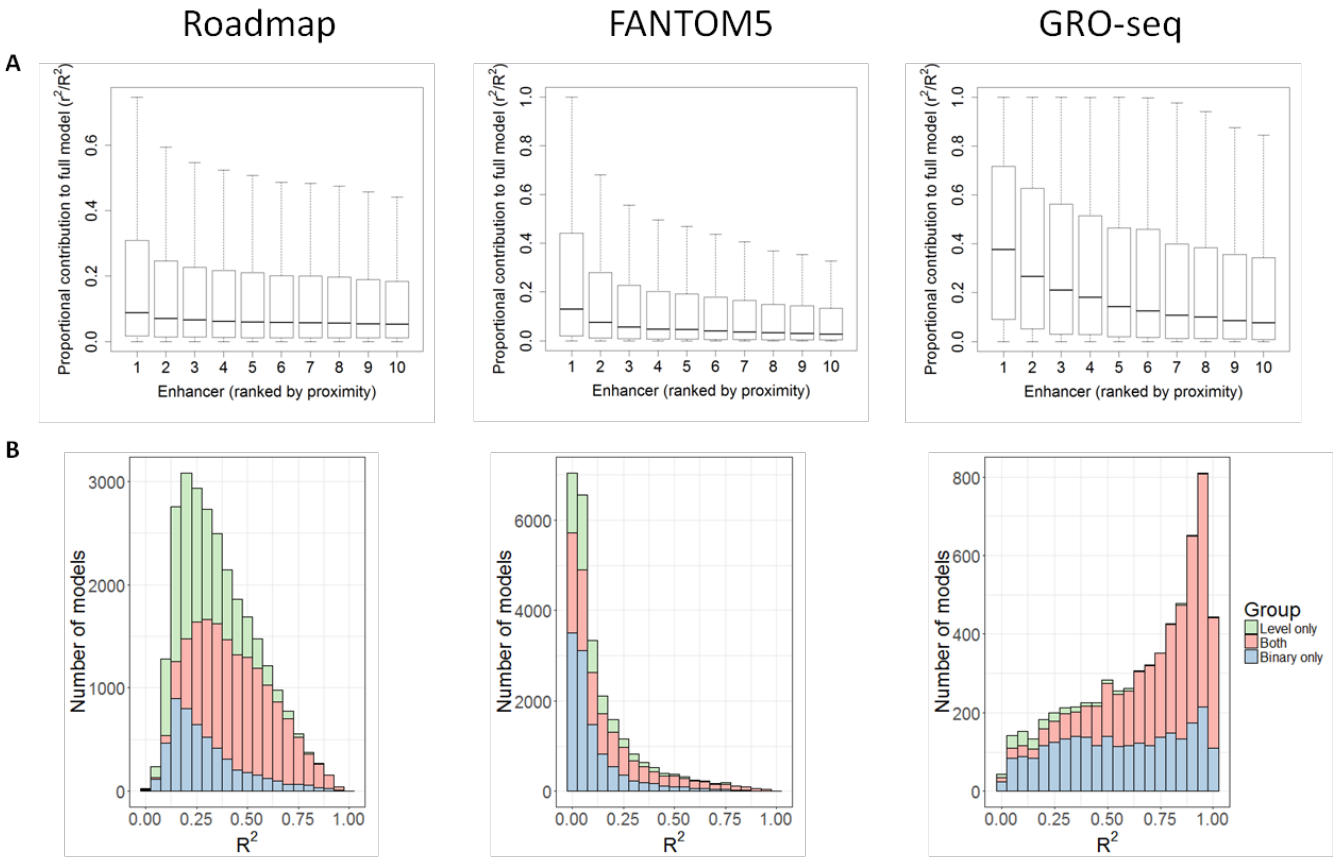


Figure S2.6. Configuration of promoter regulation by enhancers. (A). The proportional contribution of the 10 most proximal enhancers (within a distance of ±500kb from the target promoter; for FANTOM5 the distance was ±250kb from the target promoter) to the regression model, in each dataset (Roadmap Epigenomics, FANTOM5 and GRO-seq). The X axis indicates the order of the enhancers by their relative distance from the promoter, with 1 being the closest. (B) R^2 values of the models that passed one or both CV tests, in each dataset.

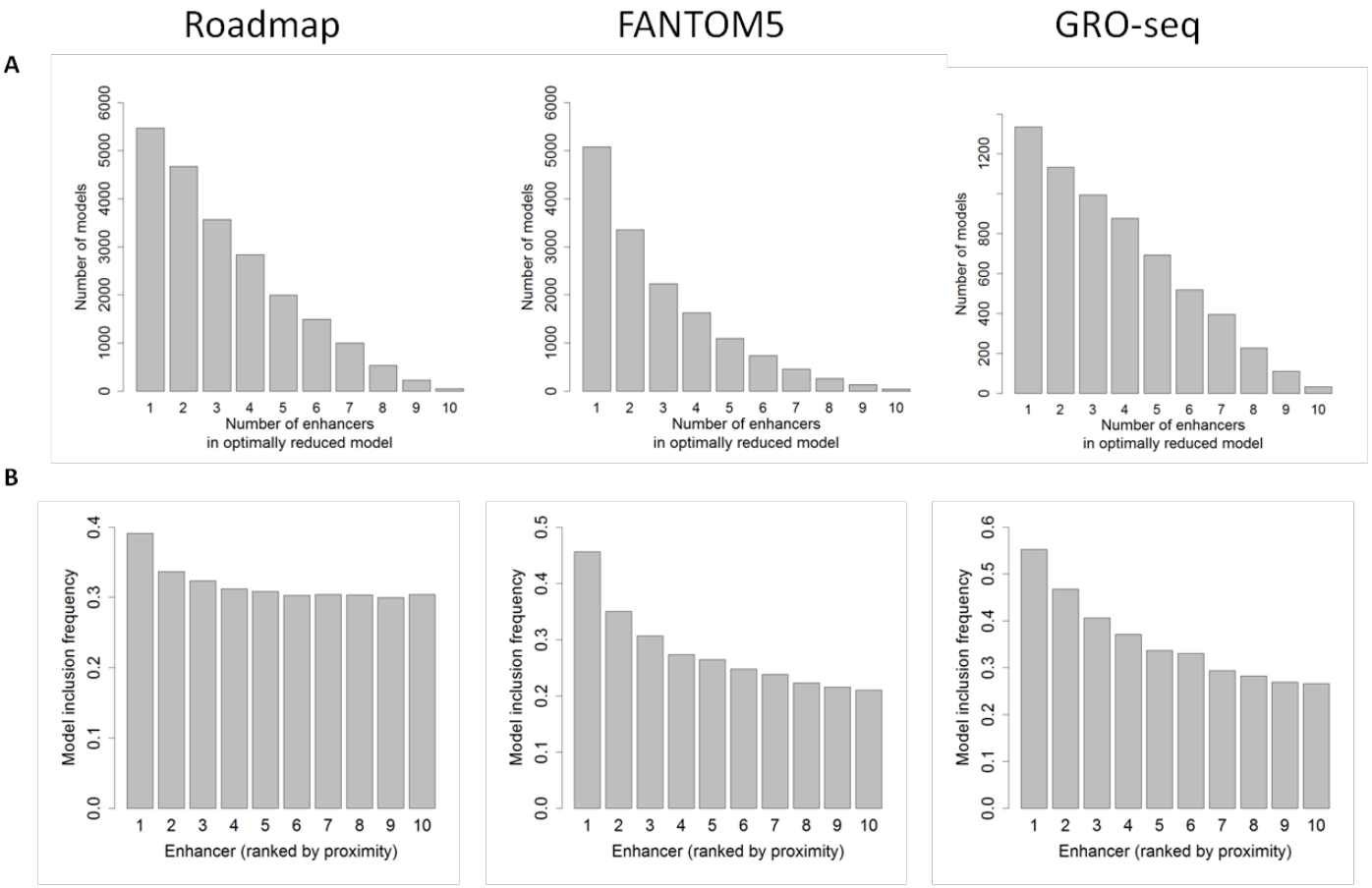


Figure S2.7. Configuration of shrunken promoter models. (A) Distribution of the number of enhancers included in the validated, optimally-reduced models (i.e. after elastic net shrinkage). (B) Inclusion frequency of enhancers in the reduced models as a function of their proximity ranking to the target promoter.

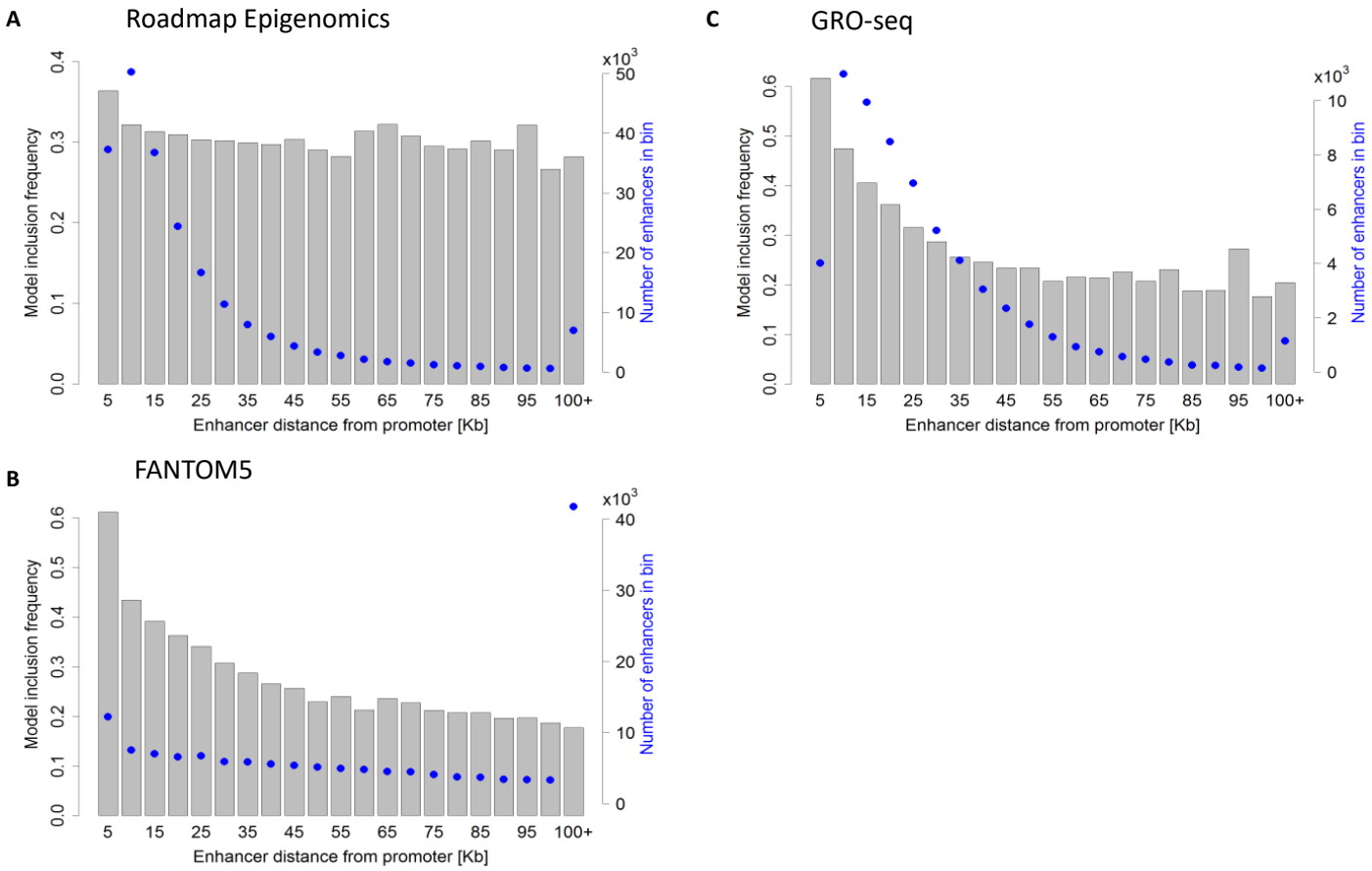


Figure S2.8. Inclusion frequency of enhancers as function of EP distance. Inclusion frequency of enhancers in the reduced models as a function of their distance from the target promoter for (A) Roadmap Epigenomics, (B) FANTOM5 and (C) GRO-seq datasets. Blue dots denote the number of enhancers (right y-axis) in each bin before the shrinkage step.

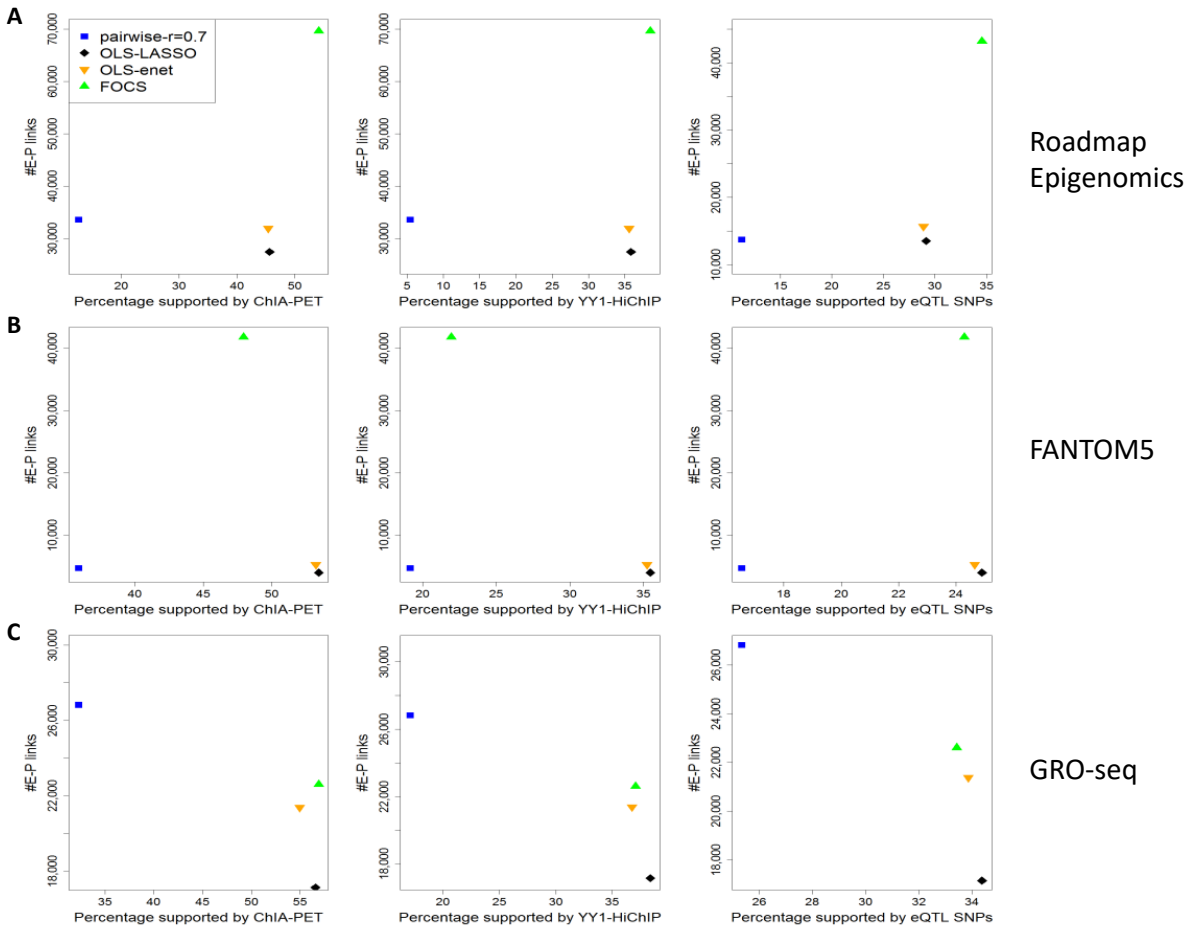


Figure S2.9. Comparison of the performance of different methods for predicting EP links using ChIA-PET, YY1-HiChIP and eQTL data as external validation. As in Fig. 2.4, but for Roadmap Epigenomics (A), FANTOM5 (B) and GRO-seq (C) datasets.

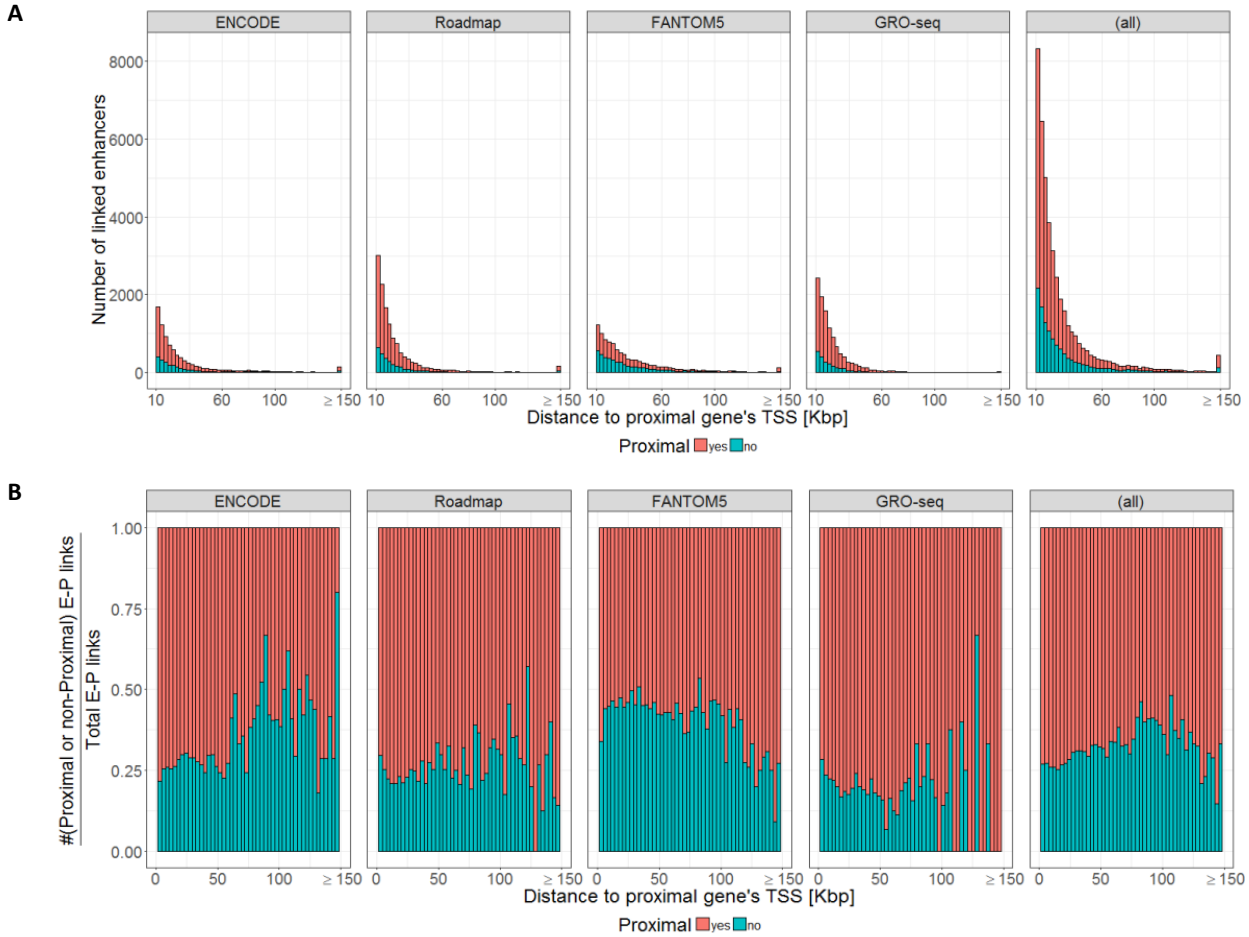


Figure S2.10. Enhancers are frequently linked to genes more distal to the nearest one. The number (A) and proportion (B) of enhancers that are linked to nearest/more distal promoter as a function of their distance to the nearest promoter.

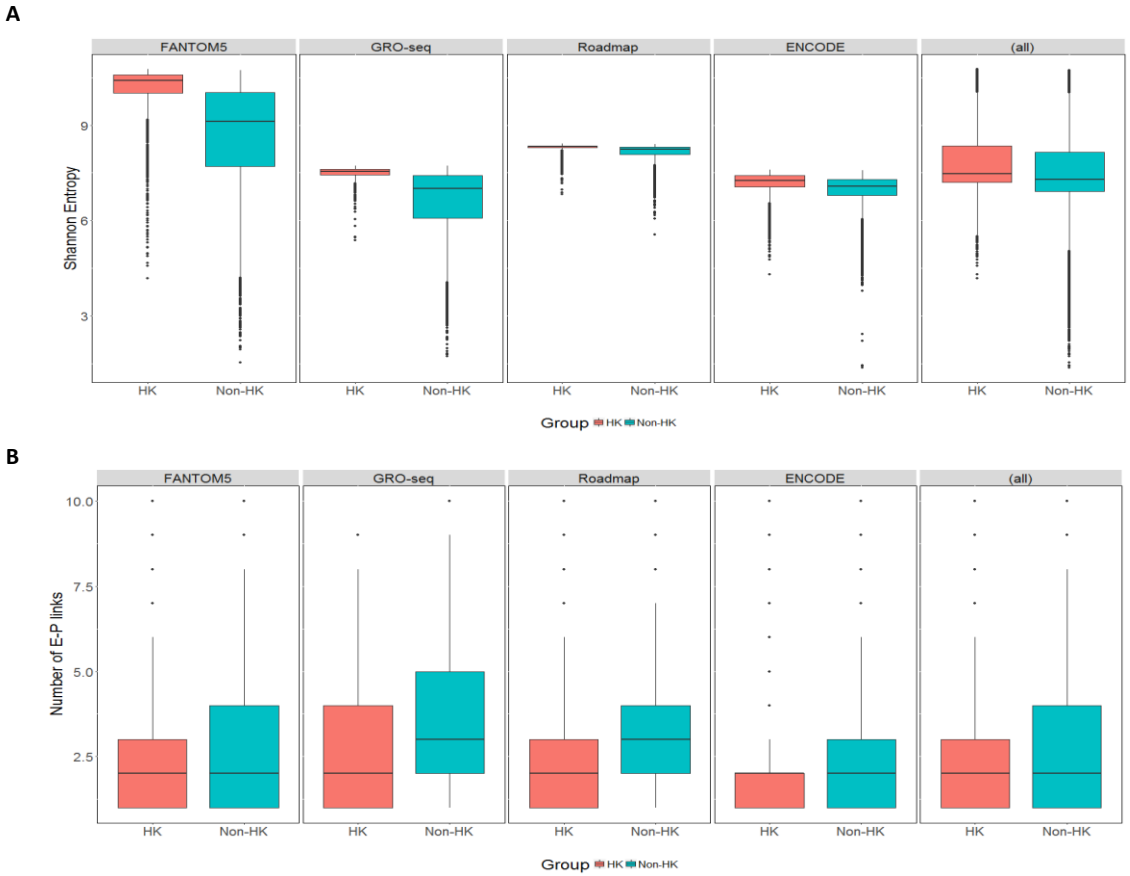
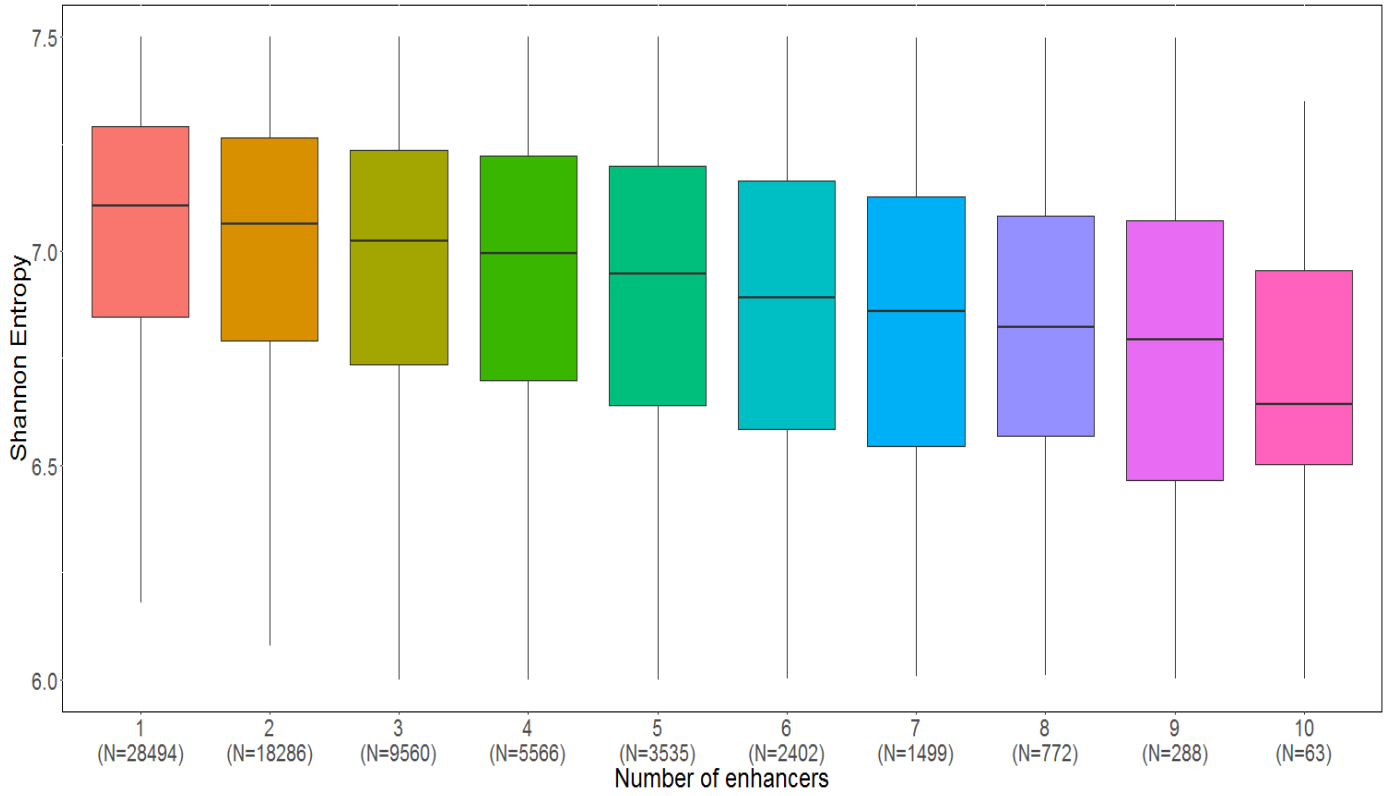


Figure S2.11. Housekeeping genes show simpler pattern of EP interactions. (A). Ubiquitous vs. cell-type specific expression pattern is quantified by Shannon Entropy. In all datasets, housekeeping (HK) genes show significantly higher Shannon Entropy than the rest of genes, reflecting their more uniform activity pattern over the examined cell panel. (B). Promoters of HK genes are involved in significantly lower number of EP interactions than other genes (in all cases, p-value << 0.001; calculated by one-sided Wilcoxon rank-sum test).

a


7.5
7.0
6.5
6.0
Shannon Entropy
1 (N=28494)
2 (N=18286)
3 (N=9560)
4 (N=5566)
5 (N=3535)
6 (N=2402)
7 (N=1499)
8 (N=772)
9 (N=288)
10 (N=63)
Number of enhancers

b

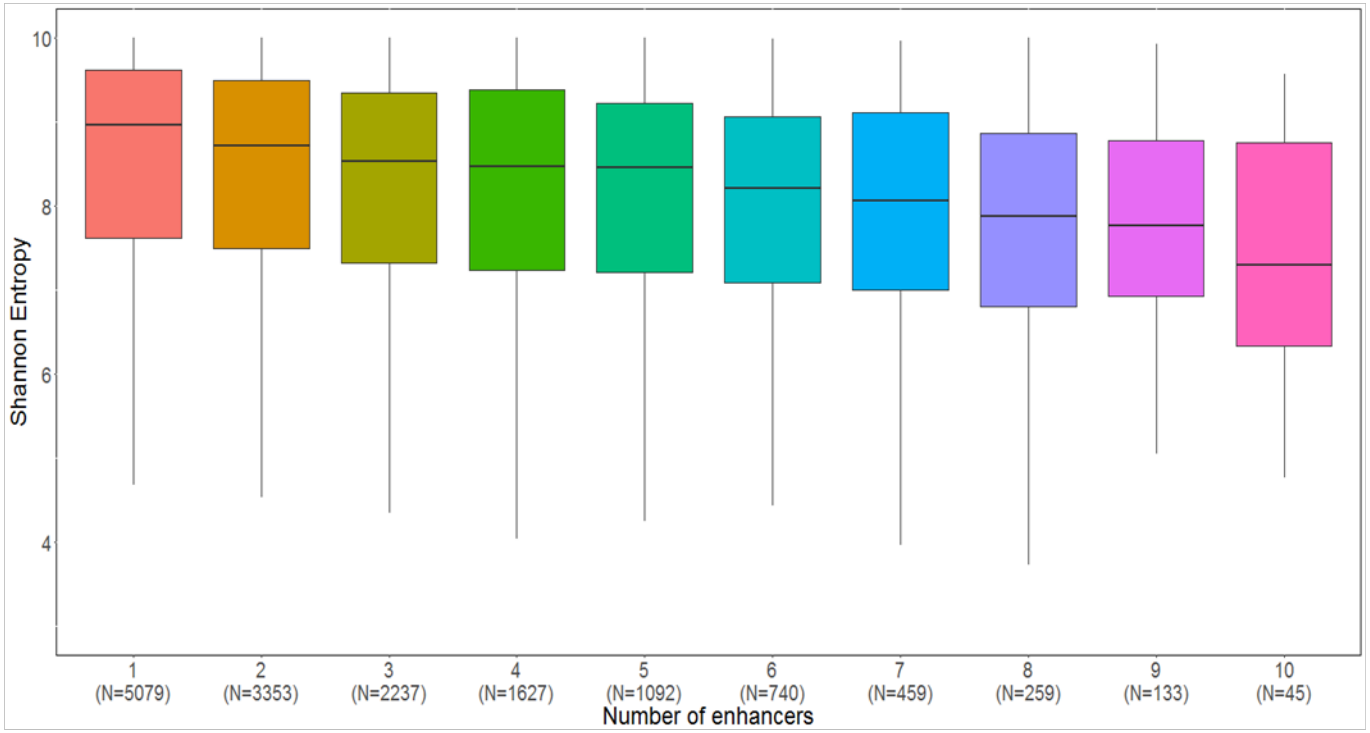


Figure S2.12. Opposite relationship between breadth of promoter activity over cell types and complexity of transcriptional regulation. We quantified the breadth of promoter activity over different cell types by Shannon entropy. Promoters were divided into bins according to the number of enhancers included in their optimally reduced models and the distribution of Shannon entropy was calculated for each bin (the number of promoters assigned to each bin is indicated in parentheses). A marked inverse relationship is observed. (a) ENCODE DHS data. (b) FANTOM5 CAGE data.

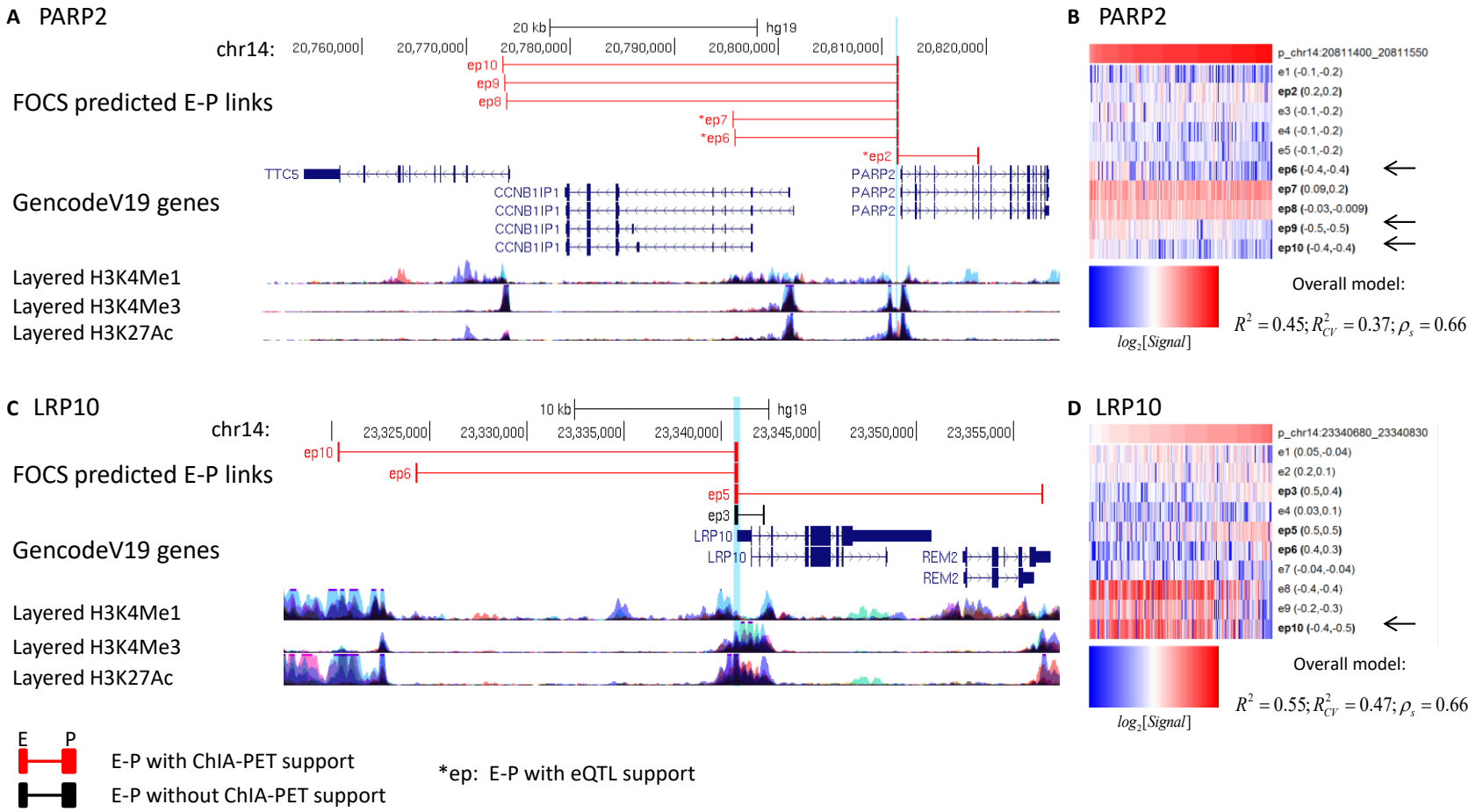


Figure S2.13. Examples for promoter models that include negatively correlated enhancers. (see legend of Fig. 2.5). In the heatmap, negatively correlated enhancers (indication of a repressor function) are indicated by an arrow.

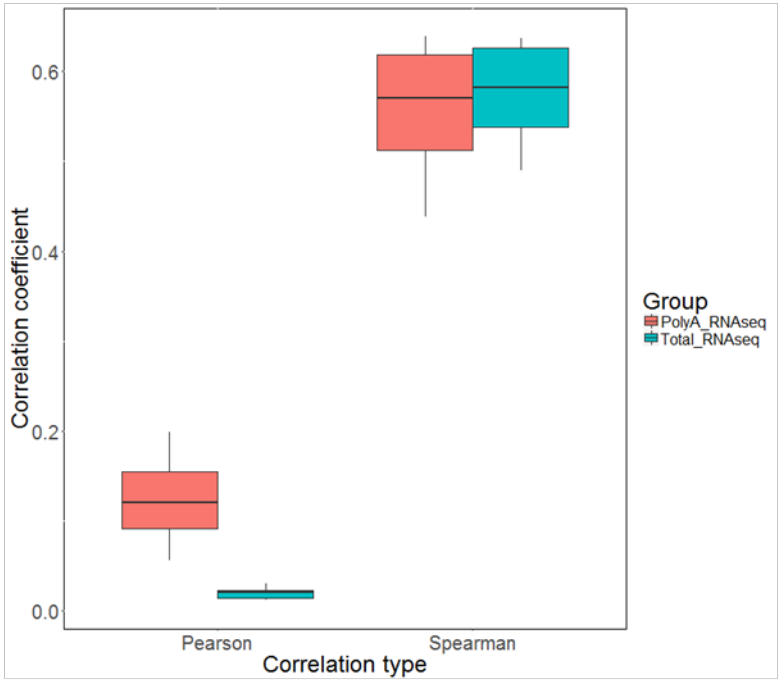


Figure S2.14. Correlation between promoter DHS signal and gene expression. We examined the correlation between DHS signal at promoters and gene expression levels using ENCODE cell lines for which both DHS and RNA-seq dataset were available (this included 11 cell-lines with polyA RNA-seq and 6 cell lines with total RNA-seq). In all cases, we observed high Spearman but low Pearson correlation indicating strong monotonic, non-linear relationship.

Supplementary Tables

Table S2.1. Number of promoter models in each regression method					
Method	**Data**	**Both**	**Activity level only**	**Binary only**	**None**
OLS (FDR≤0.1)	ENCODE	52,658	17,807	15,437	7,007
GLM.NB(FDR≤0.1)	ENCODE	33,286	20,233	17,950	21,440
ZINB(FDR≤0.1)	ENCODE	41,336	19,919	12,672	18,982
OLS (FDR≤0.2)	ENCODE	55,975	17,083	14,036	5,815
GLM.NB(FDR≤0.2)	ENCODE	37,094	19,879	17,549	18,387
ZINB(FDR≤0.2)	ENCODE	44,240	19,742	12,384	16,543
OLS (FDR≤0.1)	Roadmap	12,315	9,526	5,242	5,546
GLM.NB(FDR≤0.1)	Roadmap	6,752	7,493	5,369	13,045
ZINB(FDR≤0.1)	Roadmap	8,728	7,646	4,550	11,705
OLS (FDR≤0.2)	Roadmap	13,124	9,530	5,053	4,922
GLM.NB(FDR≤0.2)	Roadmap	7,570	7,929	5,428	11,702
ZINB(FDR≤0.2)	Roadmap	9,520	8,064	4,566	10,479
OLS (FDR≤0.1)	FANTOM5	9,943	5,081	11,043	30,223
GLM.NB(FDR≤0.1)	FANTOM5	14,197	3,221	13,758	25,114
ZINB(FDR≤0.1)	FANTOM5	13,640	3,377	13,461	25,812
OLS (FDR≤0.2)	FANTOM5	11,072	5,127	11,503	28,588
GLM.NB(FDR≤0.2)	FANTOM5	15,396	3,210	13,530	24,154
ZINB(FDR≤0.2)	FANTOM5	14,719	3,308	13,429	24,834
OLS (FDR≤0.1)	GRO-seq	3,507	236	2,580	2,037
GLM.NB(FDR≤0.1)	GRO-seq	606	377	2,659	4,718
ZINB(FDR≤0.1)	GRO-seq	1,334	657	2,844	3,525
OLS (FDR≤0.2)	GRO-seq	3,745	249	2,509	1,857
GLM.NB(FDR≤0.2)	GRO-seq	798	453	2,830	4,279
ZINB(FDR≤0.2)	GRO-seq	1,566	681	2,907	3,206
Each promoter model contained 10 enhancers as features. The number of EP links is $\boldsymbol{y} \cdot \mathbf{10}$ links where $\boldsymbol{y}$ is the number of promoter models in each category					

Table S2.2. Number of statistically validated promoter models and EP links predicted by FOCS on four genomic resources					
Data type	#promoter models	*#E-P links*	#Unique enhancers	% intronic E-P links *	# known genes**
ENCODE - DHS	70,465	167,988	92,603	74	12,256
Roadmap - DHS	21,841	69,619	49,327	67	10,668
FANTOM5 - eRNA	15,024	41,836	18,656	55	8,666
GRO-seq - eRNA	6,323	22,607	20,650	79	6,323
(*) E-P links whose E is located within an intron of a gene (not necessarily the target gene) (**) Number of Entrez genes associated with promoters					

Table S2.3. Summary of inferred EP links				
Method type	**Data**	**# promoter models**	**#Links to enhancers**	**#Unique enhancers**
Pair-wise	ENCODE	92,080	2,396,287	326,184
Pair-wise-$\boldsymbol{r = 0.7}$	ENCODE	39,372	139,170	53,950
OLS-LASSO[1]	ENCODE	39,368	122,064	74,104
OLS-enet[1]	ENCODE	39,407	150,158	85,926
FOCS*	ENCODE	70,465	167,988	92,603
Pair-wise	Roadmap	32,000	1,023,409	106,231
Pair-wise-$\boldsymbol{r = 0.7}$	Roadmap	8,606	33,598	24,657
OLS-LASSO[2]	Roadmap	6,783	27,414	21,062
OLS-enet[2]	Roadmap	6,788	31,923	24,167
FOCS*	Roadmap	21,841	69,619	49,327
Pair-wise	FANTOM5	42,234	228,908	45,936
Pair-wise-$\boldsymbol{r = 0.7}$	FANTOM5	2,224	4,681	2,449
OLS-LASSO[3]	FANTOM5	1,680	3,970	2,219
OLS-enet[3]	FANTOM5	1,684	5,239	2,771
FOCS*	FANTOM5	15,024	41,836	18,656
Pair-wise	GRO-seq	7,825	113,817	81,040
Pair-wise-$\boldsymbol{r = 0.7}$	GRO-seq	4,347	26,827	24,247
OLS-LASSO[4]	GRO-seq	4,570	17,141	16,121
OLS-enet[4]	GRO-seq	4,580	21,379	19,796
FOCS**	GRO-seq	6,323	22,607	20,650
FOCS-randCV	GRO-seq	7,004	23,960	21,679
(1) The number of OLS promoter models ($\boldsymbol{R^2 \geq 0.5}$) was 39,892 before model selection (2) The number of OLS promoter models ($\boldsymbol{R^2 \geq 0.5}$) was 6,807 before model selection (3) The number of OLS promoter models ($\boldsymbol{R^2 \geq 0.5}$) was 1,951 before model selection (4) The number of OLS promoter models ($\boldsymbol{R^2 \geq 0.5}$) was 4,851 before model selection (*) Selected promoter models passed either both validation tests or the activity level test only (**) Selected promoter models passed either binary test and/or the activity level test				

Supplemental Methods

GRO-seq data preprocessing

We downloaded raw sequence data of 245 GRO-seq samples from the Gene Expression Omnibus (GEO) database (**Additional file 3: Table S5**). First, we applied read quality control on each profile using the Trimmomatic tool (default parameters) [190]. From each read we trimmed (1) bases from Illumina Tru-seq adapters, and (2) bases with low base quality scores from both ends. We excluded reads with net length <30 bases. Finally, we cropped each read to the first 30 bases from the 5' end. Second, we aligned the trimmed read to a set of known ribosomal RNA (rRNA) genes (FASTA sequences taken from NCBI: RN18S1, RN28S1, RN5, and RN5S17) using bowtie2 [191] (default parameters), and discarded reads aligned to rRNA genes. Third, we aligned the rest of the reads to hg19 reference genome using bowtie2 (default parameters). For subsequent analyses we used only reads that had a MAPQ score greater than 10. Fourth, we merged aligned reads from multiple profiles with the same sample id (via GEO GSM id) into a single sample. In total, our collected GRO-Seq database covered 40 studies encompassing 245 samples from 23 cell lines, each assayed under control and stress conditions (**Additional file 3: Table S5**).

We quantified gene transcription activity by counting the number of reads mapped within each (unspliced) gene. As gene models we used a single transcript per gene, constructed using groHMM's makeConsensusAnnotations function [192] and hg19 UCSC refGene table, producing 22,891 consensus genes. We only used reads mapped to the gene's transcript body in the range 0.5kb to 20kb downstream of the TSS. If the transcript's length was less than 20kb then we used only the region up to the transcript termination site (TTS).

To identify active enhancers in each sample, we applied dREG [21] on the aligned reads. dREG detects "*transcriptional regulation elements*" (TREs) based on symmetric forward and reverse read coverage relative to their center position. This symmetry is a known mark of short putative enhancers [193]. We merged overlapping TREs (taking the union of their locations) detected in different samples to create *merged TREs* (mTREs). We defined as enhancers mTREs that are either: (1) intergenic: mTREs whose center is located at least 5kb from the closest gene's TSS and does not overlap any gene's transcript body, or (2) intronic: mTREs that are not exonic and have overlap with an intron of a gene. We counted the number of reads in each intergenic enhancer (in both strands) and intronic enhancer (only in antisense strand) in each sample using BEDTools [109].

The gene and enhancer expression matrices were further filtered to include only genes/enhancers (rows) with at least one sample (columns) with RPKM ≥ 1, in order to preserve only expressed genes/enhancers. Next, to focus of the analysis on differential genes, we calculated for each the coefficient of variation (CoV) (the ratio between the gene's standard deviation σ to the mean μ), and selected the most variable ones as follows: (1) we partitioned the genes according to their mean RPKM expression into 20 bins. (2) In each bin we retained the genes with CoV above the bin's median level. These two steps also reduce preference to highly or lowly expressed genes. The final gene matrix contained 8,360 genes, and the final enhancer matrix contained 255,925 enhancers.

We defined for each gene the set of k=10 candidate enhancers located within a window of ±500Kb from its TSS.

FOCS Model Implementation

The input to FOCS is two activity matrices, one for enhancers (M_e) and the other for promoters (M_p), measured across the same samples. Activity is measured by DHS signal in ENCODE and Roadmap data, and by expression level in FANTOM5 and GRO-seq data. Samples were labeled with a cell-type label out of C cell-types. The output of FOCS is predicted E-P links.

First, FOCS builds for each promoter an OLS regression model based on the k enhancers whose center positions are closest to the promoter's center position (in ENCODE, Roadmap, and FANTOM5) or TSS (in GRO-seq). Formally, let y_p be the promoter p normalized activity pattern (measured in CPM - counts per million; y_p is a row from M_p) and let X_p be the normalized activity matrix of the corresponding k enhancers (CPM; k rows from M_e). We build an OLS linear regression model $\mathrm{y_p} = \mathrm{X_p}\beta_\mathrm{p} + \varepsilon_\mathrm{p}$, where ε_p is a vector that denotes the errors of the model and β_p is the $(k+1)\ x\ 1$ vector of coefficients (including the intercept) to be estimated.

Second, FOCS performs leave-cell-type-out cross validation (LCTO CV) by training the promoter model based on samples from $C-1$ cell types and testing the predicted promoter activity of the samples from the left out cell type. This step is repeated C times. The result is a vector of predicted activity values y_p^{model} for all samples.

FOCS tests the predicted activity values using two validation tests: (1) The *binary test*. This test examines whether y_p^{model} discriminates between the samples in which p was active (observed activity $y_p \geq 1$ RPKM) and the samples in which p was inactive ($y_p < 1$ RPKM). (2) The *activity level test*. This test calculates, for the active samples, the significance of the Spearman correlation between y_p^{model} and y_p. Spearman correlation compares the ranks of the original and predicted activities. We obtain two vectors of p-values, one for each test, of length n (the number of promoter models).

Third, to correct for multiple testing, FOCS applies on each p-value vector the Benjamini - Yekutieli (BY) FDR procedure [104]. Promoter models with q-value≤ 0.1 in either both tests or in the activity level test were included in further analyses. In GRO-seq analysis, we also included models that passed only the binary test (m=2,580) since 57% of them had $R^2 \geq 0.5$ (**Fig. S2.6B**). For promoters that passed these CV tests final models are trained again using all samples.

FOCS next selects informative enhancers for each final promoter model. First, to control the FDR due to multiple hypotheses we used the BY correction. We call this process *enhancer BY FDR filtering* (**eBY**). The OLS results provide for each model P-values for the coefficients of its 10 closest enhancers. FOCS applies BY correction on the P-values produced by all models together and selects enhancers with q-value ≤ 0.01. To identify the most important ones out of the selected (≤ 10) enhancers for each promoter model, FOCS applies elastic-net model shrinkage (**enet**) with a regularization parameter λ, using the glmnet R function [194] with mixing parameter α=0.5, giving equal weights for Lasso and Ridge regularizations. We require that all the enhancers that survived eBY filtering will be included in the shrunken model. To achieve this we take the maximum λ satisfying this property. For models in which no enhancer survived the eBY filtering, we took the maximum λ yielding a shrunken model with at least one enhancer. This ensures that every promoter that passes the CV tests also has a model following the enet step.

Alternative regression methods

We compared the performance of OLS method with GLM.NB and ZINB regression methods. We repeated the FOCS steps but in the first step, instead of OLS we applied the GLM.NB or the ZINB methods. In GLM.NB/ZINB we used for y_p and X_p the raw count values instead of CPM. To correct the model according to differences in samples library sizes, we provided these sizes as an offset vector to GLM.NB and ZINB methods.

FANTOM5 E-P linking using OLS regression was followed by Lasso shrinkage (defined as OLS-LASSO) as described in [23]. Briefly, promoter models were created using OLS and models with $R^2 \geq 0.5$ were accepted for further analyses. Next, penalized Lasso regression was used to reduce the number of enhancers in the models. Optimal models were selected using 100-fold cross validation and the largest value of lambda such that the mean square error was within one standard error of the minimum, using the cv.glmnet() function in R glmnet package [194]. OLS followed by enet (called OLS-enet) was run with mixing parameter α = 0.5 in the cv.glmnet() function. OLS followed by LASSO (OLS-LASSO) was run with $\alpha = 1$.

GO enrichment analysis

GO enrichments were calculated using topGO R package [115] (algorithm="classic", statistic="fisher", minimum GO set size=10). We split the genes into target and background sets using their enhancer bin sets. Genes belonging to bins with 1-3/1-4/4-10/5-10 enhancers were considered as target set and compared to all genes from all bins as background set. Correction for multiple testing was performed using BH procedure [67].

External validation of predicted EP links

We used three external data resources for validating FOCS E-P link predictions: (1) RNAPII ChIA–PET interactions, (2) YY1-HiChIP interactions, and (3) eQTL SNPs.

We downloaded 922,997 ChIA-PET interactions (assayed with RNAPII, on four cell lines: MCF7, HCT-116, K562 and HelaS3) from the chromatin–chromatin spatial interaction (CCSI) database [116] (GEO accession numbers of the ChIA-PET samples are provided in supplementary table S6). We used the liftOver tool (from Kent utils package provided by UCSC) to transform the genomic coordinates of the interactions from hg38 to hg19. HiChIP interactions mediated by YY1 TF (cell types: HCT116, Jurkat, and K562) were taken from [105] (GEO accession id: GSE99521). As done in [105], we retained 911,190 YY1-HiChIP interactions with origami probability>0.9. Origami is a method that aims to find high confident interactions. For eQTL SNPs, we used the significant SNP-gene pairs from GTEx analysis V6 and V6p builds. 2,283,827 unique eQTL SNPs covering 44 different tissues were downloaded from GTEx portal [106].

We used 1Kbp intervals (±500 bp upstream/downstream) for the promoters (relative to the center position in ENCODE/Roadmap/FNATOM5 or to the TSS position in GRO-seq) and the enhancers (±500 bp from the enhancer center). An E-P pair is considered supported by a particular capture interaction if both the promoter and enhancer intervals overlap different anchors of an interaction. An E-P pair is considered supported by eQTL SNP if the SNP is located within the enhancer's interval and is associated with the expression of the promoter's gene. For each predicted E-P pair we checked if the promoter and enhancer intervals are

supported by capture interactions and eQTL data. We then measured the fraction of E-P pairs supported by these data resources.

To get an empirical P-value for the significance of the fraction, we performed 100 permutations on the data (100 permutations were sufficient as in all methods we got empirical P-value<0.01). In each permutation, for each promoter independently, if it had l E-P links, then l enhancers on the same chromosome with similar distances from the gene's TSS as the l linked enhancers were selected randomly. For this purpose we used the R 'Matching' package [195]. The fraction of overlap with the external data was computed on each permuted data.

Statistical tests, visualization and tools used

All computational analyses and visualizations were done in the R statistical language environment [117]. We used the two-sided Wilcoxon rank-sum test implemented in wilcox.test() function to compute the significance of the binary test. We used the cor.test() function to compute the significance of the Spearman correlation in the activity level test. Spearman/Pearson correlations were computed using the cor() function. To correct for multiple testing we used the p.adjust() function (method='BY'). We used 'GenomicRanges' package [118] for finding overlaps between genomic positions. We used 'rtracklayer' [119] and 'GenomicInteractions' [120] packages to import/export genomic positions. Counting reads in genomic positions was calculated using BEDTools [109]. OLS models were created using lm() function in 'stat' package[117]. GLM.NB models were created using glm.nb() function in 'MASS' package [121]. ZINB models were created using zeroinfl() function in 'pscl' package [122]. Graphs were made using graphics[117], ggplot2 [123], gplots [124], and the UCSC genome browser (https://genome.ucsc.edu/).

6.2. Supplement 2: Predicting cell-type specific enhancer-promoter maps

Supplementary Results

Validations against experimentally detected chromatin loops

GM12878 and K562 predicted ct-links on FANTOM5 and ENCODE data were benchmarked against experimental 3D loops in terms of precision-recall curves. We compared CT-FOCS with MAD-FOCS and JEME [133] on FANTOM5 data (**Methods; Supplementary Figure S3.11**) and with the ABC model [135,136] and TargetFinder [134] on ENCODE data (Methods; **Supplementary Figure S3.12**). Note that the very small number of cell types assayed for 3D chromatin loops does not allow us to identify true cell type-specific loops and exclude those common to many cell types. Therefore, the benchmark does not provide a gold standard of positive and negative ct-links. Thus, the validations answer only on whether predicted the ct-links are supported by experimental loops or not, but not on their cell type-specificity (that is, tend to occur in one or a few cell types compared to other cell types).
To this end, the results in **Supplementary Figures S3.11-3.12** show: (1) validations against two-step loop sets (TLSs; Methods) had 2-3 fold increase in precision compared to validations against single loops in all methods. (2) ct-links predicted by all methods had a relatively low support from 3D chromatin loops with CT-FOCS achieving higher precision on K562 (FANTOM5 and ENCODE) and GM12878 (ENCODE) than the other tested methods.

Loops involving enhancers active in a single cell type

We wished to evaluate the prevalence of experimentally detected EP links that involve enhancers that are active only in one single cell type. We performed the following analysis:

(1) We identified FANTOM5 enhancers present only in a single cell type as follows: first, each enhancer, x, (out of n≈43k enhancers) is sorted by its signals across the cell types from highest to lowest. Second, a fold-change, FC(x), is computed between the first ranked cell type T_1 and second ranked cell type T_2. Lastly, given a threshold y, if FC(x) is above the y percentile of all n FCs, then enhancer x is considered as **uniquely active** (**ua**) in cell type T_1 (termed **ua-enhancer**). We varied y between 40% and 95%. The higher the y threshold is - the more likely the enhancer that passed this criterion to be really active in a single cell type.
Notably, even for the y=95 percentile, the fold-change was a mere 2.3, very far from what could be considered a strictly unique cell-type enhancer. Moreover, with that threshold, only 12 enhancers were identified as uniquely active in GM12878.

(2) We took GM12878 POL2 ChIA-PET loops (m=95,269 loops) from [137] and resized the loop anchors to 5kb around their center position. We searched for loops whose anchors overlap with GM12878-ua enhancers. We counted how many ua-enhancers had an overlap with a loop anchor, and how many loops had anchors overlapping a ua-enhancer. We termed these loops as GM12878 specific E-loops (termed **sE-loops**).
For y=95, only 9 enhancers met this criterion, and they involved 45 sE-loops. So the vast majority of the experimentally obtained GM12878 loops (95,224/95,269 or 0.9995) did not involve a ua-enhancer, suggesting that this is a very rare case.

(3) Finally, we counted how many of the sE-loops involve anchors annotated as promoters (termed **sEP-loops**).
For y=95, only four loops met this criterion.

Even with the lowest tested threshold (y=40%, corresponding to a mere 1.0 fold-change), only 64 GM12878-ua enhancers were found. Out of them, 46 overlapped with 316 experimentally validated loops, and 70 loops involved an anchor annotated as promoter. So a truly unique enhancer is very uncommon.

We repeated the analysis for the cell type K562 using m=41,452 K562 ChIA-PET loops, and the numbers were even lower.

All the results are summarized in **Supplementary Table S5.**

Our analysis suggests that truly uniquely active enhancers are very rare. Therefore, we argue that it will be hard for any computational method to identify ct-links involving enhancers active in a single cell type.

We also tested a relaxed version of CT-FOCS that skips the leave-cell type-out cross validation step of FOCS and applies CT-FOCS to all 24,048 available promoters with their ten closest enhancers in FANTOM5. The latter method is termed 'CT-FOCS no filtering'. The reasoning was to allows a potential unique cell type not to be excluded in the cross validation. The results are shown in the table below. While CT-FOCS linked fewer promoters (~14K) than 'CT-FOCS no filtering' (~21K), both methods linked a similar number of enhancers (~27K). Also, both methods linked ~5K enhancers whose region appear in a single cell type (row 5 in the table – set A). Out of set A (row 6 in the table – set B), only ~200 enhancers (~4%) were ranked first by their signal in the same cell type compared to other cell types. The average and median FC (described in point 1 above) of the enhancers in set B was similar in both methods.

This analysis suggests that it is unlikely to observe enhancers that are strictly active in a single cell type. This may be because similar cell types from the same tissue have the same active enhancer. The more cell types are used in the analysis - the less likely we are to observe an enhancer active in a single tissue. Thus, the leave-cell-type-out-cross-validation step in FOCS has a minor effect on the identification of EP links that are strictly unique and active in a single cell type.

	CT-FOCS	**CT-FOCS no filtering**
#candidate promoter models	21,468	24,048
#candidate enhancers	36,244	37,193
#linked promoters	13,873	21,068
#linked enhancers	27,463	27,062
A - #linked enhancers in a single cell type	4,933 (18%)	5,557 (20.5%)
B - #enhancers from A that were ranked first by signal in the same single cell type vs. other cell types	196 (4%)	175 (3.1%)
Avg/Median enhancer FC b/w first and second ranked cell types by enhancer signal (on the enhancers in B)	1.5/1.3	1.8/1.2

The effect of sample size on the quality of predictions

CT-FOCS requires multiple replicates per cell types. Cell types with a single replicate are also included in our linear mixed effect model as they can contribute to estimating the fixed effect coefficients (β values). Cell types with at least two replicates can help improve variance estimate for the random effect, which is the cell type group level. Using this estimate one can predict a random slope and intercept for each cell type (**Methods**).

In FANTOM5 dataset, 179, 110, and 20 out of 472 cell types had at least 2, 3, and 4 replicates, respectively. In ENCODE, 88 out 106 cell types had at least 2 replicates.

To analyze the sample effect size on the quality of the predictions we applied CT-FOCS only on 179 FANTOM5 cell types that had at least two replicates. We name the resulting solutions as CT-FOCS-2rep. We also analyzed the properties of the original CT-FOCS solutions when restricted to the same set of 179 cell types. Properties of these solutions are summarized in **Supplementary Figure S3.13**. Focusing on these 179 cell types, the original CT-FOCS resulted, on average per cell type, with more significant ct-links, linked enhancers and promoters compared to CT-FOCS-2rep (**Table A** below).

Table A. Average number of ct-links, linked enhancers and promoters on FANTOM5 dataset

Method	# cell types	Avg. ct-links (Median)	Avg. Enhancers (Median)	Avg. Promoters (Median)
CT-FOCS	472	414 (94)	318 (72)	146 (73)
CT-FOCS	179	600 (122)	451 (90)	200 (87)
CT-FOCS-2rep	179	269 (115)	216 (96)	131 (58)

Next, we compared the two solutions using TLS support ratio as done in **Figure 4**. To this end, we focused on 100 cell types that had at least 50 ct-links in both approaches. Also, as previously done, we restricted the number of predictions on these cell types to be the same between CT-FOCS and CT-FOCS-2rep in order to control sensitivity (top EP links selected using $logEP$ value). CT-FOCS-2rep results had better median support ratios on 4 out of 5 cell types compared to the original CT-FOCS (**Table B** below). CT-FOCS was better only on K562 cell line.

Table B. The particularity of each algorithm's predictions as measured by ChIA-PET, HiChIP and PCHi-C assays

	K562 (HiChIP)	GM12878 (ChIA-PET)	Hippocampus (pcHiC)	Liver (pcHiC)	Thymus (pcHiC)
CT-FOCS	**0.9**	1.2	0	0.7	0.6
CT-FOCS-2rep	0.5	**2.0**	**0.3**	**1.1**	**1.2**
p-Value*	<6.5E-13	<1.9E-13	<4.2E-4	<2.6E-6	<2.0E-13

Values in the first two rows are the median $\log_2$(ratio) values.
*One-sided Wilcoxon rank-sum test. Marked in red is the best performing method.

We report also the CT-FOCS-2rep predictions and recommend the readers to use them instead those of CT-FOCS when selecting one of the 179 cell types included in the analysis.

Supplementary Methods

ENCODE DHS data preprocessing

ENCODE DNase-seq samples (106 cell types) were downloaded from GEO dataset GSE29692 [43,107,108]. ENCODE DHS peaks of enhancers and promoters [24] were processed as in FOCS [101] with the following changes: (1) we analyzed only promoters of annotated protein-coding genes according to GencodeV10 TSS annotations (ftp://genome.crg.es/pub/Encode/data_analysis/TSS/Gencodev10_TSS_May2012.gff.gz). (2) We applied a relative-log-expression (RLE) normalization [196], as implemented in edgeR [197,198]. (3) We retained promoters and enhancers that showed robust activity in at least one cell type: signal ≥ 5 RPKM in all samples of at least one cell type. Overall, we analyzed 208 samples from 106 cell types. Our preprocessing resulted with 36,056 promoters (mapped to 13,105, 13,464, and 13,197 protein-coding genes according to HGNC_symbols, Ensembl, and Entrez, respectively) and 658,231 putative enhancers.

Enhancers closer than 10kb to the nearest promoter were discarded since we wanted to reduce false positive links due to the high signal correlation at short distances, and to predict distal interactions as suggested by Whalen et al. 2016. The candidate enhancers for each promoter were defined as the 10 closest enhancers located within a window of 1Mb (±500kb upstream/downstream) from the promoter's center position.

We first applied the FOCS pipeline, including leave-cell-type-out cross validation (LCTO CV) on the promoters and their candidate enhancers, and accepted promoter models with q-value≤ 0.1 in the activity level test (see Hait et al. 2018 for details). Unlike FOCS, we did not apply here regularization on the predicted EP links. Overall, the procedure resulted with 17,832 promoter models (mapped to 9,090, 9,320, and 9,160 HGNC_symbols, Ensembl, and Entrez protein-coding genes, respectively).

FANTOM5 CAGE data preprocessing

We downloaded the FANTOM5 CAGE data from JEME [133] repository (https://www.dropbox.com/sh/wjyqyog3p5d33kh/AACx5qgwRPIij44ImnzvpFxUa/Input%20files/FANTOM5/1_first_step_modeling?dl=0&subfolder_nav_tracking=1). Overall, the data contained 24,048 promoters (mapped to 18,986, 20,597 and 18,912 protein-coding genes according to HGNC_symbols, Ensembl and Entrez, respectively) and 42,656 enhancers, covering 808 samples. Enhancer and promoter expression matrices were RLE normalized. We used the RLE normalized data from [133]. We manually annotated the 808 samples with 472 cell types (**Supplementary Table S6**) using **Table S1** from FANTOM5 [23].

For each promoter, the candidate enhancers were defined as the 10 closest enhancers located within ±1Mb from the promoter's TSS as performed in JEME [133]. Unlike ENCODE, we did not enforce a lower bound on the distance here. We applied the same pipeline on the promoters and their candidate enhancers as described above for the ENCODE data. This resulted with 21,468 promoter models for further analysis.

Supplementary Tables

The supplemental tables are included in the online Supplemental Material in spreadsheet format. The 'Description' tab of each supplemental table file includes a table legend. Below are the titles of the tables:

Supplementary Table S1: Comparison between CT-FOCS, TargetFinder and ABC on ENCODE DHS data across 5-10 cell types

Supplementary Table S2: The particularity of each algorithm's predictions as measured by ChIA-PET, HiChIP and PCHi-C assays

Supplementary Table S3: TF overrepresentation q-values and overrepresentation factors in promoters and enhancers involved in ct-links identified by CT-FOCS on ENCODE data

Supplementary Table S4: TF overrepresentation q-values and overrepresentation factors in promoters and enhancers involved in ct-links identified by CT-FOCS on FANTOM5 data

Supplementary Table S5: The number of ChIA-PET loops supported by enhancers with signal ranked above the y^{th} percentile, for different y values

Supplementary Table S6: FANTOM5 sample annotation. The annotation maps between 808 sample IDs, 472 cell types, and three cell type categories (cell line, primary cell, or tissue)

Supplementary Figures

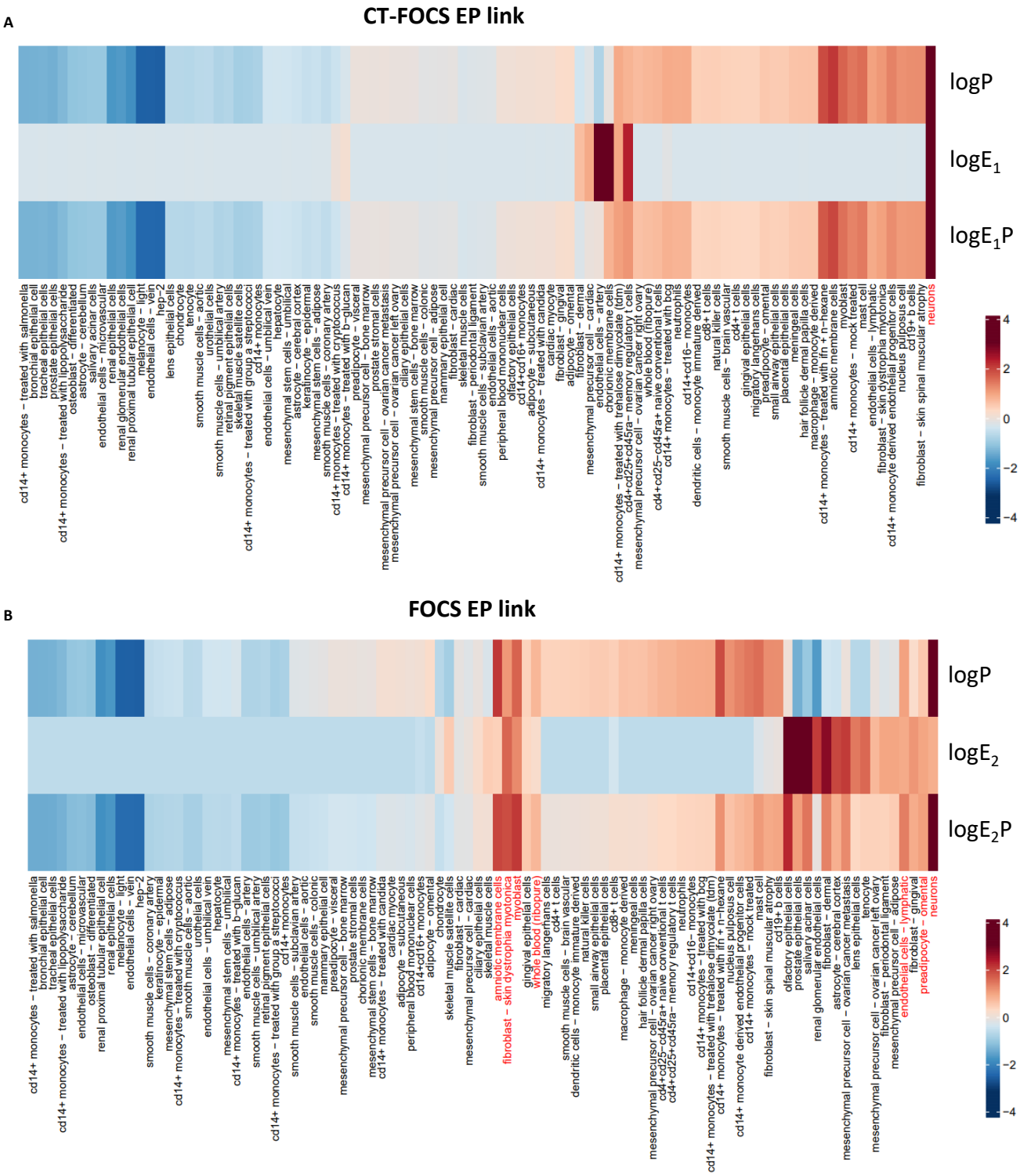
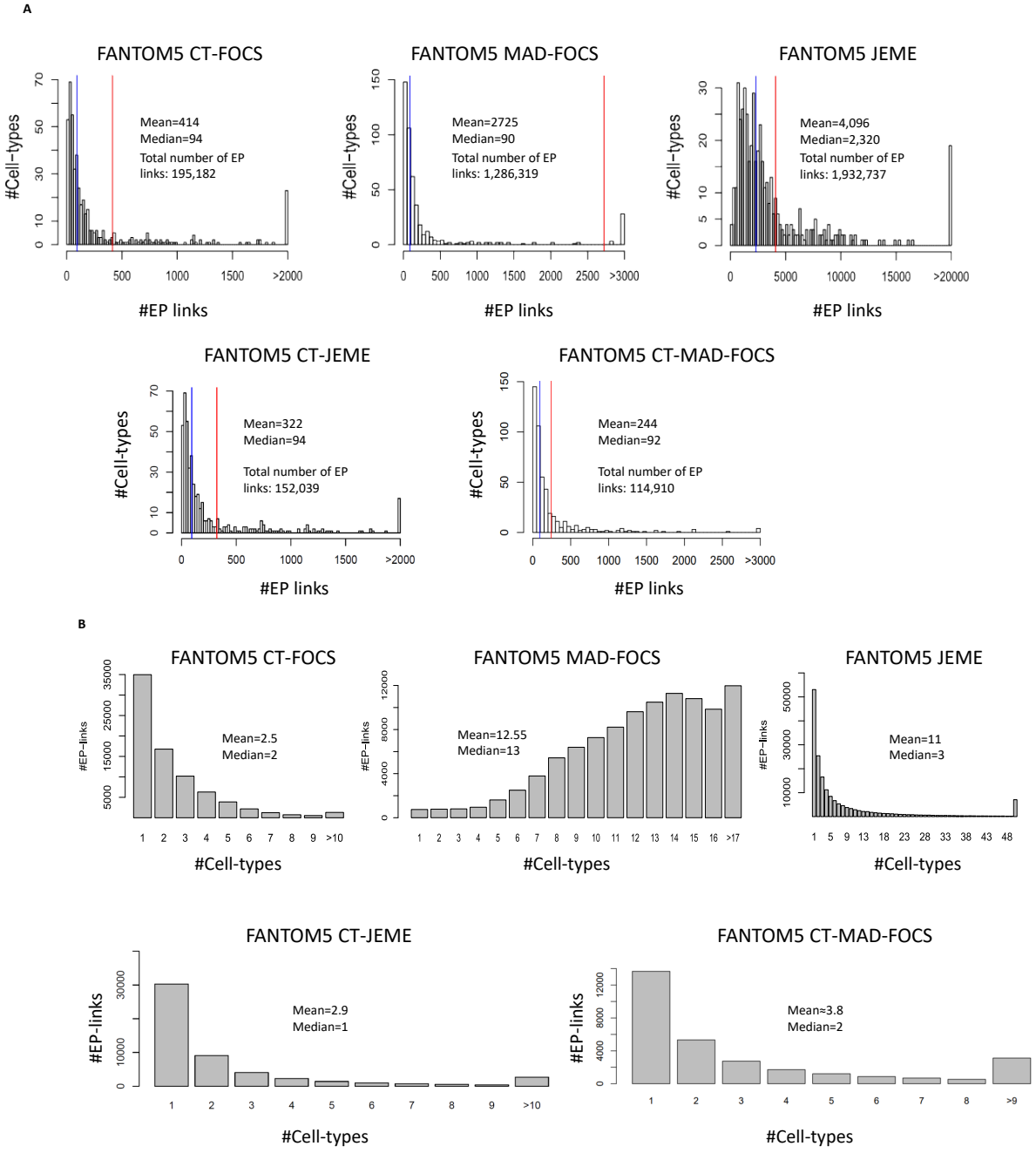


Figure S3.1. The difference between CT-FOCS and FOCS in predicting EP links. (A-B) Heatmaps of two EP links predicted by CT-FOCS and FOCS for the same promoter on FANTOM5 dataset. The links involve different enhancers, E_1 and E_2. **(A)** E_1P link, predicted by CT-FOCS as specific for neurons primary cell. **(B)** E_2P link predicted by FOCS. Cell type names marked in red have both enhancer and promoter signals above the 75% percentile across cell types. Only primary cells (n=94) with at least 3 replicates are presented. Values are in $\log_2$ median of the replicates per cell

type. LogEP is the sum of logE and logP. For visibility, values in each row were transformed to the range -4 to 4. Samples were ordered by hierarchical clustering.



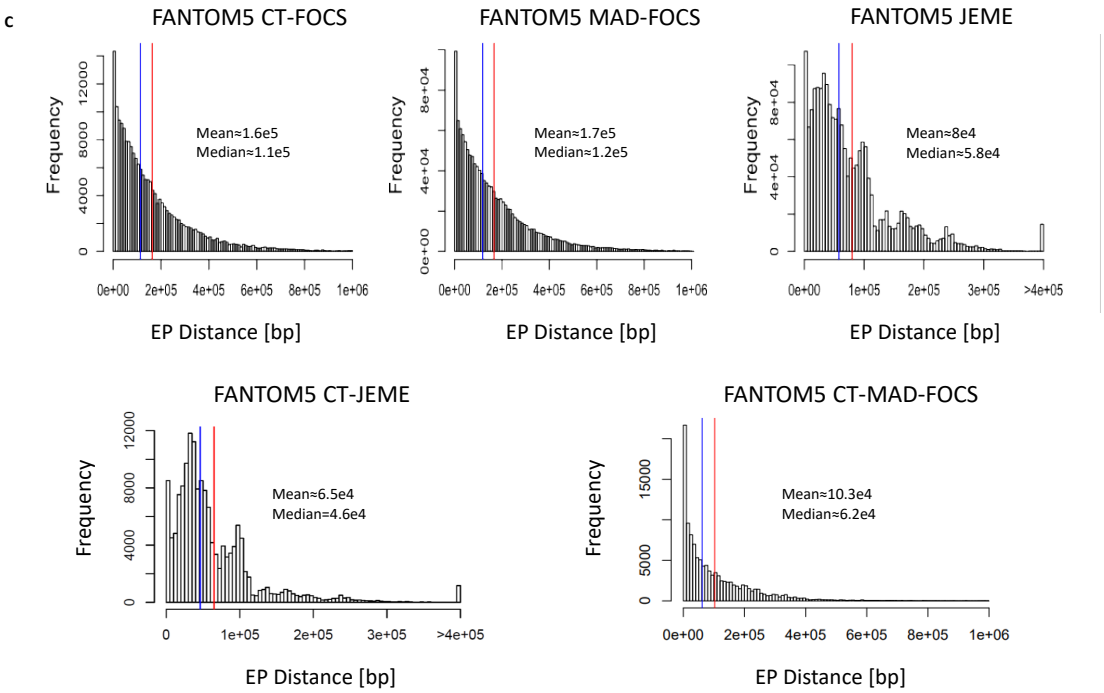


Figure S3.2. Properties of the ct-links predicted by CT-FOCS and by four other algorithms. (A) Number of links predicted per cell type. **(B)** Sharing of ct-links among cell types. **(C)** Enhancer-Promoter distances in predicted ct-links. Distances were collapsed from all cell types, i.e., repeated EP links are counted multiple times. Predictions are on 472 cell types of the FANTOM5 data.

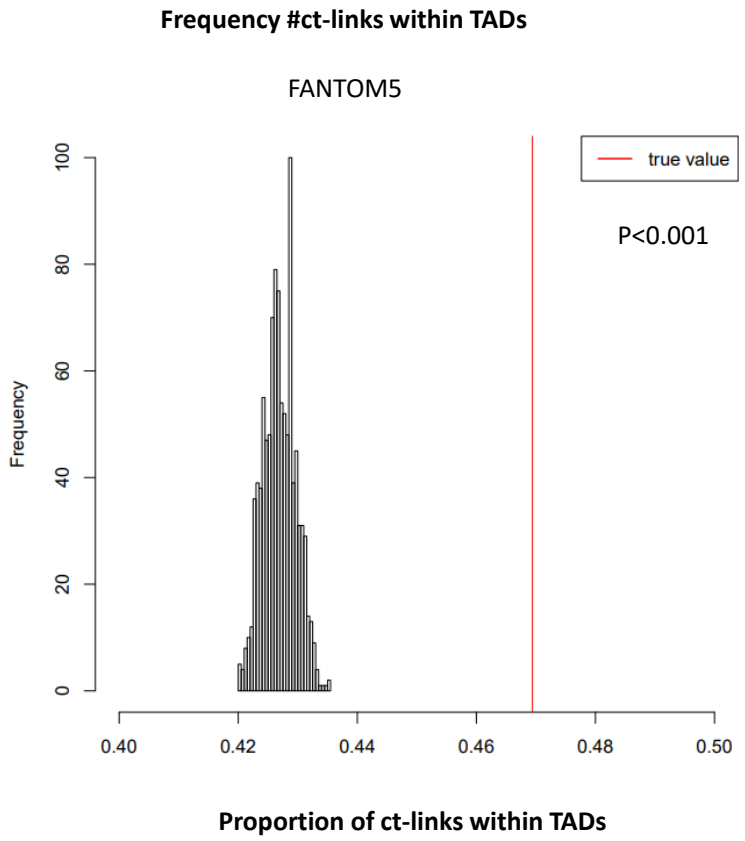


Figure S3.3. Proportion of predicted ct-links that lie within TADs. For each ct-link predicted by CT-FOCS, we randomly selected a promoter in the same chromosome and one of its 10 closest enhancers, and checked if they reside in the same TAD. The process were repeated 1,000 times, and the distribution of the fraction of links that fell within TADs is shown in black. The red line is the fraction obtained on the real data. The empirical p-value is the percentage of cases in which the fraction was higher than that observed on the real data (<0.001 here, as that percentage was zero). The 9,274 TADs reported in [199] were used.

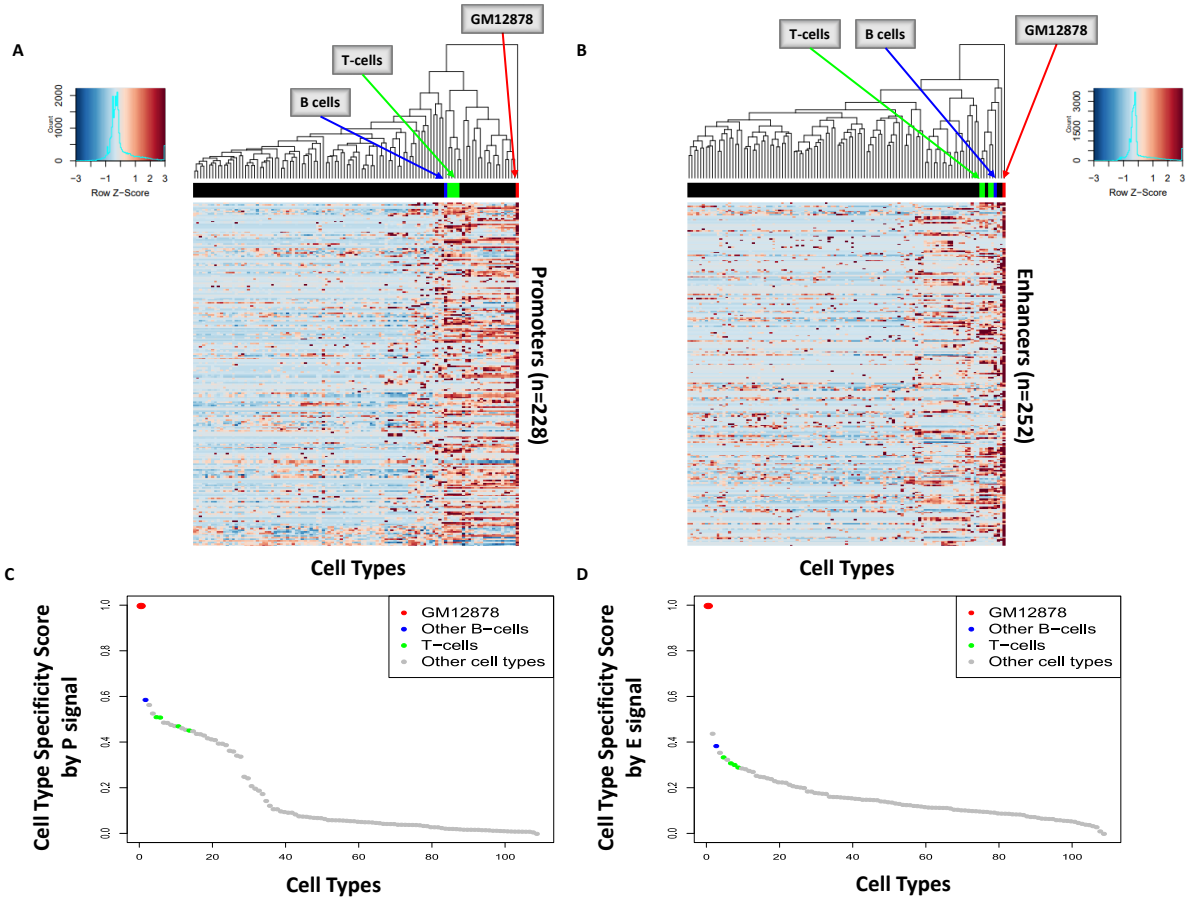


Figure S3.4. Specificity of enhancers and promoters in ct-links predicted for GM12878. (A-B) Heatmaps of linked promoter (A) and enhancer signals (B) for 340 ct-links predicted on GM12878. Columns – cell types, color – z-score of promoter and enhancer signal. B and T cell types related to GM12878 are highlighted in green and blue respectively. **(C-D)** Cell type specificity scores based on promoter (C) and enhancer signals (D). 109 cell types with at least 3 replicates each were included in the analysis (**Methods**).

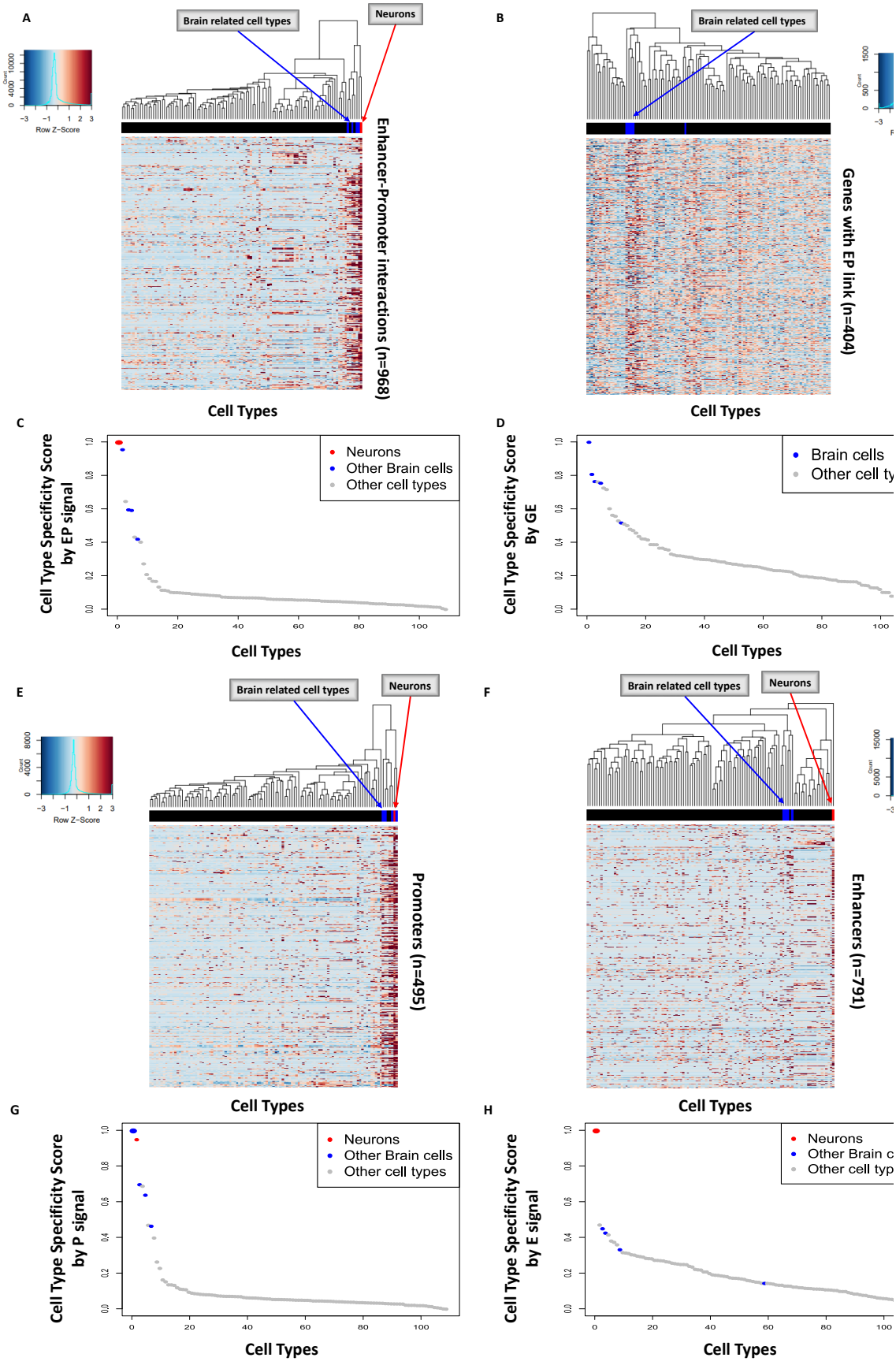
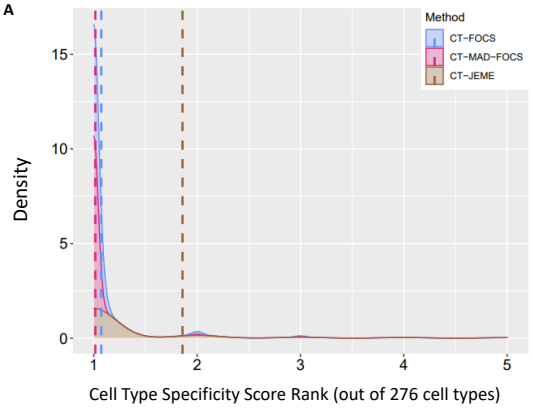

A
Brain related cell types
Neurons
Row Z-Score
Enhancer-Promoter interactions (n=968)
Cell Types
B
Brain related cell types
Genes with EP link (n=404)
Cell Types
C
Cell Type Specificity Score by EP signal
Neurons
Other Brain cells
Other cell types
Cell Types
D
Cell Type Specificity Score By GE
Brain cells
Other cell ty
Cell Types
E
Brain related cell types
Neurons
Row Z-Score
Promoters (n=495)
Cell Types
F
Brain related cell types
Neurons
Enhancers (n=791)
Cell Types
G
Cell Type Specificity Score by P signal
Neurons
Other Brain cells
Other cell types
Cell Types
H
Cell Type Specificity Score by E signal
Neurons
Other Brain c
Other cell typ
Cell Types

Figure S3.5. Specificity of ct-links predicted for Neurons. (A) Heatmap of EP signals for 968 ct-links predicted on Neurons based on FANTOM5 data. Rows – EP links, columns – cell types, color – z-score of EP signal. Brain cell types related to Neurons are highlighted in blue. **(B)** Heatmap of gene expression (GE) for 120 genes involved in the predicted ct-links. Rows – genes, columns – cell types, color – z-score of GE. **(C)** Cell type specificity scores based on the EP signals in A. **(D)** Cell type specificity scores based on expression for the gene set in B. **(E-F)** Heatmaps of linked promoter (E) and enhancer signals (F) for 968 ct-links predicted on Neurons. Columns – cell types, color – z-score of promoter and enhancer signal. **(G-H)** Cell type specificity scores based on promoter (G) and enhancer signals (H). In A, C and E-H, 109 cell types with at least 3 replicates each were included in the analysis; in B and D, 112 cell types with ENCODE gene expression are included (**Methods**).

FANTOM5 dataset

A



P<1.3E-3 (CT-FOCS vs. CT-JEME)
P<0.88 (CT-MAD-FOCS vs. CT-FOCS; not significant)
P<1.7E-4 (CT-MAD-FOCS vs. CT-JEME)

B

Cell type	CT-FOCS	CT-MAD-FOCS	CT-JEME
GM12878	1	3	1
HepG2	1	1	2
K562	1	1	2
MCF-7	1	1	3

Figure S3.6. Specificity scores of EP links predicted by CT-FOCS, CT-MAD-FOCS and CT-JEME. (A) Density plots cell type specificity scores based on EP signals on 276 FANTOM cell types. Dashed lines denote the mean rank for each method. P-values were computed using one sided Kolmogorove-Smirnov test. **(B)** Ranking of the correct cell type in terms of specificity scores of the linked genes' expression in four cell types for which expression data was available in ENCODE.

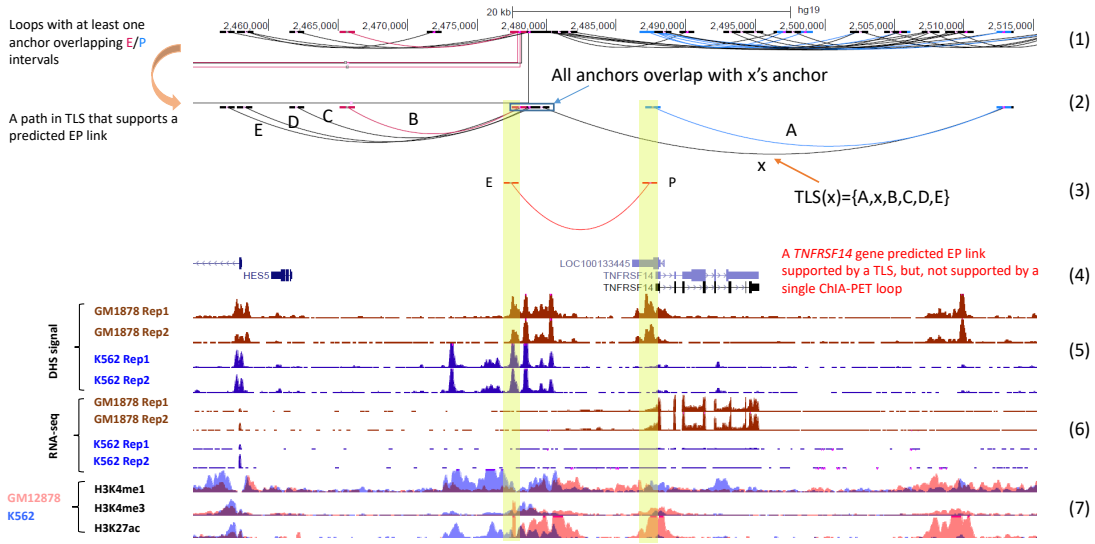


Figure S3.7. ChIA-PET TLSs support ct-links. TLS support for an EP link predicted in GM12878. From the top: (1) A track showing all ChIA-PET loops, highlighting those overlapping an enhancer (red) or a promoter (blue) in a segment of 60 kb of Chromosome 1. (2) All loops that have an anchor in common with loop x. All the anchors in the blue box have nonempty overlap. (3) The EP link predicted by CT-FOCS. That link is validated by TLS(x). (4) Gene annotation. (5) DHS signals of GM12878 and K562 (two replicates each). (6) RNA-seq levels of GM12878 and K562 (two replicates each). (7) Epigenetic marks of enhancer activity (H3K4me1+H3K27ac) and promoter activity (H3K4me3+H3K27ac). The TLS in the second track supports the single GM12878-specific EP link of *TNFRSF14* gene. This link is not supported by a single ChIA-PET loop. Tracks are shown using the UCSC genome browser.

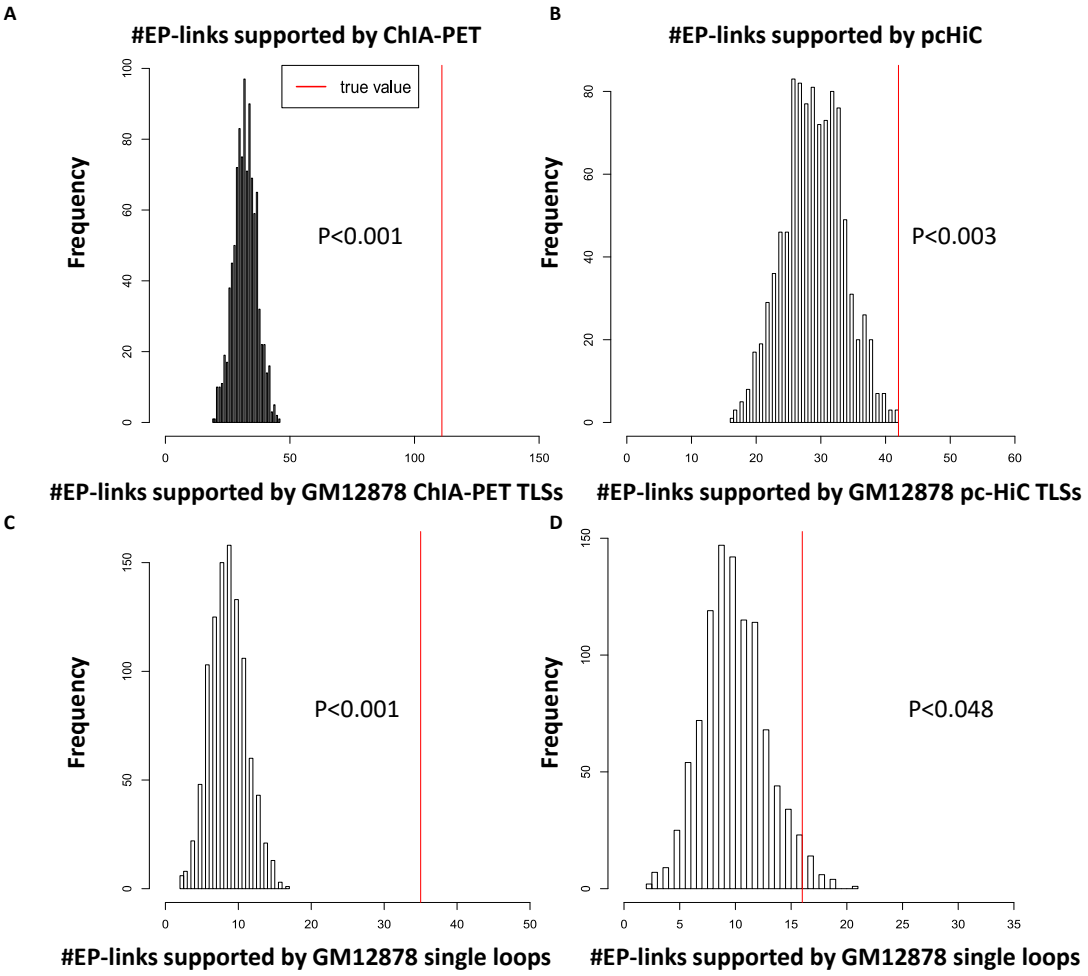


Figure S3.8. Significance of the overlaps between predicted ct-links and experimental 3D contact data. (A-B) Top – overlaps with GM12878 POL2 ChIA-PET (A) and pcHiC (B) TLSs. **(C-D)** Bottom - overlaps with GM12878 ChIA-PET (C) and pc-HiC (D) single loops. Random sets of EP links of the same number and linear distance as the true ct-links were generated. In each set, the number of random EP links that overlapped with the TLSs or single loops in 3D data was counted. The black distribution shows the counts for 1000 random sets, and the red line shows the number obtained for the links inferred by CT-FOCS. Except (D), in all cases the number of random sets with higher counts than the true set was zero.

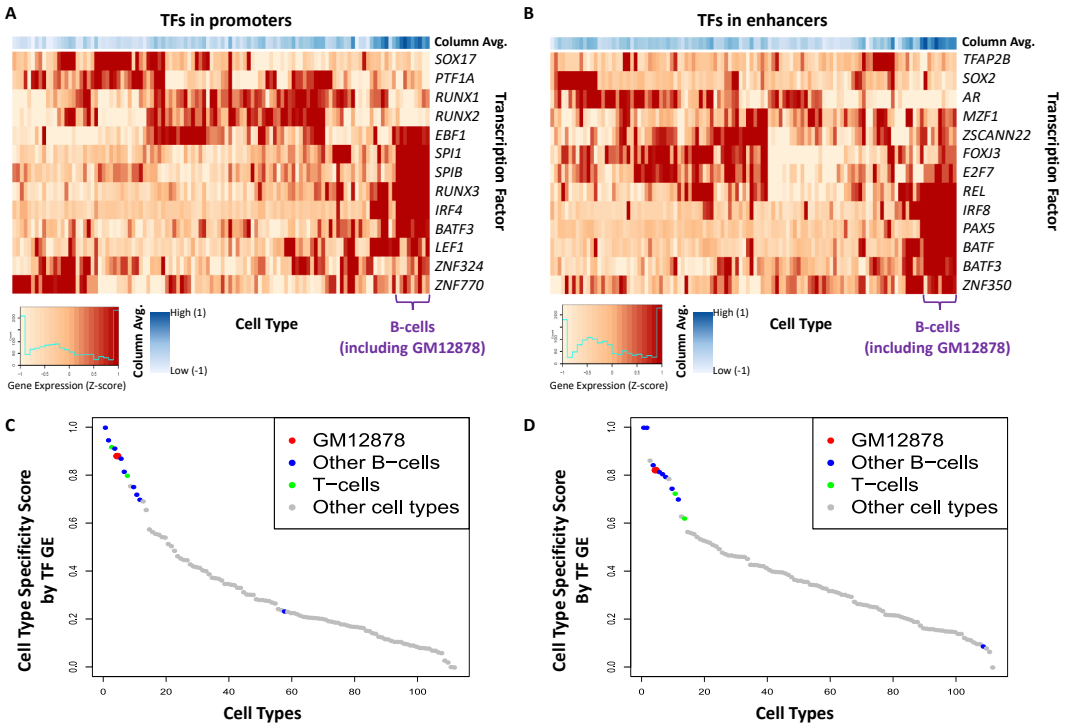


Figure S3.9. Gene expression of TFs whose motif is enriched in ct-links. (A,B) Heatmaps of the expression (after Z-score transformation) of genes encoding the TFs whose motifs were found to be enriched in promoters (A) and enhancers (B) of GM12878-specific EP links identified by CT-FOCS on ENCODE data. TFs shown had q-value < 0.1 (Hypergeometric test). **(C-D)** Cell type specificity score ranks based on GM12878-specific TF GE levels in promoters (C) and enhancers (D) compared to other cell types. 112 ENCODE cell types with ENCODE gene expression are included [44].

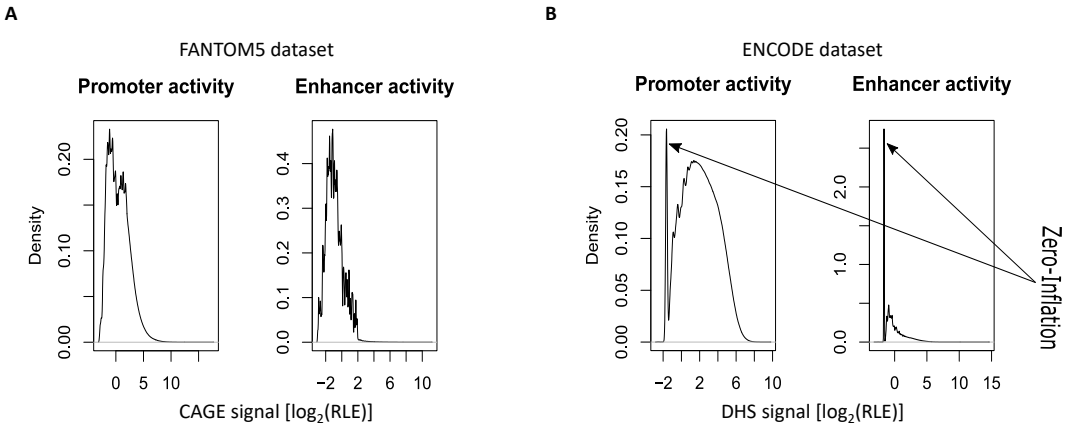


Figure S3.10. Distribution of enhancer and promoter activity signals in FANTOM5 and ENCODE data. (A) FANTOM5 (808 profiles) CAGE signal. The normalization of CAGE signals in CT-FOCS takes into account the library sizes. The activity of each enhancer or promoter was normalized using the RLE function [196] and log2 transformed. **(B)** ENCODE DHS signals (208 profiles). The sharp peaks in both promoter and enhancer indicate zero-inflated data.

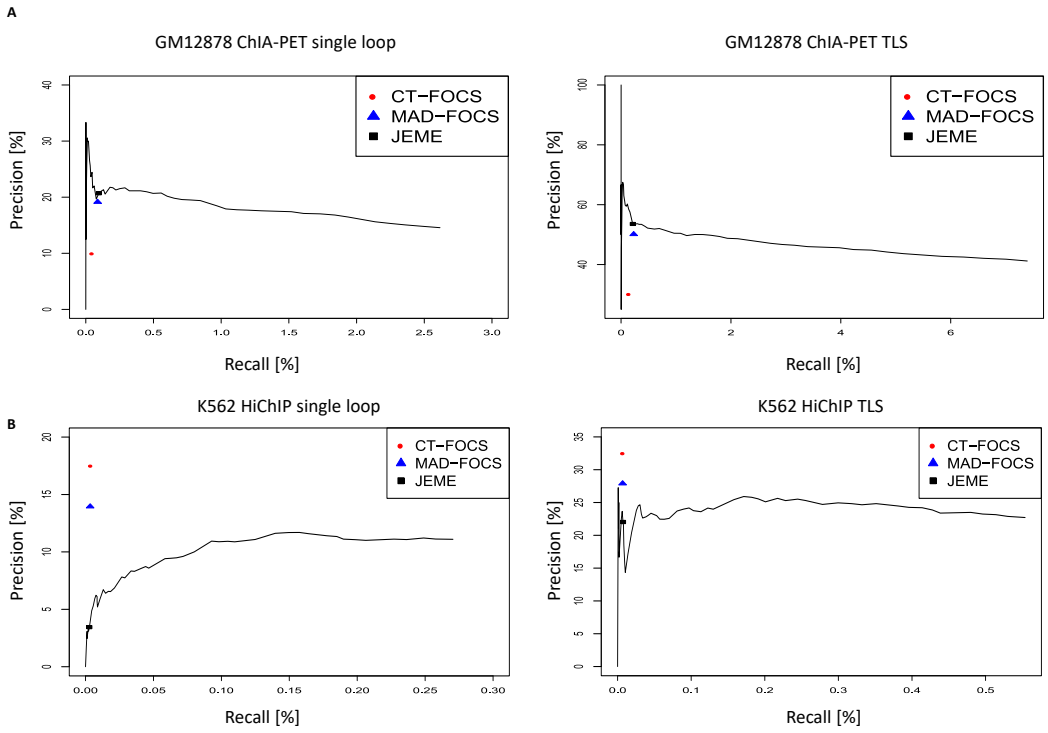


Figure S3.11. Validation of predicted EP-links against 3D chromatin loops on FANTOM5 dataset. Precision-recall curves of CT-FOCS, MAD-FOCS and JEME. Positive set included all GM12878 ChIA-PET loops (A; 92,307 loops) or K562 HiChIP loops (B; 352,359 loops). Black line denotes the JEME's curve on decreasing classification scores from the lowest to the highest recall. Left plots show the validation against single loops. Right plots show the validation against two-step connected loop sets (TLSs).

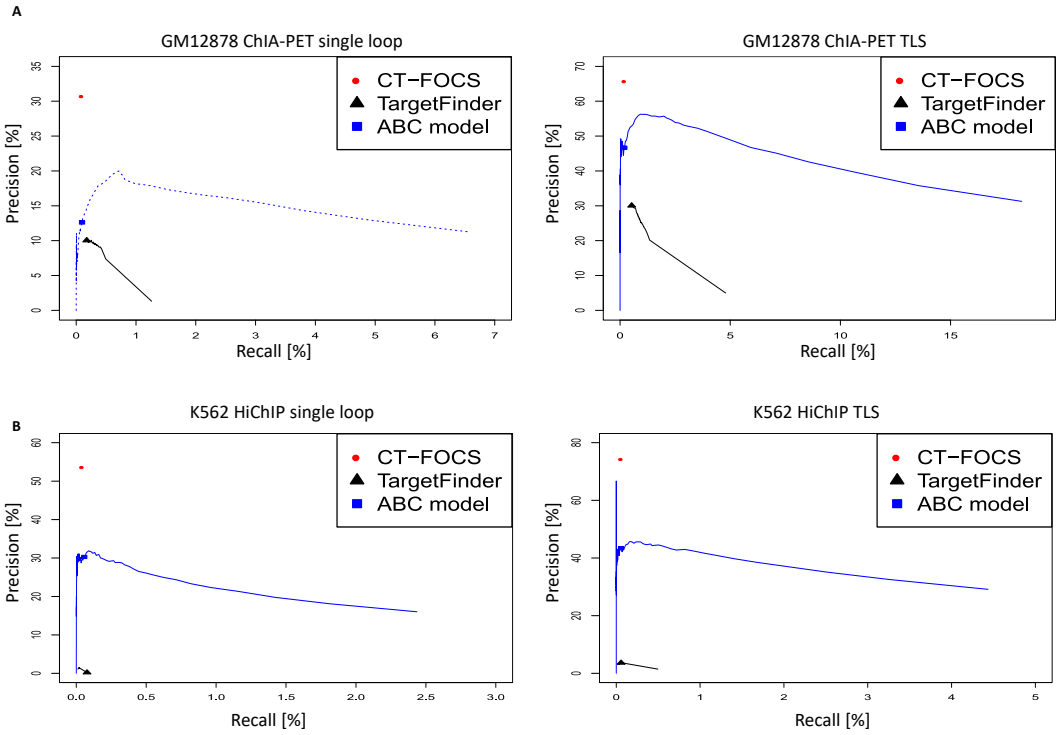


Figure S3.12. Validation of predicted EP-links against 3D chromatin loops on ENCODE dataset. Precision-recall curves of CT-FOCS, TargetFinder and ABC model. Positive set included all GM12878 ChIA-PET loops (A; 92,307 loops) or K562 HiChIP loops (B; 352,359 loops). Black line denotes the ABC model's curve on decreasing ABC scores from the lowest to the highest recall. Left plots show the validation against single loops. Right plots show the validation against two-step connected loop sets (TLSs).

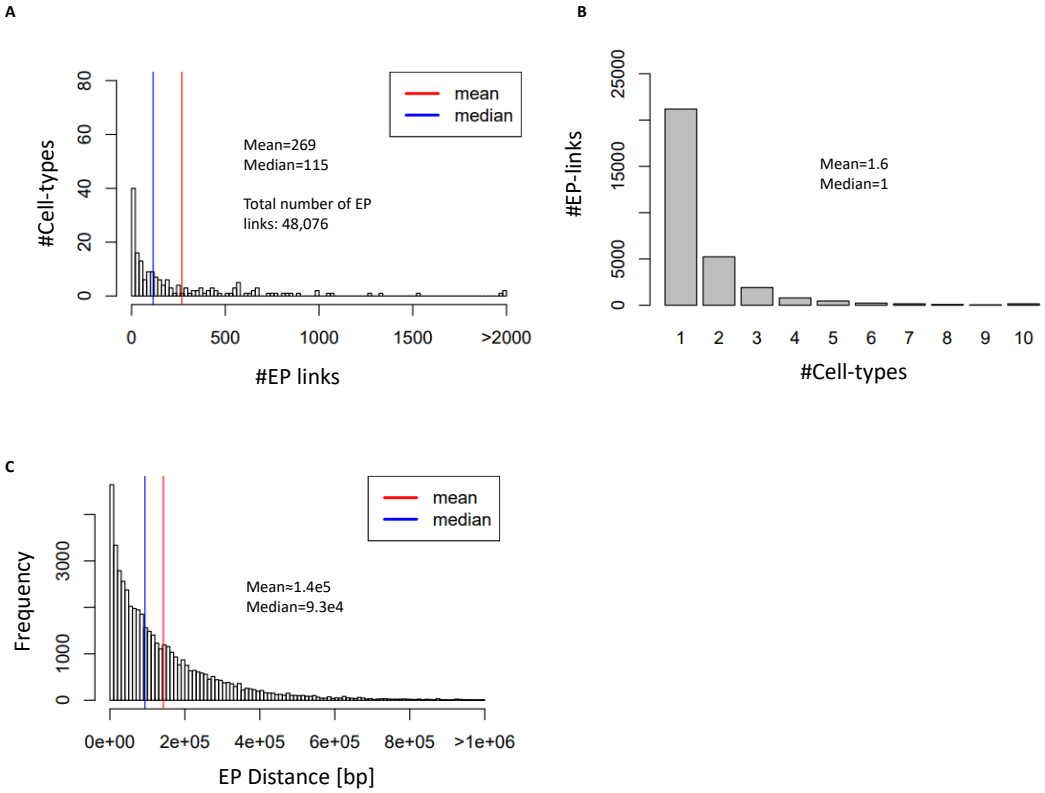


Figure S3.13. Properties of the ct-links obtained when CT-FOCS is applied on 179 FANTOM5 cell types with at least two replicates (solution sets CT-FOCS-2rep). (A) Number of predicted EP links per cell type (n=179). **(B)** Sharing of ct-links among cell types. **(C)** Enhancer-Promoter distances in predicted ct-links. EP distances were collapsed from all cell types, i.e., repeated EP links may be included.

6.3. Supplement 3: Enhancer and silencer inference

Selection of epigenetic datasets

To find the epigenetic features to be used in our model, we implemented a DeepTACT classification model with different inputs. Each input was composed of the sequence alone or the sequence with additional epigenetic signals (**Supplementary Table 4.1**). We analyzed the same GM12878 dataset described in the Methods section and measured the classification accuracies for silencers and enhancers. We evaluated the classification quality using the AUROC and AUPRC indexes.

Table S4.1. Summary of accuracies for different combinations of epigenetic datasets

Combination	AUROC enhancer	AUPRC enhancer	AUROC silencer	AUPRC silencer
Sequence (Seq.)	0.76	0.29	0.94	0.76
Seq. + Methylation (Meth.)	0.81	0.40	0.94	0.83
Seq. + H3K27ac	0.87	0.49	0.95	0.81
Seq. + H3K27me3	0.81	0.37	0.95	0.84
Seq. + H3K4me1	0.90	0.47	0.96	0.82
Seq. + Meth. + H3K27ac	0.90	0.54	0.96	0.85
Seq. + Meth. + H3K27me3	0.85	0.46	0.96	0.84
Seq. + Meth. + H3K4me1	0.91	0.52	0.96	0.83
Seq. + H3K27ac + H3K27me3	0.89	0.52	0.95	0.84
Seq. + H3K27ac + H3K4me1	0.90	0.52	0.96	0.84
Seq. + H3K27me3 + H3K4me1	0.90	0.49	0.96	0.84
Seq. + Meth. + H3K27ac + H3K4me1	0.92	0.54	0.96	0.85

Supplemental Figures

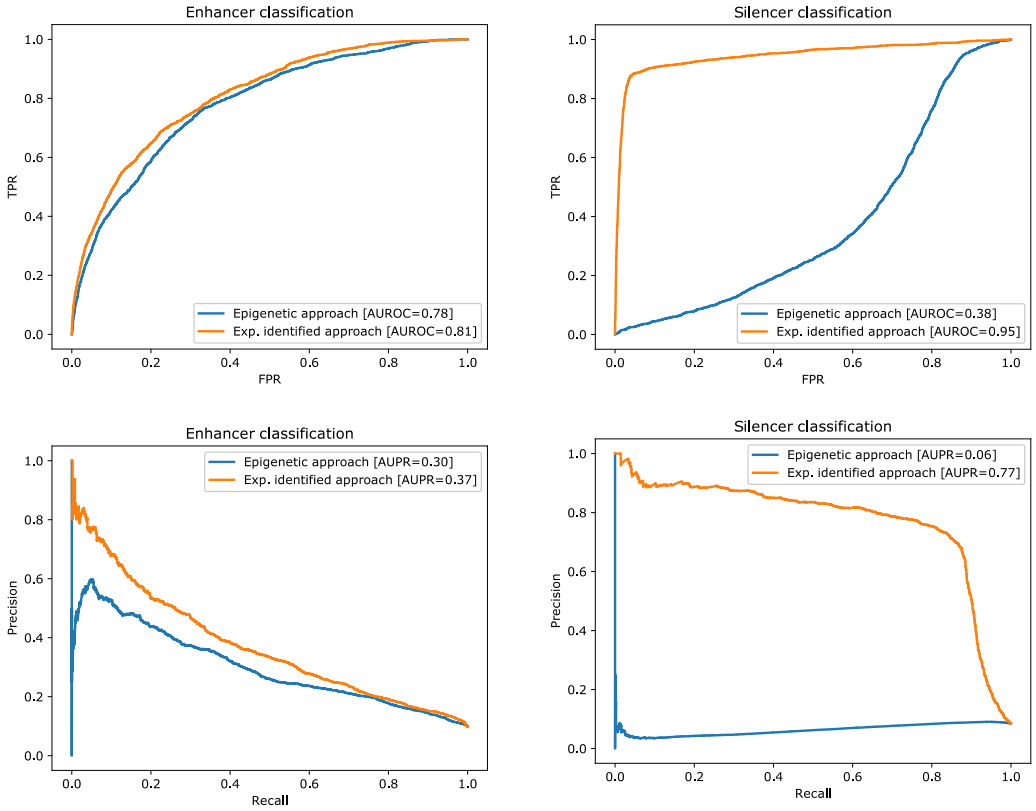


Figure S4.1. Performance of CNN models in predicting experimentally identified regulatory elements in GM12878. We trained the CNN models twice: first using putative enhancers and silencers defined based on epigenetic marks, and second, using the elements experimentally identified by ATAC-STARR-Seq in this cell line. Both approaches were tested for their ability to correctly detect experimentally identified elements.

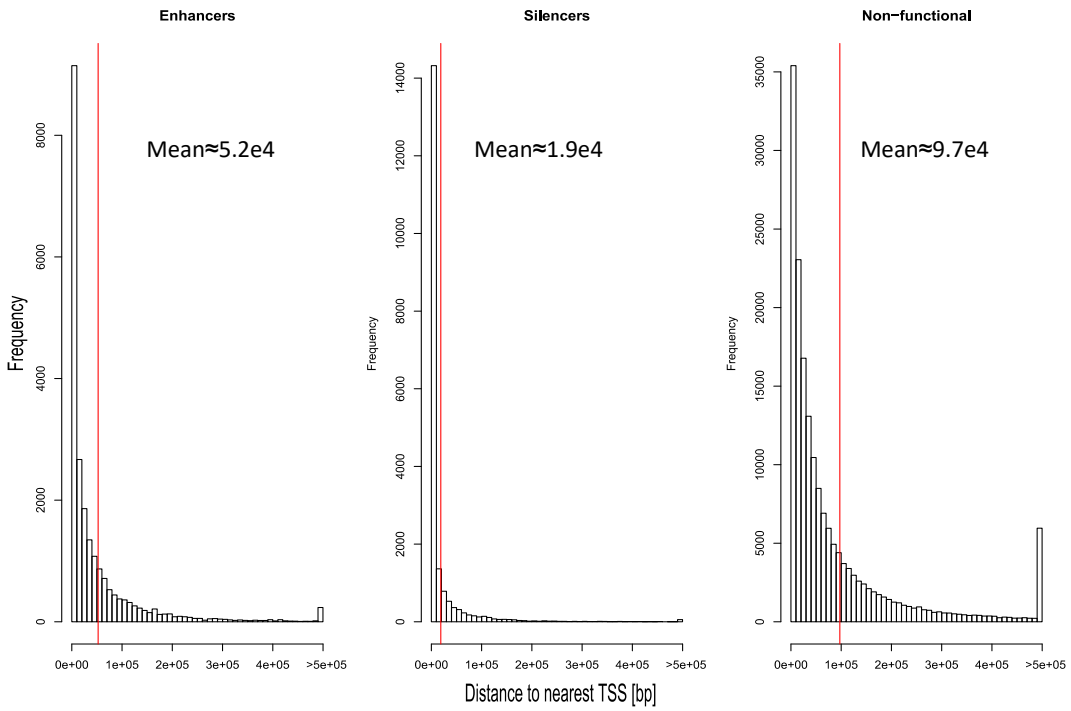


Figure S4.2. Distance to nearest TSS. Enhancers, silencers and non-functional distances to the TSS of their nearest gene. Vertical red line denotes the mean distance.

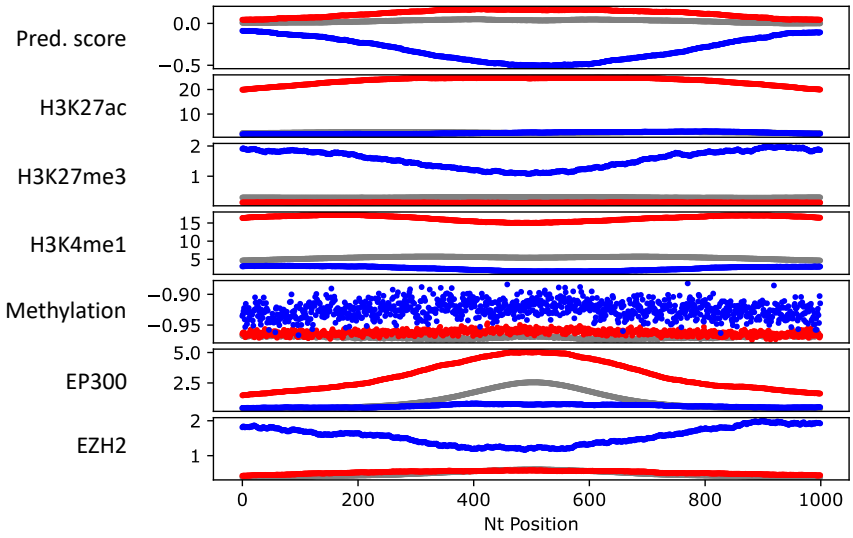


Figure S4.3. Summary of epigenetic markers in the exploratory set. Top to bottom: predicted scores (output of Step1 – the regression model), H3K27ac, H3K27me3, H3K27me1, Methylation, EP300 and EZH2. Predicted enhancers, silencers and nonfunctional are marked by red, blue and grey colors, respectively. In each predicted class and each track, the average signal per position in the 1kb sequences is shown. In b, the grey curve overlaps the blue curve for H3K27ac and the red curve for the EZH2.

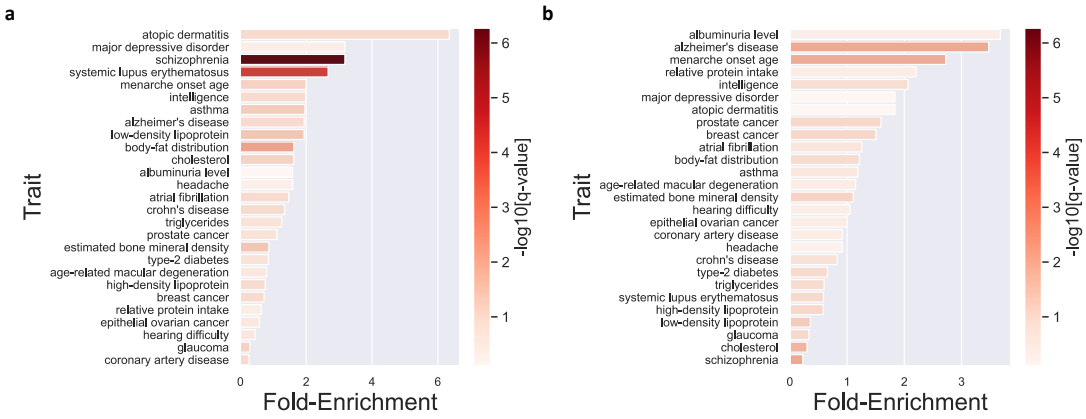


Figure S4.4. GWAS enrichment in the experimentally identified (a) enhancers and **(b)** silencers.

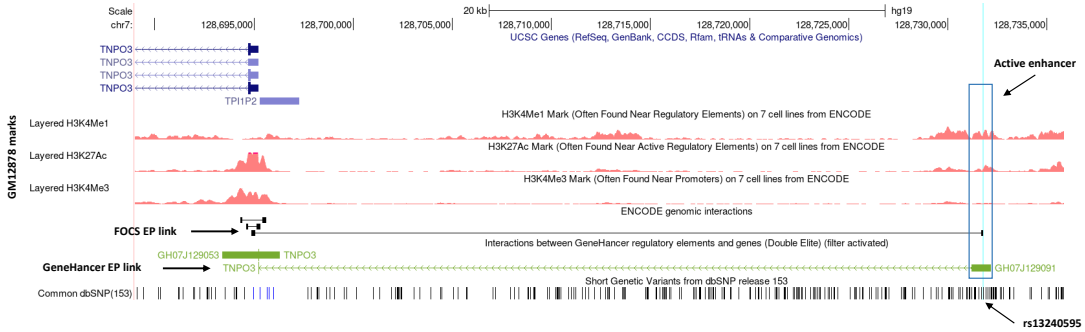
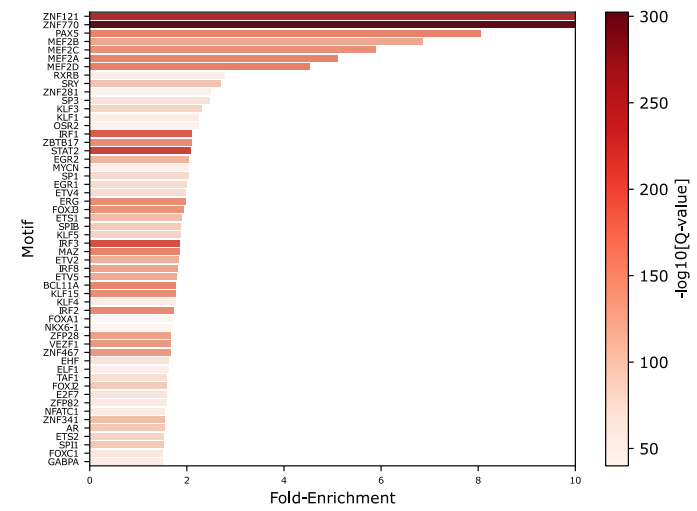


Figure S4.5. UCSC genome browser tracks of SLE risk SNP, rs13240595 (marked in arrow), falling within a predicted active enhancer linked to TNPO3 gene.

a



b

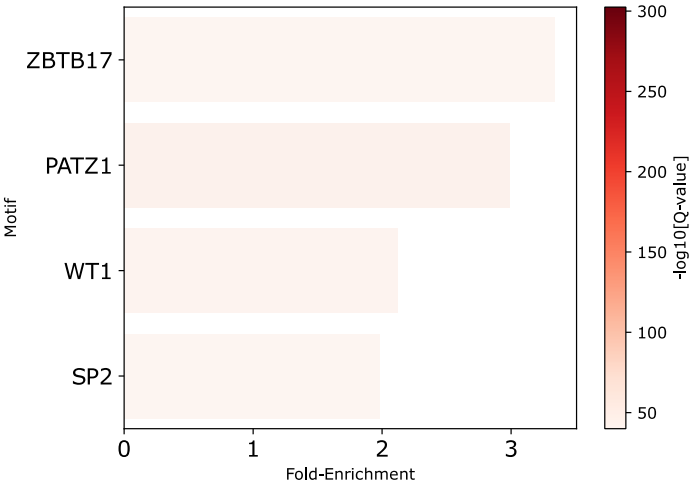


Figure S4.6. Motif enrichment. Enriched motifs within predicted (a) enhancers and (b) silencers.

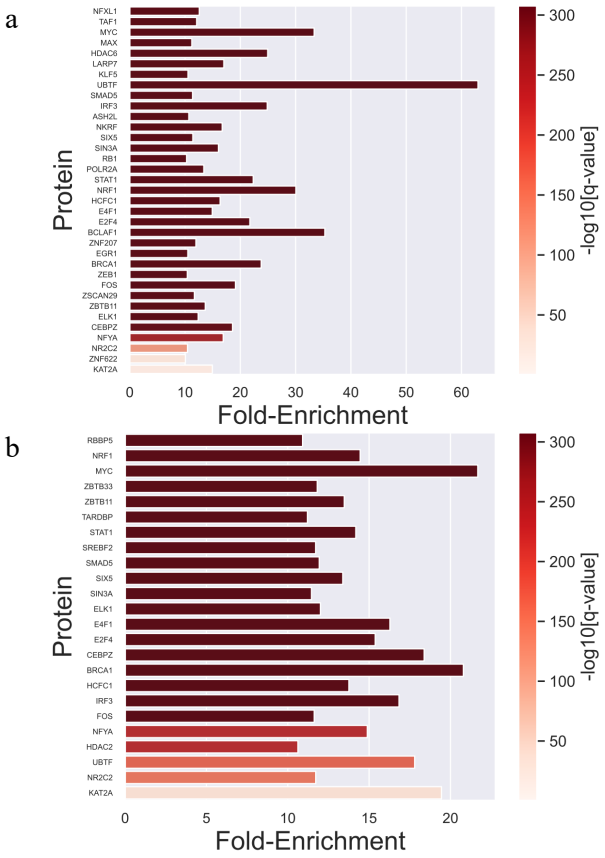


Figure S4.7. ChIP-seq enrichment in the experimentally identified sets of **(a)** Silencers and **(b)** Enhancers.

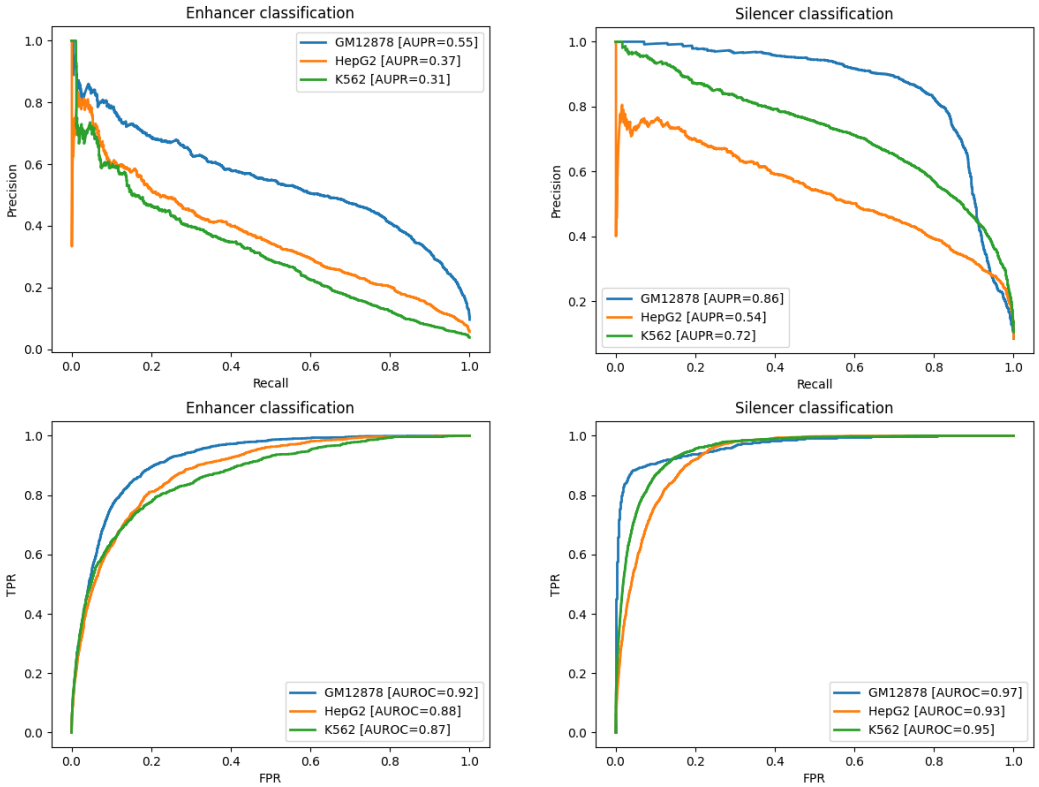


Figure S4.8. Performance of transfer learning of GM12878 model to K562 and HepG2 cell lines. Original GM18278 performances are in blue curves.

7. References

1. Alberts BJ, Lewis A, Raff J, others. Molecular biology of the cell. 2008.

2. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014;15:272–86.

3. Segert JA, Gisselbrecht SS, Bulyk ML. Transcriptional silencers: Driving gene expression with the brakes on. Trends in Genetics. 2021;37:514–27.

4. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nat Genet. 2008;40:340–5.

5. Matlin AJ, Clark F, Smith CWJ. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 2005;6:386–98.

6. Huang D, Petrykowska HM, Miller BF, Elnitski L, Ovcharenko I. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. Genome Res. 2019;29:657–67.

7. Zhang Y, See YX, Tergaonkar V, Fullwood MJ. Long-Distance Repression by Human Silencers: Chromatin Interactions and Phase Separation in Silencers. Cells. 2022;11:1–17.

8. Cai Y, Zhang Y, Loh YP, Tng JQ, Lim MC, Cao Z, et al. H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. Nat Commun. 2021;12.

9. Olins AL, Olins DE. Spheroid chromatin units (v bodies). Science (1979). 1974;183:330–2.

10. Felsenfeld G, Groudine M. Controlling the double helix. Nature. 2003;421:448–53.

11. Han M, Grunstein M. Nucleosome loss activates yeast downstream promoters in vivo. Cell. 1988;55:1137–45.

12. Lorch Y, LaPointe JW, Kornberg RD. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. Cell. 1987;49:203–10.

13. Yun M, Wu J, Workman JL, Li B. Readers of histone modifications. Cell Res [Internet]. 2011;21:564–78. Available from: https://doi.org/10.1038/cr.2011.42

14. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nat Rev Genet. 2013;14:288–95.

15. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. Science (1979). 1998;281:60–3.

16. Hartenstein V, Jan YN. Studying Drosophila embryogenesis with P-lacZ enhancer trap lines. Roux's archives of developmental biology. 1992;201:194–220.

17. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132:311–22.

18. Janky R, Verfaillie A, Imrichová H, van de Sande B, Standaert L, Christiaens V, et al. iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. PLoS Comput Biol. 2014;10.

19. Visel A, Bristow J, Pennacchio LA. Enhancer identification through comparative genomics. Semin Cell Dev Biol. 2007. p. 140–52.

20. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, et al. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proceedings of the National Academy of Sciences. 2002;99:757–62.

21. Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, et al. Identification of active transcriptional regulatory elements from GRO-seq data. Nat Methods. 2015;12:433–8.

22. Suryamohan K, Halfon MS. Identifying transcriptional cis-regulatory modules in animal genomes. Wiley Interdiscip Rev Dev Biol. 2015;4:59–84.

23. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.

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

25. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Kheradpour P, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317.

26. Perry MW, Boettiger AN, Bothma JP, Levine M. Shadow enhancers foster robustness of Drosophila gastrulation. Current Biology. 2010;20:1562–7.

27. Ron G, Moran D, Kaplan T. Promoter-Enhancer Interactions Identified from Hi-C Data using Probabilistic Models and Hierarchical Topological Domains. Nat Commun. 2017;8:2237.

28. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503:290–4.

29. Zhang Y, Wong C-H, Birnbaum RY, Li G, Favaro R, Ngan CY, et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature. 2013;504:306–10.

30. Sawada S, Scarborough JD, Killeen N, Littman DR. A lineage-specific transcriptional silencer regulates CD4 gene expression during T lymphocyte development. Cell. 1994;77:917–29.

31. Siu G, Wurster AL, Duncan DD, Soliman TM, Hedrick SM. A transcriptional silencer controls the developmental expression of the CD4 gene. EMBO J. 1994;13:3570–9.

32. Pang B, Snyder MP. Systematic identification of silencers in human cells. Nat Genet. 2020;52:254–63.

33. Ngan CY, Wong CH, Tjong H, Wang W, Goldfeder RL, Choi C, et al. Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. Nat Genet. 2020;52:264–72.

34. Doni Jayavelu N, Jajodia A, Mishra A, Hawkins RD. Candidate silencer elements for the human and mouse genomes. Nat Commun. 2020;11:1–15.

35. Hansen TJ, Hodges E. ATAC-STARR-seq reveals transcription factor--bound activators and silencers within chromatin-accessible regions of the human genome. Genome Res. 2022;32:1529–41.

36. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences. 1977;74:5463–7.

37. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype--phenotype interactions. Nat Rev Genet. 2015;16:85–97.

38. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science (1979). 2007;316:1497–502.

39. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.

40. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.

41. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 2007;17:877–85.

42. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10:1213.

43. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.

44. Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, et al. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. 2013;23:777–88.

45. Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res. 2019;47:e60–e60.

46. Zheng L, Liu L, Zhu W, Ding Y, Wu F. Predicting enhancer-promoter interaction based on epigenomic signals. Front Genet. 2023;14.

47. Ma S, Zhang Y. Profiling chromatin regulatory landscape: Insights into the development of ChIP-seq and ATAC-seq. Molecular biomedicine. 2020;1:1–13.

48. Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, et al. Global reference mapping of human transcription factor footprints. Nature. 2020;583:729–36.

49. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome Biol. 2019;20:1–21.

50. Hait TA, Elkon R, Shamir R. CT-FOCS: a novel method for inferring cell type-specific enhancer–promoter maps. Nucleic Acids Res. 2022;gkac048.

51. Park PJ. ChIP--seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10:669–80.

52. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010;2010:pdb--prot5384.

53. Hashimshony T, Zhang J, Keshet I, Bustin M, Cedar H. The role of DNA methylation in setting up chromatin structure during development. Nat Genet. 2003;34:187–92.

54. Gopalakrishnan S, Van Emburgh BO, Robertson KD. DNA methylation in development and human disease. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2008;647:30–8.

55. Cho J-W, Shim HS, Lee CY, Park SY, Hong MH, Lee I, et al. The importance of enhancer methylation for epigenetic regulation of tumorigenesis in squamous lung cancer. Exp Mol Med. 2022;54:12–22.

56. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science (1979). 2017;356:eaaj2239.

57. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacology [Internet]. 2013;38:23–38. Available from: https://doi.org/10.1038/npp.2012.112

58. Kunde-Ramamoorthy G, Coarfa C, Laritsky E, Kessler NJ, Harris RA, Xu M, et al. Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. Nucleic Acids Res. 2014;42:e43--e43.

59. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed Y Bin, et al. An oestrogen-receptor-α-bound human chromatin interactome. Nature. 2009;462:58.

60. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods. 2016;13:919.

61. De Wit E, De Laat W. A decade of 3C technologies: insights into nuclear organization. Genes Dev. 2012;26:11–24.

62. Han J, Zhang Z, Wang K. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. Mol Cytogenet [Internet]. 2018;11:21. Available from: https://doi.org/10.1186/s13039-018-0368-2

63. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102:15545–50.

64. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.

65. Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1:59.

66. Dunn OJ. Multiple comparisons among means. J Am Stat Assoc. 1961;56:52–64.

67. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995;289–300.

68. Linhart C, Halperin Y, Shamir R. Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. Genome Res. 2008;18:1180–9.

69. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202–8.

70. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017–8.

71. Kulakovskiy I V, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2017;46:D252–9.

72. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2022;50:D165--D173.

73. Lindstrom MJ, Bates DM. Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. J Am Stat Assoc. 1988;83:1014–22.

74. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.

75. Chollet F. Deep learning with Python. Simon and Schuster; 2021.

76. Stevens E, Antiga L, Viehmann T. Deep learning with PyTorch. Manning Publications; 2020.

77. Xing W, Du D. Dropout prediction in MOOCs: Using deep learning for personalized intervention. Journal of Educational Computing Research. 2019;57:547–70.

78. Gupta A, Anpalagan A, Guan L, Khwaja AS. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. Array. 2021;10:100057.

79. Tahir A, Munawar HS, Akram J, Adil M, Ali S, Kouzani AZ, et al. Automatic target detection from satellite imagery using machine learning. Sensors. 2022;22:1147.

80. Xue P, Wang J, Qin D, Yan H, Qu Y, Seery S, et al. Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis. NPJ Digit Med. 2022;5:19.

81. Kraljevic Z, Shek A, Bean D, Bendayan R, Teo J, Dobson R. MedGPT: Medical concept prediction from clinical narratives. arXiv preprint arXiv:210703134. 2021;

82. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51:12–8.

83. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60:84–90.

84. Su C, Xu Z, Pathak J, Wang F. Deep learning in mental health outcome research: a scoping review. Transl Psychiatry. 2020;10:116.

85. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16:321–32.

86. Luo Z, Zhang J, Fei J, Ke S. Deep learning modeling m6A deposition reveals the importance of downstream cis-element sequences. Nat Commun. 2022;13:1–16.

87. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. Nat Methods. 2020;17:1111–7.

88. Routhier E, Pierre E, Khodabandelou G, Mozziconacci J. Genome-wide prediction of DNA mutation effect on nucleosome positions for yeast synthetic genomics. Genome Res. 2021;31:317–26.

89. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. International conference on machine learning. 2017. p. 3319–28.

90. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. International conference on machine learning. 2017. p. 3145–53.

91. Park Y, Kellis M. Deep learning for regulatory genomics. Nat Biotechnol. 2015;33:825–6.

92. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26:990–9.

93. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44:e107--e107.

94. Zeng H, Edwards MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA--protein binding. Bioinformatics. 2016;32:i121--i127.

95. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33:831–8.

96. Liu F, Li H, Ren C, Bo X, Shu W. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. Sci Rep. 2016;6:28517.

97. Kleftogiannis D, Kalnis P, Bajic VB. DEEP: a general computational framework for predicting enhancers. Nucleic Acids Res. 2015;43:e6--e6.

98. Li Y, Shi W, Wasserman WW. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics. 2018;19:1–14.

99. Wang Y, Liu T, Xu D, Shi H, Zhang C, Mo Y-Y, et al. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. Sci Rep. 2016;6:1–15.

100. Hait TA. Using large-scale high-throughput data for enhancer-promoter network inference. Tel-Aviv University; 2017.

101. Hait TA, Amar D, Shamir R, Elkon R. FOCS : a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. Genome Biol. 2018;19:59.

102. Lawless JF. Negative binomial and mixed Poisson regression. The Canadian Journal of Statisitcs. 1987;15:209–25.

103. Greene WH. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. 1994;

104. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29:1165–88.

105. Weintraub AS, Li CH, Zamudio A V., Sigova AA, Hannett NM, Day DS, et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell. 2017;171:1573-1579.e28.

106. Consortium Gte, others. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (1979). 2015;348:648–60.

107. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.

108. Polak P, Lawrence MS, Haugen E, Stoletzki N, Stojanov P, Thurman RE, et al. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. Nat Biotechnol. 2014;32:71.

109. Quinlan AR, Hall IM. BEDTools : a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

110. Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

111. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. Nat Biotechnol. 2010;28:1045–8.

112. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012;489:83–90.

113. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523:212.

114. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22:1760–74.

115. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. R package version 2.28.0; 2016.

116. Xie X, Ma W, Songyang Z, Luo Z, Huang J, Dai Z, et al. Original article CCSI : a database providing chromatin – chromatin spatial interaction information. 2016;1–7.

117. R Core Team. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2018.

118. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol. 2013;9:e1003118.

119. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25:1841–2.

120. Harmston, N., Ing-Simmons, E., Perry, M., et al. GenomicInteractions: An R/Bioconductor package for manipulating and investigating chromatin interaction data. BMC Genomics. 2015;16:963.

121. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth. New York; 2002.

122. Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. J Stat Softw. 2008;27:1–25.

123. Wickham H. ggplot2: elegant graphics for data analysis. Springer-Verlag New York; 2016.

124. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots: various R programming tools for plotting data. 2016.

125. Kumasaka N, Knights AJ, Gaffney DJ. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. Nat Genet. 2019;51:128.

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

127. Rajarajan P, Borrman T, Liao W, Schrode N, Flaherty E, Casiño C, et al. Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. Science. 2018;362:eaat4311.

128. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell. 2019;176:377–90.

129. Krijger PHL, de Laat W. Regulation of disease-associated gene expression in the 3D genome. Nat Rev Mol Cell Biol. 2016;17:771–82.

130. Pombo A, Dillon N. Three-dimensional genome architecture: players and mechanisms. Nat Rev Mol Cell Biol. 2015;16:245–57.

131. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014;159:1665–80.

132. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell. 2016;167:1369–84.

133. Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, et al. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. Nat Genet. 2017;201:7.

134. Whalen S, Truty RM, Pollard KS. Enhancer – promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet. 2016;48:488.

135. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer--promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51:1664–9.

136. Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, et al. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021;593:238–43.

137. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell. 2015;163:1611–27.

138. Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet. 2019;51:1442–9.

139. Olivares-Chauvet P, Mukamel Z, Lifshitz A, Schwartzman O, Elkayam NO, Lubling Y, et al. Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature. 2016;540:296.

140. Song W, Sharan R, Ovcharenko I. The first enhancer in an enhancer chain safeguards subsequent enhancer-promoter contacts from a distance. Genome Biol. 2019;20:197.

141. Kumasaka N, Knights AJ, Gaffney DJ. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. Nat Genet. 2019;51:128.

142. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38:576–89.

143. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods. 2009;6:283.

144. Nechanitzky R, Akbas D, Scherer S, Györy I, Hoyler T, Ramamoorthy S, et al. Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells. Nat Immunol. 2013;14:867.

145. Wang H, Lee CH, Qi C, Tailor P, Feng J, Abbasi S, et al. IRF8 regulates B-cell lineage specification, commitment, and differentiation. Blood, The Journal of the American Society of Hematology. 2008;112:4028–38.

146. Zhang K, Li N, Ainsworth RI, Wang W. Systematic identification of protein combinations mediating chromatin looping. Nat Commun. 2016;7:12249.

147. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J Exp Soc Psychol. 2013;49:764–6.

148. Hou C, Li L, Qin ZS, Corces VG. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Mol Cell. 2012;48:471–84.

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

150. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148:458–72.

151. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.

152. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015;47:598.

153. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012;148:84–98.

154. Csárdi G, Nepusz T. The igraph software package for complex network research. InterJournal, complex systems. 2006;1695:1–9.

155. Xi W, Beer MA. Local epigenomic state cannot discriminate interacting and non-interacting enhancer-promoter pairs with high accuracy. PLoS Comput Biol. 2018;14:e1006625.

156. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw. 2011;42:1–28.

157. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing. Vienna, Austria; 2020 [cited 2019 Mar 1]. p. https://www.R-project.org. Available from: http://www.r-project.org

158. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: Linear and Nonlinear Mixed Effects Models Description. 2021.

159. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J. 2017;9:378–400.

160. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32:2847–9.

161. Huang D, Ovcharenko I. Enhancer-silencer transitions in the human genome. Genome Res. 2022;32:437–48.

162. Wang J, Yu X, Gong W, Liu X, Park K-S, Ma A, et al. EZH2 noncanonically binds cMyc and p300 through a cryptic transactivation domain to mediate gene activation and promote oncogenesis. Nat Cell Biol. 2022;24:384–99.

163. Charlet J, Duymich CE, Lay FD, Mundbjerg K, Sørensen KD, Liang G, et al. Bivalent regions of cytosine methylation and H3K27 acetylation suggest an active role for DNA methylation at enhancers. Mol Cell. 2016;62:422–31.

164. Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, et al. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. 2012;22:2399–408.

165. Dekker J. GC-and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. Genome Biol. 2007;8:1–14.

166. Shrikumar A, Tian K, Avsec Ž, Shcherbina A, Banerjee A, Sharmin M, et al. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5. arXiv preprint arXiv:181100416. 2018;

167. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8:1–9.

168. Groenewoud D, Shye A, Elkon R. Incorporating regulatory interactions into gene-set analyses for GWAS data: A controlled analysis with the MAGMA tool. PLoS Comput Biol. 2022;18:e1009908.

169. van Mierlo HC, Broen JCA, Kahn RS, de Witte LD. B-cells and schizophrenia: A promising link or a finding lost in translation? Brain Behav Immun. 2019;81:52–62.

170. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017;2017.

171. Lu X, Chen X, Forney C, Donmez O, Miller D, Parameswaran S, et al. Global discovery of lupus genetic risk variant allelic enhancer activity. Nat Commun. 2021;12:1611.

172. Kottyan LC, Zoller EE, Bene J, Lu X, Kelly JA, Rupert AM, et al. The IRF5--TNPO3 association with systemic lupus erythematosus has two components that other autoimmune disorders variably share. Hum Mol Genet. 2015;24:582–96.

173. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics. 2016;54:1–30.

174. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. Nature. 2011;469:343–9.

175. Grozinger CM, Hassig CA, Schreiber SL. Three proteins define a class of human histone deacetylases related to yeast Hda1p. Proceedings of the National Academy of Sciences. 1999;96:4868–73.

176. Pérez-Olivares M, Trento A, Rodriguez-Acebes S, González-Acosta D, Fernández-Antorán D, Román-García S, et al. Functional interplay between c-Myc and Max in B lymphocyte differentiation. EMBO Rep. 2018;19:e45770.

177. Oganesyan G, Saha SK, Pietras EM, Guo B, Miyahira AK, Zarnegar B, et al. IRF3-dependent type I interferon response in B cells regulates CpG-mediated antibody production. Journal of Biological Chemistry. 2008;283:802–8.

178. Bailey TL, Grant CE. SEA: Simple Enrichment Analysis of motifs. bioRxiv. 2021;

179. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends in Genetics. 2013;29:569–74.

180. Melton C, Reuter JA, Spacek D V, Snyder M. Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat Genet. 2015;47:710–6.

181. Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet. 2014;46:1160–5.

182. Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proceedings of the National Academy of Sciences. 2017;114:E4914–23.

183. Pang B, Snyder MP. Systematic identification of silencers in human cells. Nat Genet. 2020;52:254–63.

184. Eberwine J, Sul J-Y, Bartfai T, Kim J. The promise of single-cell sequencing. Nat Methods. 2014;11:25–7.

185. Saliba A-E, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. 2014;42:8845–60.

186. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523:486–90.

187. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Mol Cell. 2018;71:858–71.

188. Chen H, Albergante L, Hsu JY, Lareau CA, Lo Bosco G, Guan J, et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. Nat Commun. 2019;10:1903.

189. West J, Ventura D, Warnick S. Spring research presentation: A theoretical foundation for inductive transfer. Brigham Young University, College of Physical and Mathematical Sciences. 2007;1.

190. Bolger AM, Lohse M, Usadel B. Genome analysis Trimmomatic : a flexible trimmer for Illumina sequence data. 2014;30:2114–20.

191. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. 2012;9:357–60.

192. Chae M, Danko CG, Kraus WL. groHMM : a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. BMC Bioinformatics. 2015;9–11.

193. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465:182–7.

194. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1.

195. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. 2011;

196. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.

197. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

198. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40:4288–97.

199. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.

200. Hait TA, Elkon R, Shamir R. Inferring transcriptional activation and repression activity maps in single-nucleotide resolution using deep-learning. Res Sq. 2023;

לאחרונה זיהה ניסוי מסוג ATAC-STARR-seq בצורה ניסויית עשרות אלפי מעצמים ומשתיקים בתאי B (תאים מסוג GM12878). על סמך הניסוי, בכל בסיס ברצפים שנתגלו חושב ציון תרומה חיובית או שלילית לשיעתוק גן המטרה שלהם. מטרתנו הייתה לפתח באמצעות המידע הנ"ל שיטת סיווג אמינה יותר לניבוי מעצמים ומשתיקים. כדי לזהות מה מגדיר מבחינה אפיגנטית משתיקים פונקציונאליים הפעלנו שיטות לזיהוי תכונות חשובות (Feature importance) על המודל המאומן.

ראשית, חקרנו את השאלה הבאה, האם מודל למידה עמוקה שאומן על רצפי דנ"א שסומנו כמעצמים ומשתיקים באמצעות מעצמים ומשתיקים שאומתו ניסויית יהיה מדוייק יותר או פחות לעומת אותו מודל למידה עמוקה שאומן על רצפי דנ"א שסומנו כמעצמים וכמשתיקים באמצעות מידע אפיגנטי בלבד (ללא אישוש נסיוני)? התשובה שלנו הראתה יתרון ברור לשימוש במידע הניסויי מסוג ATAC-STARR-seq לזיהוי משתיקים.

בנוסף, שיטות קודמות שניבאו משתיקים השתמשו רק ברצף הדנ"א כקלט. שאלתנו השנייה בעבודה זו הייתה האם הוספת המידע האפיגנטי כקלט בנוסף לרצף הדנ"א תשפר את ביצועי הניבוי? גם כאן התשובה הייתה חיובית.

לבסוף, בדקנו האם ניסוי ה-ATAC-STARR-seq שתואר לעיל החמיץ מעצמים ומשתיקים אמיתיים מסויימים בתאי ה-B שנבחנו. לצורך כך, אימנו מודל למידה עמוקה על מעצמים ומשתיקים שנתגלו ניסויית וחזינו 3,752 מעצמים ו-518 משתיקים חדשים מתוך קבוצה של רצפים גנומיים שלא חפפו שום מעצם או משתיק שנתגלו ניסויית בתאים אלו. קיבלנו תמיכה במהימנות הרצפים החדשים שנתגלו באמצעות העשרות ביולוגיות של TFs ושל וריאנטים בגנום הקשורים למחלות (GWAS).

הכלי זמין ב: https://github.com/Shamir-Lab/EnhancerSilencerDL. העבודה כעת בביקורת עמיתים ופורסמה כ-preprint בתור:

Hait TA, Elkon R, Shamir R**. Inferring single nucleotide activation and repression maps using deep-learning.** *Research Square*. 23 August 2023, https://doi.org/10.21203/rs.3.rs-3270775/v1

בפרק 3 אני מציג את שיטתנו למציאת אינטראקציות מעצם-מקדם הספציפיות למספר קטן של סוגי תאים. שיטת FOCS שהוצגה בפרק 2 חוזה אינטראקציות מעצם-מקדם גלובליות על בסיס קורלציית תבניות הפעילות על פני דגימות רבות המכסות מאות סוג תאים שונים. אי לכך, האינטראקציות שנחזו הן גלובליות ואינן בהכרח ספציפיות למספר קטן של סוגי תאים. אתגר מרכזי הוא לזהות מי מבין האינטראקציות הגלובליות שנחזו הן באמת פונקציונאליות וספציפיות למספר קטן של תאים (לאינטראקציות כאלו קראנו ct-links). לצורך כך, פיתחנו את שיטת CT-FOCS (cell-type-FOCS), שיטה המבוססת על linear mixed effect model (LMM) שחוזה אינטראקציית מעצם-מקדם המתקיימות במספר קטן של סוגי תאים ספציפיים על בסיס מספר חזרות של דגימות לכל סוג תא. יישמנו את CT-FOCS על נתונים מסוג CAGE מהמאגר FANTOM5 ומסוג DNase-seq מהמאגרים ENCODE ו-Roadmap Epigenomics, וחזינו מספר כולל של 229,518 ct-links ל-651 תאים שונים.

השווינו את ביצועי CT-FOCS לעומת שיטות קיימות על ידי בדיקת ה-ct-links אל מול אינטראקציות שנמצאו בצורה ניסויית ועל ידי בדיקת ספציפיות תאית של ביטוי גני המטרה של המעצמים. הדרך הישירה לבדיקת נכונות ה-ct-links באמצעות אינטראקציות ניסיוניות (Loops או לולאות) היא לבדוק האם המעצם והמקדם חופפים את שני העוגנים (anchors) של אותה לולאה. יחד עם זאת, בהינתן שלולאות מעידות על קרבה מרחבית של עוגניהן, עוגנים חופפים מלולאות שונות עשויים להעיד גם כן על קרבה מרחבית של אותם עוגנים. לקיחה בחשבון של אותם עוגנים חופפים מלולאות שונות עשויה לעזור באימות נכונותן של ct-links הנפרשות במרחק לינארי של לא יותר מ-20kb, מרחק בו לא סביר למצוא לולאות מטכנולוגית ChIA-PET. לצורך כך, הגדרנו את סט הלולאות החופפות (Two-step connected loop set – TLS) כדי להרחיב את סט העוגנים שנחשבים קרובים מרחבית לאימות ct-links.

בחלק האחרון של עבודה זו שאלנו האם ה-ct-links שנחזו מגדירים בקרה תאית ספציפית על גנים. לצורך כך, מדדנו את הספציפיות של 402 פקטורי שיעתוק (Transcription Factors – TFs) ידועים בתוך המעצמים והמקדמים של ct-links. הראינו ש-ct-links שנחזו על ידי CT-FOCS מניבים TFs שהם מאוד ספציפיים לתאים ובצורה מובהקת לעומת שיטות קיימות. לפיכך, הראינו ש-CT-FOCS מסוגלת לחזות ct-links בעלי משמעות ביולוגית רלוונטית לבקרה תאית ספציפית של גנים.

הכלי זמין ב: http://acgt.cs.tau.ac.il/ct-focs/. העבודה פורסמה בתור:

Hait TA, Elkon R, Shamir R. **CT-FOCS: a novel method for inferring cell type-specific enhancer–promoter maps.** *Nucleic Acids Research*, Volume 50, Issue 10, 10 June 2022, Page e55, https://doi.org/10.1093/nar/gkac048.

בפרק 4 אני מציג את שיטתנו לניבוי אזורי בקרה לא מקודדים בגנום ובפרט לניבוי של משתיקים. אזורי בקרה המבקרים שיעתוק גנים כגון מעצמים ומקדמים נחקרו בצורה מעמיקה לאורך שני העשורים האחרונים. לעומת זאת, משתיקים, אשר מורידים או משתיקים שיעתוק של גן המטרה שלהם, קיבלו תשומת לב מועטה בעיקר בגלל שקשה לזהותם ניסויית.

שיטות סיווג שמטרתן לנבא מעצמים ומקדמים הספציפיים לתאים אומנו בעיקר על רצף הדנ"א ומידע אפיגנטי. מעצמים ומקדמים שהוגדרו כקבוצה החיובית עבור האימון, האימות והמבחן סומנו בעיקר באמצעות מידע אפיגנטי. לדוגמא, מעצמים פונקציונאליים ידועים כמסומנים באזורם בגנום באמצעות המודיפיקציות של היסטונים, H3K4me1 ו- H3K27ac. לעומתם, עדיין לא ברור מה מגדיר מבחינה אפיגנטית משתיקים פונקציונאליים.

איסוף מידע המכסה אלפי דגימות של סוגי תאים שונים מאחת או יותר מטכנולוגיות אלו הינו תהליך מורכב מאוד המכיל רמות גבוהות של רעש סטטיסטי, ומקשה על פירוש תוצאות הניסויים. בנוסף, לפעמים קיימת סתירה בין תוצאות מניסויים שונים שבדקו השערה ביולוגית דומה, כגון ניסויים שונים לגילוי משתיקים באותו סוג תא שלא הראו חפיפה בין המשתיקים שנמצאו בניסויים השונים [200]. כדי לטפל בבעיות המוזכרות לעיל, הדגש בתזה זו הוא פיתוח שיטות חישוביות לניתוח משולב של נתונים מהרבה ניסויים יחד. מטרתנו היא לספק לקהילה המדעית כלים חדשים לזיהוי רצפי בקרה והאינטראקציות ביניהם תוך שימוש במגוון רחב של ניסויים מטכנולוגיות שונות. השיטות שנבחנו בתזה זו נבחנו והוכחו כעדיפות על פני שיטות קיימות ואחרות. בכל יישום של שיטה אנו מציגים את הערך הביולוגי המוסף של הניתוח בעזרת ניתוח מקיף של הנתונים תוך כדי שימוש בידע ביולוגי מהספרות.

תוצאות

בפרק 2 אני מציג את עבודתנו הראשונה שעוסקת בחיזוי אינטראקציות מעצם-מקדם. דפוסי ביטוי גנים מבוקרים בעיקר על ידי פעילות של מעצמים אשר יוצרים קשרים פיזיים עם גן המטרה שלהם. בעוד שניתן למצוא אינטראקציות מעצם-מקדם בצורה ניסויית בכל תא, מספר הניסויים שנעשו בתחום זה הוא מועט. יחד עם זאת, מאגרים גדולים כגון ENCODE, Roadmap Epigenomics ו- FANTOM5 מספקים מקור מידע אפיגנטי נרחב המכסה מאות סוגי תאים (לדוגמא, ניסויים המודדים אזורים פתוחים בגנום כגון DNase-seq). מאגרים אלו יכולים לשמש לצורך פיתוח שיטות חישוביות למיפוי מעצמים לגן המטרה שלהם.

עבודה זו החלה בלימודי התואר השני שלי ונמשכה לתוך שנתי הראשונה בלימודי הדוקטורט. בלימודי התואר השני, פיתחנו שיטה חישובית הנקראת FOCS (*F*DR-Corrected *O*LS with *C*ross-validation and *S*hrinkage). שיטה זו חוזה אינטראקציות מעצם-מקדם על בסיס קורלציה בין תבניות הפעילות של המעצם והמקדם לאורך דגימות שמקורן ממאות סוגי תאים. FOCS חוזה אינטראקציות מעצם-מקדם גלובליות, כלומר אינטראקציות עם קורלציה גדולה ומובהקת בין המעצם ובין המקדם. בעבודת המוסמך יישמנו את השיטה על מאגר מידע קטן המכיל 246 דגימות המכסות 23 סוגי תאים מטכנולוגית GRO-seq שמזהה בצורה ניסויית מעצמים ומקדמים פונקציונאליים.

במהלך הדוקטורט, חקרתי את יעילות וביצועי שיטת FOCS על עוד 2,384 דגימות (פי עשרה מידע לעומת המידע שנבדק בלימודי התואר השני) מניסויי DNase-seq ו-CAGE שנאספו מ- ENCODE, Roadmap Epigenomics ו-FNATOM5. בנוסף, בדקנו את טיב האינטראקציות שנחזו על בסיס אינטראקציות ניסיוניות מטכנולוגית HiChIP (בנוסף לאימות מול ChIA-PET ו-eQTL מלימודי התואר השני). הראינו שביצועי FOCS טובים יותר לעומת שיטות קיימות וביצענו אנליזות נוספות על האינטראקציות שנחזו כדי להראות תובנות ביולוגיות רלוונטיות.

הכלי זמין ב: http://acgt.cs.tau.ac.il/focs. העבודה פורסמה בתור:

Hait TA, Amar D, Shamir R, Elkon R**. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map.** *Genome Biol* 2018, 19:56. https://pubmed.ncbi.nlm.nih.gov/29716618/

תקציר

רקע כללי

האורגניזמים החיים מורכבים מתאים המשמשים כיחידות הבסיסיות של החיים. תחום הביולוגיה התאית מתמקד בפיענוח המבנה, הפונקציה וההתנהגות של התאים. הבקרה על התפתחות ופונקציה תאית נשלטת על ידי הדנ"א. הדנ"א מתחלק לדנ"א מקודד ודנ"א לא מקודד. דנ"א מקודד מוגדר כרצפים המתורגמים לחלבונים. רצפים אלו נקראים גנים והם מהווים פחות מ-2% מכלל הגנום. הדנ"א הלא מקודד (כ-98% מכלל הגנום) מוגדר כרצפים שאינם מקודדים לחלבונים. בדנ"א הלא מקודד ישנם רצפי בקרה שתפקידם הוא בקרה על רמות השיעתוק של גנים. השכיחים והנחקרים ביותר שברצפים אלו הם המקדמים (promoters), הממוקמים בקרבת נקודת תחילת השיעתוק של הגנים, והמעצמים (enhancers), הממוקמים ברובם, על הרצף הלינארי של הדנ"א, הרחק מהגנים אותם הם מבקרים. מטרת שני סוגים אלו היא לקדם ולהגביר שיעתוק של גנים. קבוצה פחות נפוצה ונחקרת של רצפי בקרת גנים הינה משתיקים (silencers) שתפקידה הפחתה ואף הפסקה של שיעתוק גנים. מקדמים, מעצמים ומשתיקים יוצרים אינטראקציות מרחביות בגנום על מנת לבקר שיעתוק גנים. אינטראקציות אלו יכולות להיות שונות בין סוגי תאים שונים בגופנו ולהגדיר תכונות ספציפיות לתאים.

השוואה בין אוכלוסיות שונות של אנשים באזורי הדנ"א המקודדים והלא מקודדים, כגון מעצמים ומשתיקים, יכולה לתרום לגילוי שינויים ברמת רצף הדנ"א שמשפיעים על הבקרה. תשומת לב מיוחדת ניתנה לגילוי שינויים תורשתיים ברמת הבסיס הבודד ברצף (SNPs) באוכלוסייה. שינויים אלו נפוצים יותר באזורים הלא מקודדים מאשר במקודדים [4]. לשינויים גנטיים אלו יש פוטנציאל להשפיע על המועדות של פרט באוכלוסייה למחלות ותגובתו לפתוגנים, תרכובות כימיות ועוד. על ידי הבנה של התהליכים הגורמים לרצפי בקרה לשלוט בגנים וברמותיהם באופן ספציפי לתאים, נוכל למצוא קשרים סטטיסטיים בין השינויים הגנטיים באזורי הבקרה ובין רמת השיעתוק של גנים. תובנות אלו תוכלנה להוביל להבנה מעמיקה יותר של הגורמים הגנטיים הגורמים לסיכון מוגבר למחלות.

טכנולוגיות ניסיוניות רבות פותחו על מנת לעזור לחוקר לאתר רצפי בקרת גנים ולמצוא אינטראקציות מרחביות ביניהם. יחד עם זאת, מיעוט מספר הניסויים שבוצעו על מנת לאתר אינטראקציות מרחביות בין רצפי הבקרה דורש פיתוח שיטות על מנת לאתרן בצורה חישובית על בסיס מידע מטכנולוגיות לא מרחביות. כמו כן, זיהוי משתיקים הוא משימה קשה משום שקבוצה זו קשה לזיהוי בצורה ניסויית בהשוואה למקדמים ומעצמים. קיימות טכנולוגיות רבות ומגוונות המודדות מאפיינים אפיגנטיים שונים של הגנום כגון רמות מתילציה, אתרי קשירה של חלבונים לרצפי הבקרה, מדידת מודיפיקציות על גבי נוקלאוזומים סביב רצפי הבקרה, מציאת רצפי בקרה באזורים פתוחים בגנום שלא מאוכלסים על ידי נוקלאוזומים ועוד. הודות לירידה במחירי הניסויים, מספר הניסויים שנעשו מטכנולוגיות שונות מכסה מאות סוגי תאים שונים תחת תנאים שונים, דבר שהוביל ליצירת מאגרי מידע ענקיים כגון ENCODE, Roadmap Epigenomics ו-FANTOM5.

תמצית

אנו חיים בתקופה מרגשת בה מידע עתק ביולוגי וביורפואי זמין למחקר לגילוי תובנות ביולוגיות חדשות על בקרה גנומית. בקרה זו קובעת כיצד תאי גופנו שולטים בכמות ובהרכב המדוייק של החלבונים הנוצרים מכל גן תחת תנאים מסויימים. מאמץ עיקרי במחקר זה מכוון לפיתוח שיטות חישוביות לגילוי אזורי בקרה גנומיים לא מקודדים ולהבנה כיצד הם מאורגנים מבחינה מרחבית בגנום על מנת לבקר שיעתוק גנים. יתר על כן, ההבנה כיצד תאים יוצרים מבנה גנומי מרחבי מסויים עשויה לספק תובנות חשובות על האירועים הייחודיים לתאים אלו הקובעים את גורלם. בחיבור זה חקרנו אספקטים שונים של אזורי בקרה ומיקומם המרחבי בגנום, תוך שימוש במקורות מידע נרחבים מתאים רבים ושונים, על ידי שימוש בטכניקות כגון מידול הסתברותי, למידה סטטיסטית ולמידה עמוקה (deep learning).

פיתחנו מספר שיטות חישוביות חדשות להסקת קשרי מעצם-מקדם (enhancer-promoter interactions). השיטה הראשונה, שהורחבה מתיזת התואר השני, חוזה קשרי מעצם-מקדם המראים קורלציה גבוהה בפעילות האקטיבית בין המעצם ובין המקדם על פני מספר גדול של תאים שונים. השיטה השנייה חוזה מי מהקשרים שנחזו בשיטה הראשונה הינם ספציפיים לסוגי תאים מסויימים. הראינו ששתי השיטות בעלות ביצועים טובים יותר משיטות קודמות ומספקות תובנות ביולוגיות חדשות. לבסוף, באמצעות טכניקות בלמידה עמוקה, ענינו על שאלות מגוונות לגבי כיצד נכון לחזות משתיקים (silencers) פונקציונאליים ומה מגדיר אותם מבחינה אפיגנטית.

הפקולטה למדעים מדוייקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע"ש בלבטניק

שיטות חישוביות להסקה ושימוש ברשתות מעצם-מקדם

חיבור לשם קבלת תואר "דוקטור לפילוסופיה"

מאת **תום אהרן עיט**

בהנחייתם של פרופ' רון שמיר ופרופ' רן אלקון

הוגש לסנאט של אוניברסיטת תל אביב

ספטמבר 2023