Tel-Aviv University

Raymond and Beverly Sackler Faculty of Exact Sciences

The Blavatnik School of Computer Science

Using large-scale high-throughput data for enhancer-promoter network inference

Thesis submitted in partial fulfillment of the requirements for

M.Sc. degree in the School of Computer Science, Tel-Aviv University

By

Tom Aharon Hait

The research work for this thesis has been carried out at Tel-Aviv University

under the supervision of

Prof. Ron Shamir

Dr. Ran Elkon

May 2017

Acknowledgements

Three people guided and supported me through this thesis, both my advisors and Dr. David (Didi) Amar. From each one of them I have learned how to properly conduct a biomedical research with integrity and reliability.

First, I would like to thank to my collaborator and my friend, Didi, for sharing with me his exceptional knowledge in machine learning, for his close guidance in my thesis, in Expander, and in ADEPTUS project, for his critical and practical remarks, and for giving me mental support when most needed. I would like to thank to Dr. Rani Elkon for showing me how deep and wonderful the biology world is, and for his broad creativity and ideas in recent biological and bioinformatics trends. I would like to thank to Prof. Ron Shamir for giving me the opportunity to join to an amazing lab first as a software programmer of Expander and then as a researcher. The computational and biological aspects taught by Ron had proven invaluable. It was truly a privilege to be guided by the abovementioned researchers.

I also would like to thank to my recent and former lab fellows and friends. Adi Maron-Katz for guiding me during my work on Expander software tool. Yaron Orenstein for assisting me with computational problems within Expander and AMADEUS. Dvir Netanely, Kobi Perl, Idan Nurick, Roye Rozov, David Pellow, Gal Dinstag, and Ron Zeira for the fruitful scientific and non-scientific discussions (especially on Veganism). Special gratitude goes to Gilit Zohar-Oren for her availability to any administrative or morale problems.

Last but not least, I thank my father, Avi, and my aunt, Efrat, for their love and support.

Finally, I deeply thank to the Edmond J. Safra Foundation for the financial support over the past two years.

This thesis is dedicated to the memory of my mother,

Ronit (Goldberg) Hait (1950-2008)

"We cannot solve our problems with the same thinking we used when we created them"

Albert Einstein

Abstract

Comprehensive identification of enhancer regions in the genome and mapping these regions to their target genes are key functional genomics tasks. Current predictions of enhancer regions are mainly based on epigenetic marks that are correlated with enhancer activity, including H3K4me1, H3K27ac and binding of p300. However, recent large-scale experimental tests suggested that many of these candidates are not functional.

Enhancer RNAs (eRNAs) is a recently discovered class of transcripts that are transcribed, mostly bidirectionally, from genomic regions of active enhancers. It was recently suggested that eRNA expression could serve as a superior mark of functional enhancers. Since eRNAs have relatively low stability and are mostly not polyadenylated, their expression is not detected by standard RNA-seq protocols. However, eRNAs are readily detected by the GRO-seq technique, which measures transcription rates of nascent RNAs on a genomic scale.

We developed a novel computational method called FOCS (FDR-corrected OLS with Cross-validation and Shrinkage) for mapping enhancers to their target genes. FOCS employs regressing each gene expression with the eRNA expression of proximal enhancers, along with elastic-net for model shrinkage and cross-validation across cell types to avoid information leakage. We applied FOCS on publicly available GRO-seq data from 40 studies encompassing 246 samples and covering 23 cell lines examined under control and stressed conditions. When validating the predictions on external data, including ChIA-PET and GTEx eQTLs, FOCS outperformed, both in quality and quantity, previous predictions made using the pairwise enhancer-promoter correlation and Lasso-based regression method.

Contents

Figures

Tables

Supplementary Figures

Supplementary Tables

1. Introduction and summary

Recent advances of genome-wide next generation sequencing (NGS) technologies allowing to systematically identifying active enhancers and the publicly available high number of experiments conducted can now be used to develop algorithms for linking enhancers to their target genes (or promoters) across diverse cell-types examined under control and stressed conditions. Predicted enhancer-promoter (E-P) links can be further used to develop new methods for inferring transcriptional regulators from co-expression of target genes by incorporating active enhancer regions in addition to the traditional and limited inference using promoter regions.

Large-scale genomic measurements integration across multiple and diverse samples from different cell-types aims to dissect biological phenomena that is either common to most cell-types or specific to few cell-types. For example, in genome-wide association studies (GWAS), a set of genetic variants is compared with changes in gene expression levels across different individuals to identify if any variant is associated with a trait [1], e.g., a disease or a tissue. Other methods used the identified co-expressed gene clusters from gene expression profiles to identify enriched up or down regulated functional processes (e.g., cell-cycle) [2–4], or to identify common motifs in the promoters of the genes [5] and to associate these motifs to transcription factors (TFs), which are proteins that activate or repress gene transcription.

Biological network inference is a process of making inferences and predictions between genes, proteins, and metabolites using the growing sets of high-throughput (HT) expression data. Briefly, methods that use HT expression data for inference of regulatory networks rely on searching for patterns of partial correlation or conditional probabilities that indicate casual influence [6,7]. Such patterns combined with other supplemental data on the genes or proteins, or with other information on the cell-types, form the basis upon which such algorithms work. A network is represented as a set of nodes (genes or proteins) with directed or not directed edges between them indicating an influence. Examples for such methods are: protein-protein network inference from multiple heterogeneous data [8,9] and multiple methods of gene regulatory network inference based on expression data and other biological information [10]. Network inference methods can also connect mixed nodes from different features, e.g., protein-gene interactions or enhancer-promoter interactions. These networks are usually represented as a bipartite graph $G(V, V', E)$ where V, V' are two sets of nodes representing two different genomic features and E is the set of edges connecting nodes between V and V'.

In this thesis, we aimed (1) to develop an improved statistical method for E-P network inference without assuming any underlying distribution on the data and (2) to build genome-scale E-P maps based on comprehensive meta-analysis. Previous methods for predicting enhancer-promoter links are: (1) Pearson pairwise correlation between expression patterns of enhancer and promoter [11,12] and (2) modeling gene expression with k proximal enhancers using ordinary least squares (OLS) regression followed by Lasso model selection to reduce number of enhancers [12]. These methods evaluate the performance of the gene models based on statistic scores under the assumption of normal distribution, which may not be valid for count expression data.

Here we describe FOCS (FDR-corrected OLS with Cross-validation and Shrinkage), a novel method for enhancer-promoter network inference using enhancer and gene expression data generated from global-run-on sequencing (GRO-Seq) technique. GRO-Seq measures both enhancer and gene transcription rates at the same time in a single experiment. We use different machine learning regression methods to build a gene model prediction based on proximal enhancers. We apply leave cell-type out cross validation to avoid model over-fitting to the training set. We developed two novel non-parametric statistical validation tests to evaluate each gene model performance. Given two expression matrices of enhancers and genes across multiple samples from different cell-types, our method predicts enhancer-promoter links. Our method is automatic and can be used on other types of NGS technologies.

We used external sources to validate the quality of the predicted E-P links. The first source is DNA-DNA 3D interactions from chromatin interaction analysis by paired-end tag (ChIA-PET) data and the second source is genetic variants called single nucleotide polymorphisms (SNPs) that are associated with changes in gene expression levels. We show that FOCS outperformed, both in quality and quantity, the previous mentioned methods for enhancer-promoter network inferences.

2. Background

This chapter lays out the background and terminology required for the thesis. In **Section 2.1** we introduce basic biological definitions and recent findings. We also discuss on high throughput data types that were used in this thesis, and give a brief introduction to cancer diseases. In **Section 2.2** we discuss and give formal definitions of the computational problems addressed. This section includes problems of data representation, identification of expressed genomic regions, gene clustering, motif finding, and gene sets enrichment analysis. We also give background on regression analysis and describe in detail the regression methods used in the thesis.

2.1. Biological background

This section introduces the relevant biological terms and definitions needed for understanding the goals and the computational problems addressed in this thesis. For more details on basic biology see [13], and for gene regulation, enhancers and epigenetics see [14]. We also discuss existing technologies allowing us to systematically identify genomic regions of interest and to measure their expression for our computational analyses.

2.1.1. Biological concepts

2.1.1.1. Introduction to cell biology

The living organisms are composed from basic fundamental units of life, called cells. Cell biology is the discipline that tries to answer questions on the structure, function, and behavior of cells. The cellular and organismal function and development is governed primarily by their deoxyribonucleic-acid (DNA). The DNA sequence contains functional units called genes. Genes are divided into two groups, coding and non-coding genes. Coding genes contain the code for protein translation while the no-coding genes contain the code for transcription of non-coding ribonucleic acids (ncRNAs), which are not translated into proteins.

ncRNAs include some important RNAs such as transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) that take part in the ribosome, which is a cellular particle translating mature coding RNAs (also known as messenger RNAs, mRNAs) into proteins. Current studies estimate that there are approximately 21,000 protein-coding genes and 9000 ncRNA genes in the human genome [13].

Apart from genes, there are other non-coding regions proximal to genes called *promoters* that contain sequences that can be bound by proteins and control gene transcription. An understudied type of regulatory regions is *enhancers*, which are distal regions from the gene that are also bound by activator proteins to promote gene transcription. Non-coding regions, including enhancers, are one of the sources of high sequence variation in DNA sequences of humans. *Single nucleotide polymorphism* (SNP) is a variation in a single nucleotide position

across the population. SNPs are more frequent in non-coding regions compared to coding regions [15]. These variations can affect the individual's disease risk, response to pathogens, chemicals, and other agents.

Understanding how non-coding regulatory regions interact with genes and affect on gene transcription can help associate variations in non-coding regions with gene's transcription levels and improve our understanding of genetic factors affecting disease predisposition.

2.1.1.2. Gene regulation

Cells use various mechanisms to increase or decrease the amounts of gene products, RNAs and/or proteins, in a process termed *gene regulation*. Gene regulation is essential for the organism to be adapted and versatile to its environment by allowing the cell to express specific RNAs and proteins when needed. The first discovery of the gene regulation was in 1961 when the *lac operon* was identified by Jacques Monod, showing that some enzymes involved in lactose metabolism are expressed in Escherichia coli bacteria only in the presence of lactose and absence of glucose. Gene regulation in multicellular organisms is the driver of cellular development and differentiation in the embryo, which leads to different cell types possessing different gene expression profiles from the same genome sequence. Differences in gene expression profiles can, in turn, lead to differences in RNA/protein abundance and as a consequence to differences in the phenotypic characteristics of the cells.

The central dogma of molecular biology is the model of the genetic information (see **Fig. 1a**). The transfer is from DNA to RNA (via transcription), and from RNA to protein (via translation). This genetic information flow is primarily unidirectional, although information flow from RNA to DNA also occurs.

Regulation of gene transcription is govern by sequence specific proteins called transcription factors (TFs) that bind to regulatory regions in the genome. A central regulatory region is the promoter, which initiates gene's transcription and is located near the transcription start site (TSS) of the gene (see **Fig. 1b** for gene structure). The initial product of a gene transcription is a pre-mature RNA composed of the 5UTR, introns, exons, and 3UTR sequences. The pre-mature RNA is later transformed to mature RNA by assembling one or more exons together in a process called *splicing* (see **Fig. 1a**), filtering out introns and UTRs. The splicing process is also governed by regulatory regions (located within introns and exons) and by sequence specific proteins [16].

Enhancers are another group of regulatory regions that regulate transcription by pairing with specific promoters via co-factor proteins mediating the enhancer-promoter (E-P) link (see **Fig. 2a**). The promoter sequence in the vicinity of the TSS is sufficient to assemble the POL2 machinery. However, transcription is often weak in the absence of E-P links stabilizing the POL2 machinery [14]. Understanding the function of enhancers is currently an area of great interest with a potential impact on understanding gene expression development, and on evolution and disease studies [17–19].

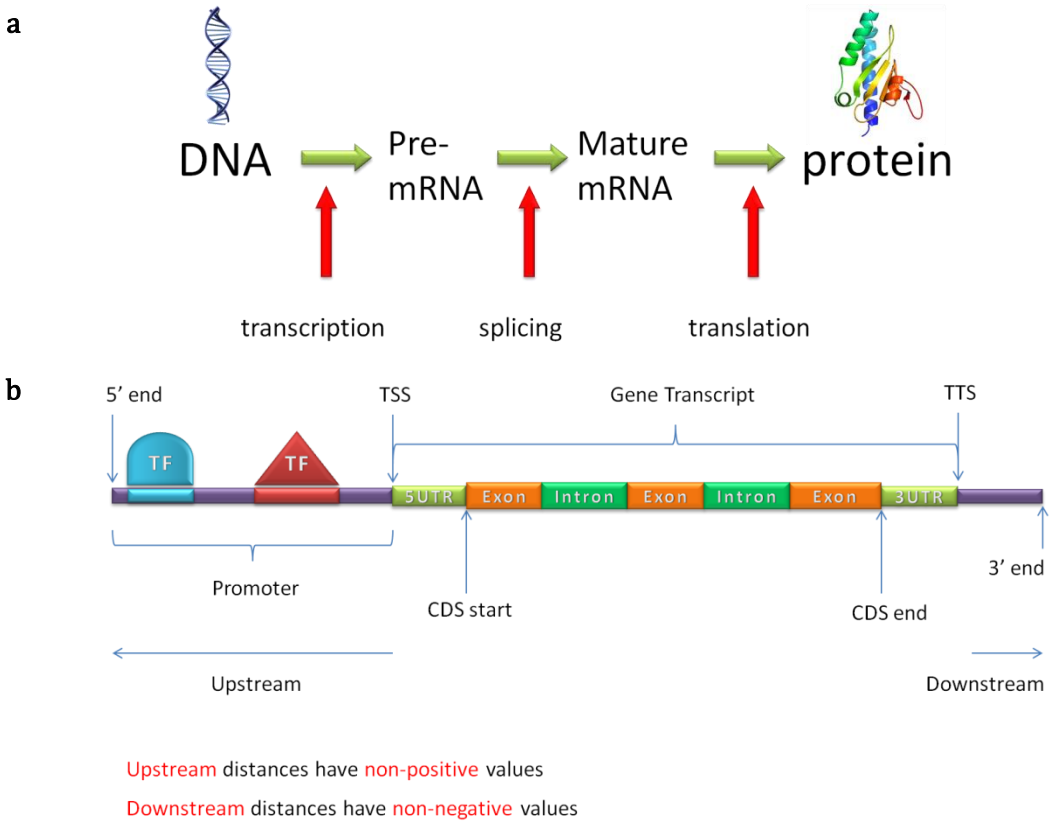


Figure 1. Gene structure and regulation. (a) The central dogma of molecular biology. The genetic information flow starts from the DNA and ends with the protein product. The gene regulation process controls the transcription step. (b) The gene structure. The promoter is bound by sequence specific TFs 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.

2.1.1.3. Chromatin organization

Chromatin is a complex of macromolecules found within the nuclei of the cells, consisting of DNA, protein, and RNA. The chromatin's primary functions are 1) DNA packaging into a more compact and denser shaping, 2) preparing the DNA macromolecule for cell division, 3) preventing DNA damage, and 4) to control DNA replication and gene expression. These functions are mediated by the chromatin organization, which is controlled by multiple factors.

One of the key players in chromatin organization are particles called *nucleosomes* observed by Don and Ada Olins in 1974 [20], which are involved in DNA organization and packaging. The nucleosome core particle consists of eight protein cores (i.e., as octamers) called *histones*. The DNA is wrapped around the core twice, forming a unit of length approximately 165 bp [21]. Series of higher order structures eventually form a *chromosome*. This added compaction of the DNA creates an additional layer of regulation of gene expression [22,23]. The DNA folding and

un-folding around the nucleosomes is controlled through chemical modifications on nucleosomes (see **Fig. 2b**).

The nucleosome core consists of eight histones: H2A, H2B, H3 and H4, each in two copies. The amount and extent of the DNA wrapping around the nucleosomes is controlled by chemical modifications on specific sites in the histone proteins. For example, nucleosomes flanking active enhancer regions (see **Fig. 2b.b**) are often acetylated or methylated at lysine (Lys or 'K' in amino acid table) sites in histone H3. These sites are acetylated by H3K27ac, i.e., acetylation of histone H3 at lysine site 27, and methylated by H3K4me1, i.e., methylation of histone H3 at lysine site 4. Active promoters bound by POL2 are flanked by nucleosomes (see **Fig. 2b.c**) with H3K27ac and H3K4me3 modifications. These histone modifications can be measured using chromatin immunoprecipitation sequencing (ChIP-Seq) high-throughput technique.

2.1.1.4. Enhancers

Enhancers are short, 50-1500 base-pairs (bp), regions of DNA that when bound by TFs increase gene's transcription [24,25]. The first enhancer discovery, 30 years ago, was a 72 bp sequence of the SV40 virus genome, which could enhance the transcription of a reporter gene in HeLa cells several hundred fold [26]. Soon after that, enhancers were discovered in animal genomes [27]. Since then, an extensive research was done to describe the biochemical and functional properties of many enhancers [14]. Enhancer sequences contain short DNA motifs that are binding sites (BSs) for TFs. 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**). These markers control the DNA packaging, accessibility for transcription, preventing DNA damage, and replication in cell division (see **Section 2.1.1.3** for further details).

Enhancers were traditionally identified using enhancer trap techniques using reporter gene assays or by comparative sequence analysis between multiple species in computational genomics. For example, in flies, lacZ gene was used as a reporter and fused into the fly genome. If the reporter gene fused near an enhancer then the lacZ expression will reflect the expression pattern driven by that enhancer [28].

The emergence of more advanced genomic and epigenetic technologies allowed large-scale identification of enhancers. *Next generation sequencing* (NGS) methods enable the large-scale identification of TF binding sites, detection of extensive epigenetic data 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 enabled identification of nucleosome-depleted, or open chromatin regions that can contain regulatory regions. Computational methods for NGS data analysis include comparative genomics via sequence conservation of non-coding regions [29,30], clustering of known or predicted TF-binding sites [31], and supervised machine-learning approaches trained on known regulatory regions [32].

All of these methods have proven effective for regulatory region discovery, but each has its own limitations, and each creates greater or lesser number of false-positive identifications [33].

In this thesis we focus on enhancer identification and linking to target genes by using high-throughput (HT) methods measuring both enhancer and gene expressions across many cell-types. Examples for such projects include the FANTOM5 consortium [12], which used cap analysis of gene expression (CAGE) deep-sequencing HT method, and projects that utilize DNase-Seq expression data [11,34] from the ENCODE consortium [35]. These projects assume that the expression patterns of an enhancer and its target gene are highly correlated. To this end, pair-wise correlation is used for linking enhancers to their target genes. However, this method does not take into consideration the possibility that multiple enhancers contribute to enhancing the same gene expression [36]. Other projects seek enhancer-promoter (E-P) links from contact interactions in the 3D genome architecture, which can be captured by Hi-C and ChIA-PET HT techniques [37–39]. However, currently a limited number profiles is available from Hi-C and ChIA-PET, and therefore, the confidence in the predicted E-P links from these experiments may be low.

We aim to use currently available global run-on sequencing (GRO-Seq) data to develop a better approach for linking enhancers to their target genes that takes into consideration also the architecture of enhancers.

Elucidating active enhancers and delineating E-P links is a challenging task that is now more achievable since the emergence of HT technologies capturing simultaneously enhancer and gene expression and interactions in the 3D genome structure. Increasing the current knowledge of annotated enhancers and E-P links can expand our understanding on gene regulation and may suggest different disease treatment approaches targeting enhancers in addition to the traditional gene-based treatments.

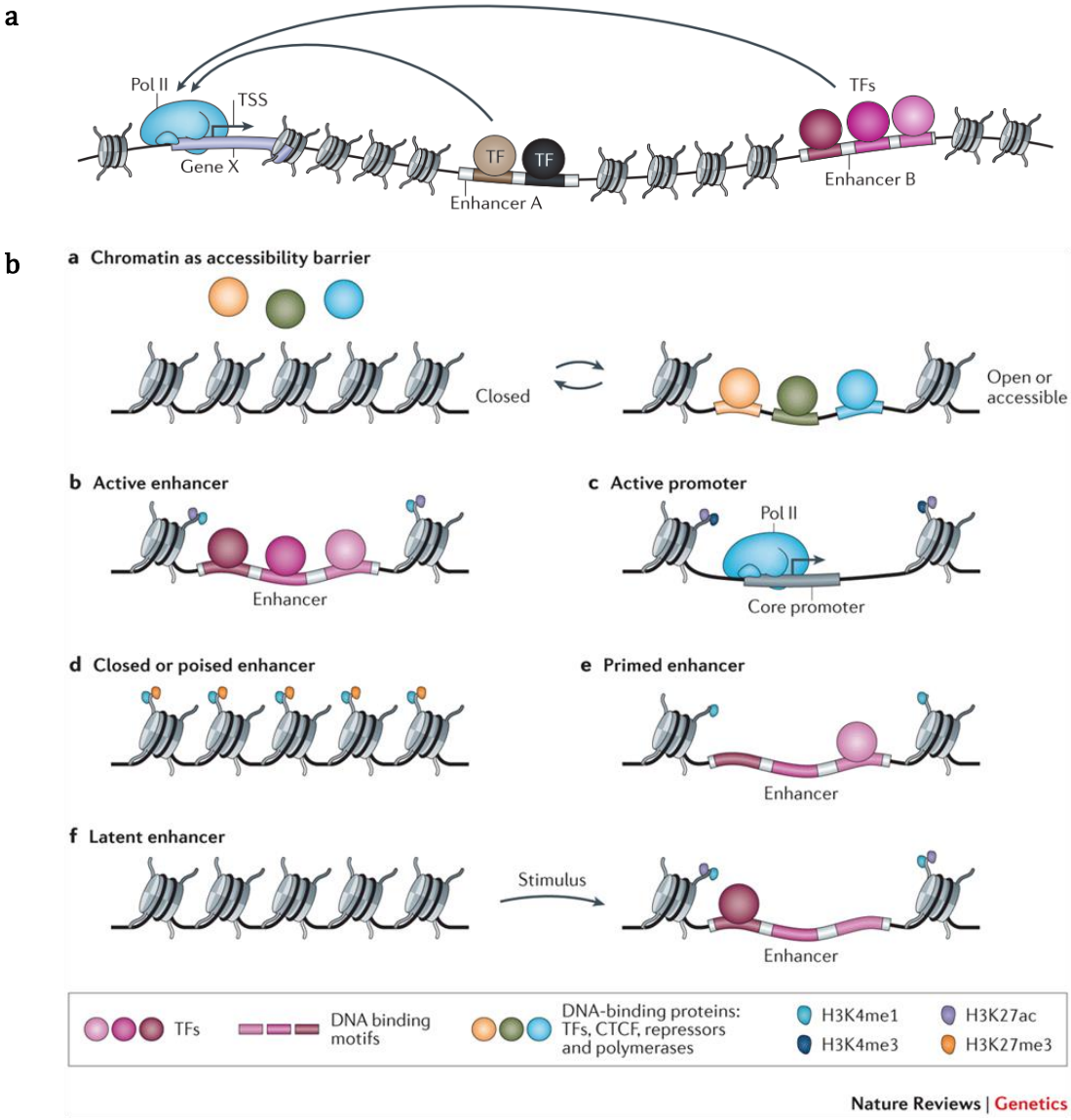


Figure 2. Enhancers and Chromatin accessibility controlled by histone marks. (a) Enhancers located distal from gene X are linked with POL2 via co-factor TFs. Enhancers contain binding sites for sequence specific TFs. Nucleosomes are located in regions between enhancers and gene X. (b) Chromatin accessibility 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: [14].

2.1.1.5. Epigenetics

The term epigenetics in its contemporary usage emerged in the 1990s, but for some years has been used in ambiguous meanings [40]. A consensus definition of the concept of epigenetic trait as "stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence" was formulated at a Cold Spring Harbor meeting in 2008 [41], although alternate definitions that include non-heritable traits are still being used as suggested by the epigenetic roadmap consortium [42]. The molecular basis of epigenetics is changes that modify the activation of certain genes, but not the genetic sequence of the DNA.

Examples for such epigenetic changes are DNA methylation and histone modification, each of which alters how genes are expressed without altering the DNA sequence. Epigenetic changes can be a result of DNA damage [43–45].

Chromatin remodeling is one way of regulating gene expression via epigenetic changes (see **Section 2.1.1.3** for further details). Epigenetic changes modify the way DNA is wrapped around the histones and thus can cause a change in the gene expression as well.

Open chromatin regions in the DNA bound by TFs, which can be both expression and location measured using DNase-Seq technique, could also be indicative of epigenetic changes. Examples for open chromatin regions are active enhancers and promoters. Active enhancers and promoters were previously identified and linked using one or more combinations of DNase-Seq, histone modifications ChIP-Seq, and other non-epigenetic high-throughput techniques (e.g., RNA-Seq) covering multiple cell-types [11,34,46].

Active enhancer and promoter identification and mapping still remain challenging tasks. The use of epigenetic data is common to all projects for both validation and identification of enhancer-promoter mapping.

2.1.1.6. Cancer cell lines

Comparative cancer research across multiple types of cancer diseases requires sources of cells that can be grown in large numbers of uniform cell-type, stored in liquid nitrogen at -193ºC for indefinite period and retain their viability when thawed [13]. Cancer cell lines are generated from transformed cancer tissues that underwent an immortalization process. These cell lines differ from normal cells/tissues in several ways. Transformed cell lines often grow without attachment to a surface and can proliferate to a much higher density in a culture dish.

Immortalized cell lines have accumulated sufficient number of mutations in their genome or were provided with viral genes deregulating the cell-cycle process allowing them to proliferate indefinitely. For example, human fibroblasts provided with the gene that encodes the catalytic subunit of telomerase, which prevents telomere shortening in chromosomes, can proliferate indefinitely.

Primary cells are also used in experiments. They differ from cell lines in that primary cells were directly isolated from the human or animal tissue and therefore still encompass the

tissue characteristics, they have a limited lifespan, and may have higher purchasing costs than cell lines.

One of the main sources of HT genetic and epigenetic data on many cell-types is the ENCODE project [35]. Currently, 182 immortalized and transformed cell lines and 142 primary cells covering many high-throughput techniques are available in the ENCODE project.

2.1.2. Next Generation Sequencing (NGS)

Next generation sequencing (NGS) is a general name to novel sequencing techniques developed over the last decade. NGS performs very HT deep sequencing within a single day that can provide hundreds of millions of short sequences. NGS has revolutionized the genomic research in term of time and cost needed to generate the sequence data compared to the previous Sanger sequencing technology [47], which was used in the human genome project.

This section introduces the relevant high-throughput techniques used in this thesis. We also describe in general how NGS data is preprocessed and used for gene and enhancer identification and quantification.

2.1.2.1. Chromatin Immunoprecipitation (ChIP-Seq)

ChIP-Seq is a method for analyzing protein interactions with the DNA [48]. ChIP-Seq combines chromatin immunoprecipitation (ChIP) with DNA deep sequencing to identify binding sites of DNA-associated proteins. The ChIP-Seq workflow is described in **Figure 3**.

ChIP-Seq can also be used to identify histone modifications sites along the genome by targeting histone marks such as H3K4me1 and H3K27ac for active enhancers, and H3K4me3 and H3K27ac for active promoters (see **Fig. 2b**). In addition, ChIP-Seq of P300/POL2 can help identify BSs in active enhancers/promoters respectively. ChIP-Seq is widely used for comparison between different cell types by identifying BSs preferences of one or more TFs. For example, a study mapped nine chromatin marks across nine cell types to identify regulatory elements and to link enhancers to their target promoters [49].

DNA-protein BSs prediction from ChIP-Seq read count data (also known as 'peak calling') requires developing computational tools that perform peak calling. The most popular method is MACS [50], which empirically models the shift size between two ChIP-Seq peaks (one on each DNA strand) flanking the same BS, and uses it to improve the spatial resolution of the predicted BSs (see **Fig. 4**).

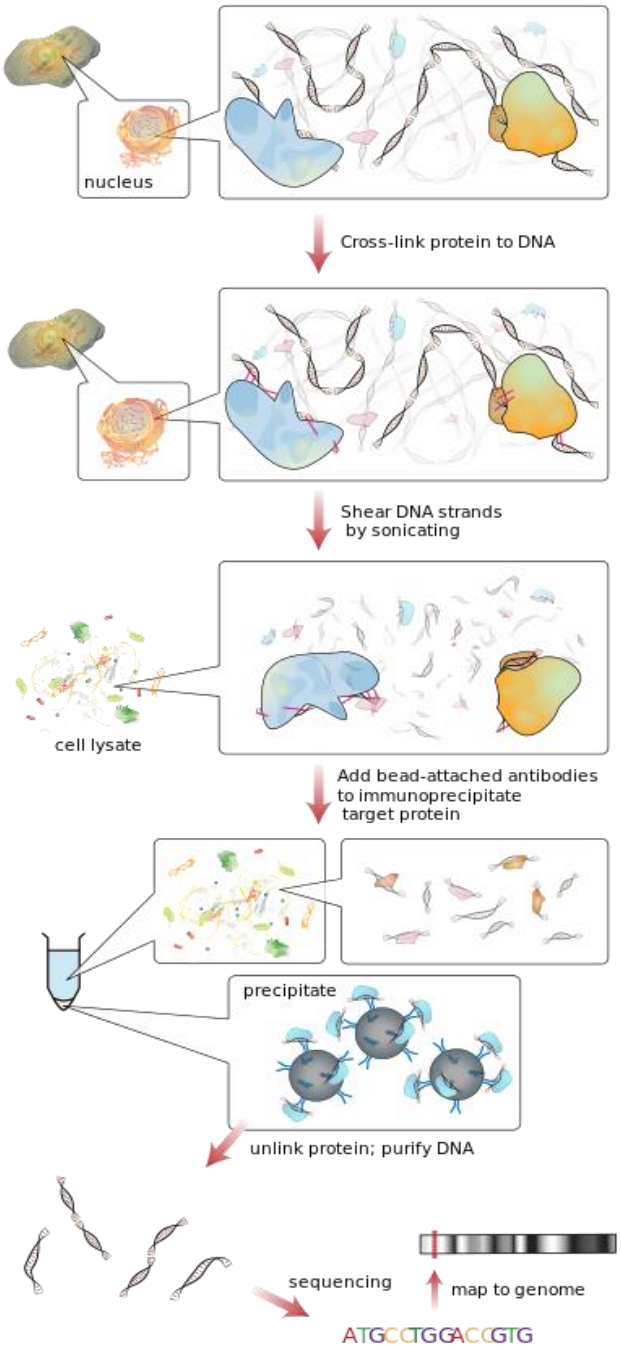


Figure 3. ChIP-Seq workflow. First, the DNA is extracted from the nucleus and cross linked to the proteins to prevent detaching during sonication process. Second, the DNA is fragmented by sonication. Third, a protein-specific antibody is added to attach to the protein of interest. Fourth, the antibody is precipitated to select only DNA fragments attached to the protein of interest. Finally, the proteins are removed from the DNA segments, the segments are sequenced and mapped to a reference genome. Source: https://en.wikipedia.org/wiki/ChIP-sequencing

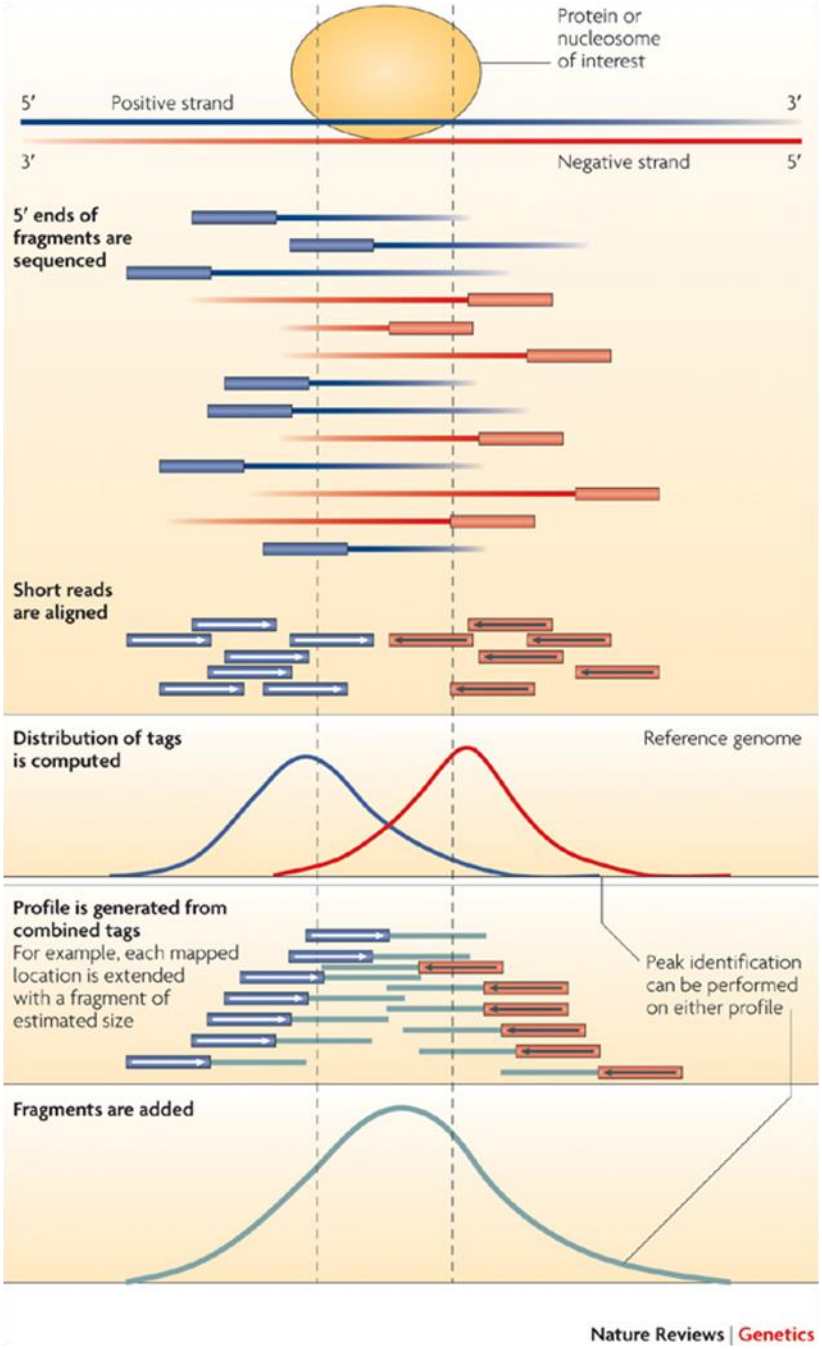


Figure 4. ChIP-Seq peak calling. DNA fragments from a ChIP experiment are sequenced from the 5' end. The alignment of these tags to the genome results in two peaks (one on each strand) that flank the binding location of the protein or nucleosome of interest. This strand-specific pattern can be used for the optimal detection of enriched regions. To create an approximate distribution of all fragments, each tag location can be extended by an estimated fragment size in the appropriate orientation and the number of fragments can be counted at each position. Source: [51].

2.1.2.2. DNase I hypersensitive sites detection (DNase-Seq)

DNase-Seq is a method for identifying locations of open chromatin regions. Such regions are known to be sensitive to DNase I cleavage [52,53] (see **Fig. 5b**). The technique is briefly described in **Figure 5**. DNase-Seq locations (also known as DHS peaks) are widely used for enhancer and promoter identification as these regions are known to be open and bound by proteins when active. Many studies developed methods for linking enhancer to promoter based on DHS peaks and expression [11,34].

Identification of open chromatin regions from DNase-Seq read count data is done computationally. Computational tools can be divided into two classes: *segmentation-based* and *site-centric* methods. Segmentation-based methods use *hidden markov models* (HMM) or sliding window methods to segment the genome into open/closed chromatin regions. Examples for such methods are HINT [54], Boyle [55] and Neph [56]. Site-centric methods identify footprints given the open chromatin profile around motif-predicted BSs, i.e., regulatory regions predicted using DNA-protein sequence information. Examples for such methods are CENTIPEDE [57] and Cuellar-Partida [58].

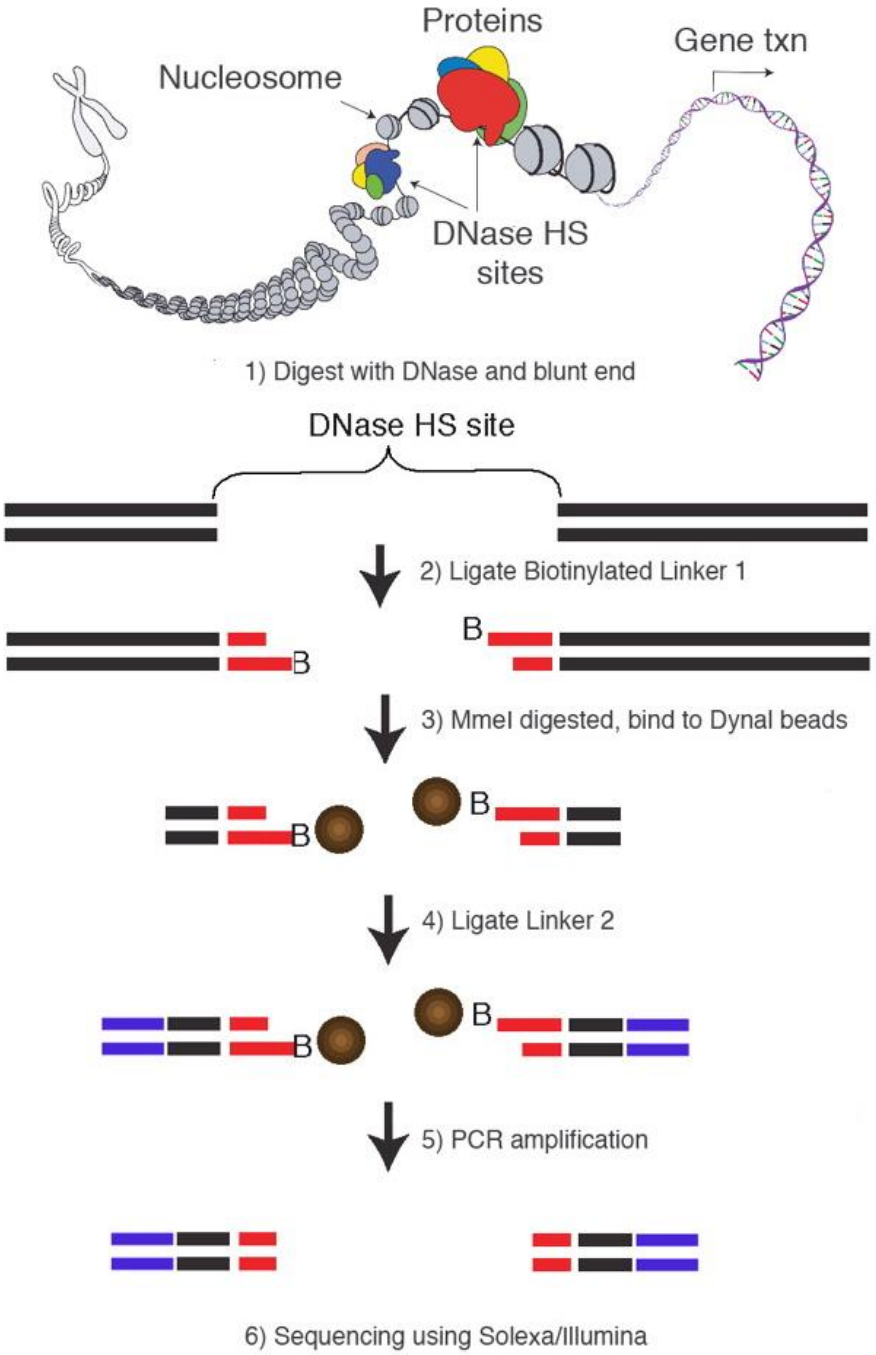


Figure 5. Schematic of DNase-Seq workflow. Cells are lysed with detergent to release nuclei, and the nuclei are digested with optimal concentrations of DNase I. DNase I-digested DNA is embedded in low-melt gel agarose plugs to reduce additional random shearing. DNA (while still in the plugs) is then blunt-ended, extracted, and ligated to biotinylated linker 1 (red bars). Excess linker is removed by gel purification. Biotinylated fragments (linker 1 plus 20 bases of genomic DNA) are digested with MmeI 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 and sequenced. Source: [59].

2.1.2.3. Chromatin Interaction Analysis by Paired-End Tag (ChIA-PET)

ChIA-PET is a technique that determines genome-wide long-range chromatin interactions. The technique incorporates ChIP-based enrichment (usually of POL2 interactions), chromatin proximity ligation, Paired-End Tags, and high-throughput sequencing.

ChIA-PET is useful in identifying E-P interactions in the 3D genome structure mediated by a protein/TF of interest. POL2 is often used in the ChIP step since it can point to expression of all active enhancers and genes. ChIA-PET can also be used to unravel mechanisms of genome control during processes such as cell proliferation, cell differentiation, and development. Using ChIA-PET one can create the interactome maps for DNA-binding regulatory proteins and promoter regions and identify unique targets for therapeutic intervention [60].

In **Figure 6** ChIA-PET is shown alongside the previous methods, ChIP-Seq and DNase-Seq. All of these methods are used to identify heterogeneity in regulatory elements that were affected due to epigenetic changes.

The key step of the ChIA-PET protocol is to link two DNA regions (also termed as paired end tags) that are close to each other by a mediator protein (e.g., POL2). The linking is done by adding a linker sequence that attaches these tags, forming a sequence of tag-linker-tag. Further details on ChIA-PET protocol and analysis are described in **Figure 7** and in [61].

ChIA-PET interactions data reveal an additional level of gene regulation that depends on the 3-dimensional genome structure. Such information could be used to identify E-P links, and to validate and compare methods for predicting E-P links based on one dimensional high-throughput techniques.

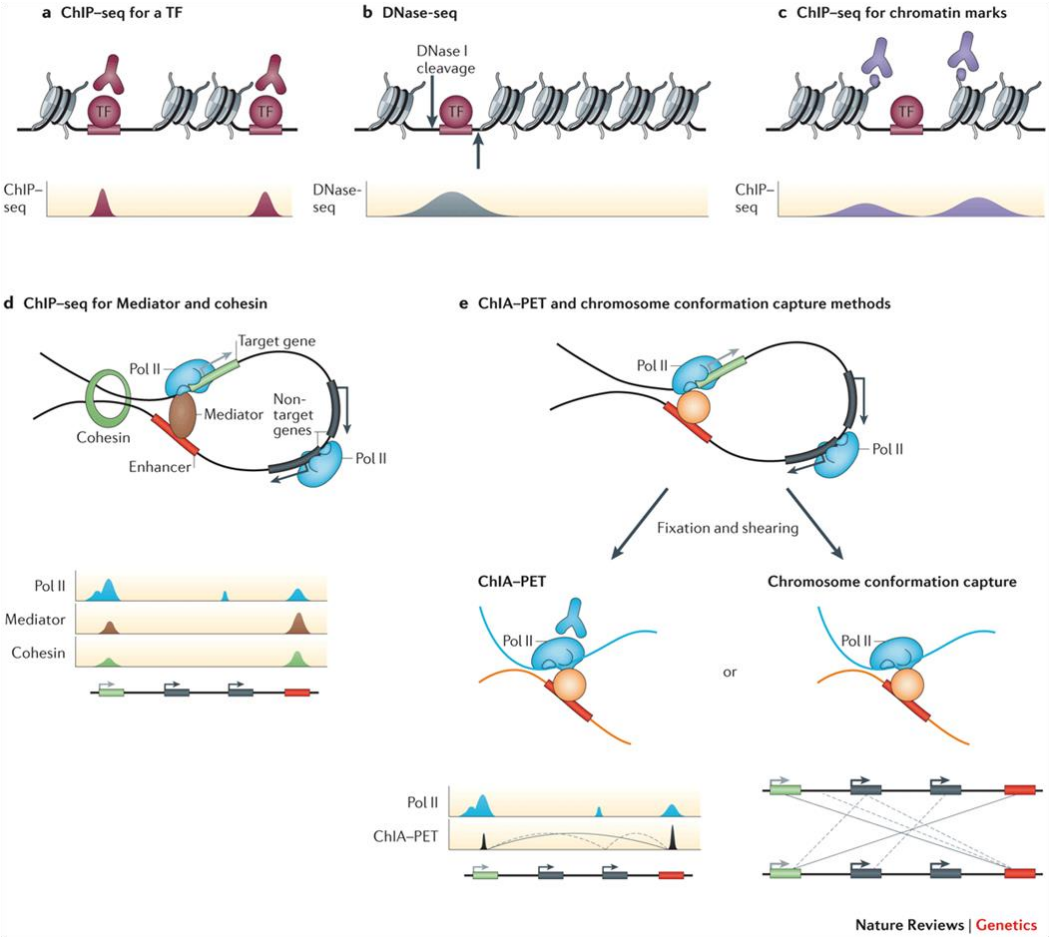


Figure 6. Regulatory elements identification techniques. (a) ChIP-Seq identifies contact points by targeting a specific TF. (b) DNase-Seq identifies regulatory elements using DNase I that cleavages DNA positions that are not bound by TFs/nucleosomes. (c) ChIP-Seq of chromatin marks such as H3K4me1 and H3K27ac identify active enhancers regulatory elements. (d) ChIP-Seq of cohesin and mediator proteins that link between the POL2 and the enhancer region. Note it only identifies the contact points of the cohesin and the mediator. (e) Two methods that identify the pairwise interactions between contact points. ChIA-PET differs from the chromatin conformation capture (3C) technique by applying the ChIP step targeting only interactions mediated by POL2 using a specific POL2-antibody. Source: [14]

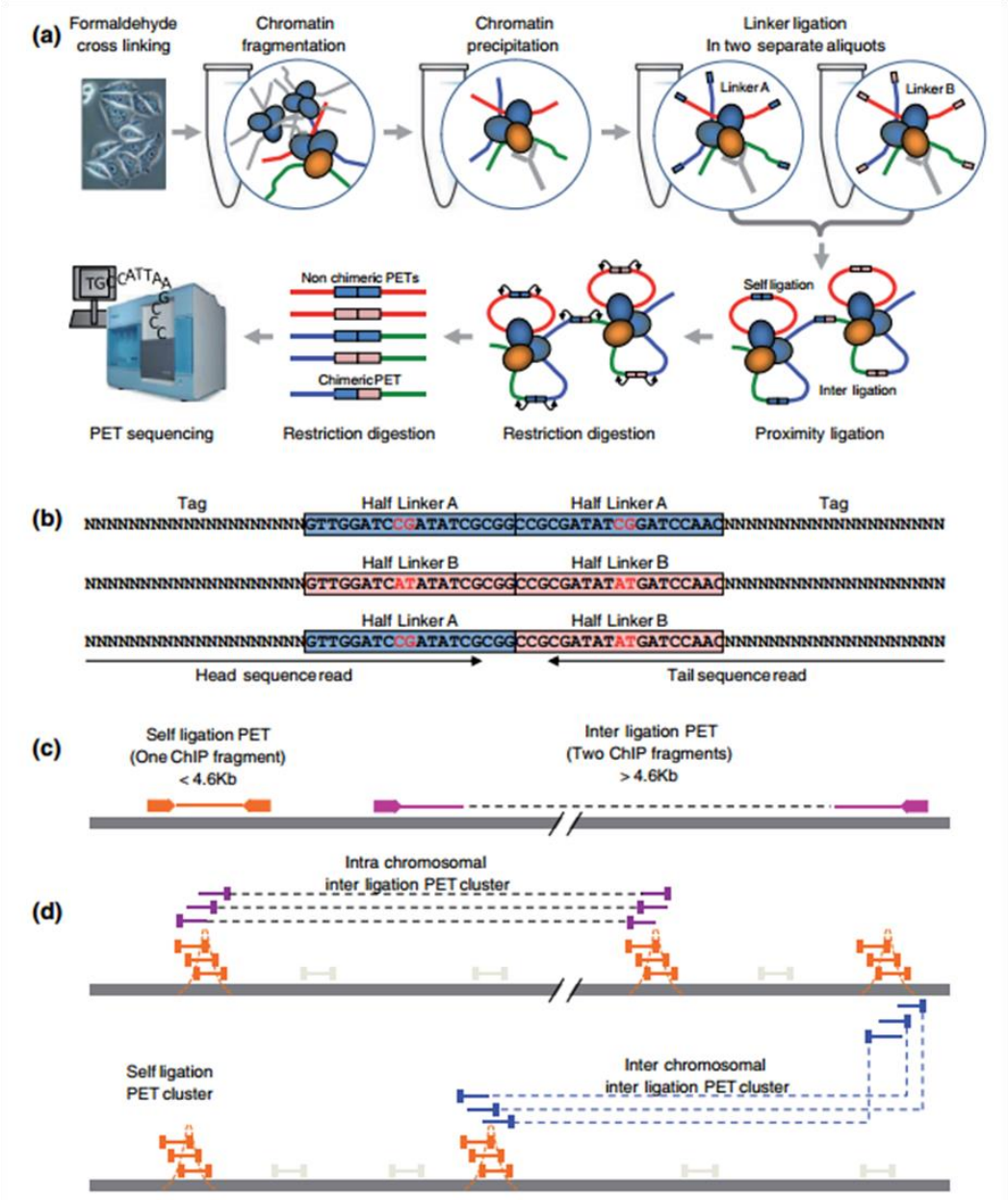


Figure 7. Schematic of ChIA-PET analysis. (a) The ChIA-PET experimental protocol, which includes chromatin preparation, ChIP, linker ligation, proximity ligation, MmeI restriction digestion, and DNA sequencing. (b) The ChIA-PET constructs prepared for sequencing analysis. Each paired-end tags (PET) construct involves a pair of tags (20 bp each) and a linker (38 bp) between the tag pairs. This full-length linker is derived from ligation of two half-linkers, A or B, each with a unique barcode nucleotide (CG for half-linker A and AT for half-linker B). The barcode nucleotides are highlighted as red letters. Linkers with AB barcodes are considered to be non-specific chimeric proximity ligation products. This reduces noise caused from non-specific chimeric products that would not have been detectable using a single half-linker. (c) Mapping tags of PET sequences to reference genome. The categories of 'self-ligation PETs' and 'inter-ligation PETs' were assigned. (d) Clustering of overlapping PET sequences in the same genomic regions to identify enriched protein binding sites by overlapping 'self-ligation PETs' and long-range chromatin interactions by overlapping 'inter-ligation PETs'. Source: [61].

2.1.2.4. Global Run-On (GRO-Seq)

The most prevalent HT techniques, microarrays and RNA-Seq, were developed to measure gene expression levels. These techniques are capable of measuring stable transcripts (or mRNAs for coding-genes), which last for relatively long time due to chemical modifications and additions to the final assembled transcripts. However, transcripts from non-coding regions are likely to degrade fast, thus, have very low steady state levels and cannot be robustly captured by those techniques.

Global run-on sequencing (GRO-Seq) was developed to identify genomic regions that are transcribed at a certain time point directly from the DNA [62]. This means that all actively transcribed regions are identified. **Figure 8** describes in detail the GRO-Seq workflow.

Since GRO-Seq allows measuring the active transcription rather than steady state mRNA levels, it can be used to measure transcription rates of non-coding genomic regions not necessarily producing stable RNAs and to identify early changes affecting primary target genes rather than both primary and secondary targets. Novel transcripts, including non-coding RNAs, can be detected using GRO-Seq. Specific genomic regions of interest are the enhancers, which can be detected and identified based on the RNA transcribed from them [63,64]. These RNAs are termed as *enhancer RNAs* (eRNAs). eRNAs have been associated with stimulus-dependent enhancers [65] and like active promoters they exhibit transcription initiation in opposing directions on each strand, a phenomenon called *divergent transcription* [62,66,67].

Novel non-coding genes can be identified by performing de-novo transcript identification with GRO-Seq aligned reads. As described for DNase-Seq, segmentation-based tools are used for de-novo transcript identification. Examples for such tools are groHMM [68], which is based on HMM, and HOMER [69], which uses a sliding window method to segment the genome into transcribed and non-transcribed regions.

GRO-cap technique [70] is a modified form of GRO-Seq that identifies TSSs in promoters and enhancers by utilizing the tagging and extensive purification of nascent RNAs from GRO-Seq and then by employing redundant enzymatic steps to enrich for nascent RNAs with 5' caps marking TSSs. Precision run on sequencing (PRO-Seq) and PRO-cap are more advanced techniques than GRO-Seq and GRO-cap allowing identification of POL2 pausing sites downstream of transcription initiation and TSSs, respectively, at base pair resolution [71,72].

The discovery of eRNAs using GRO-Seq data opened new opportunities for enhancer-promoter (E-P) linking. Expression patterns of enhancer regions and genes have been used in correlation-based methods to infer E-P links as previously done using DNase-Seq and ChIP-Seq [11,34].

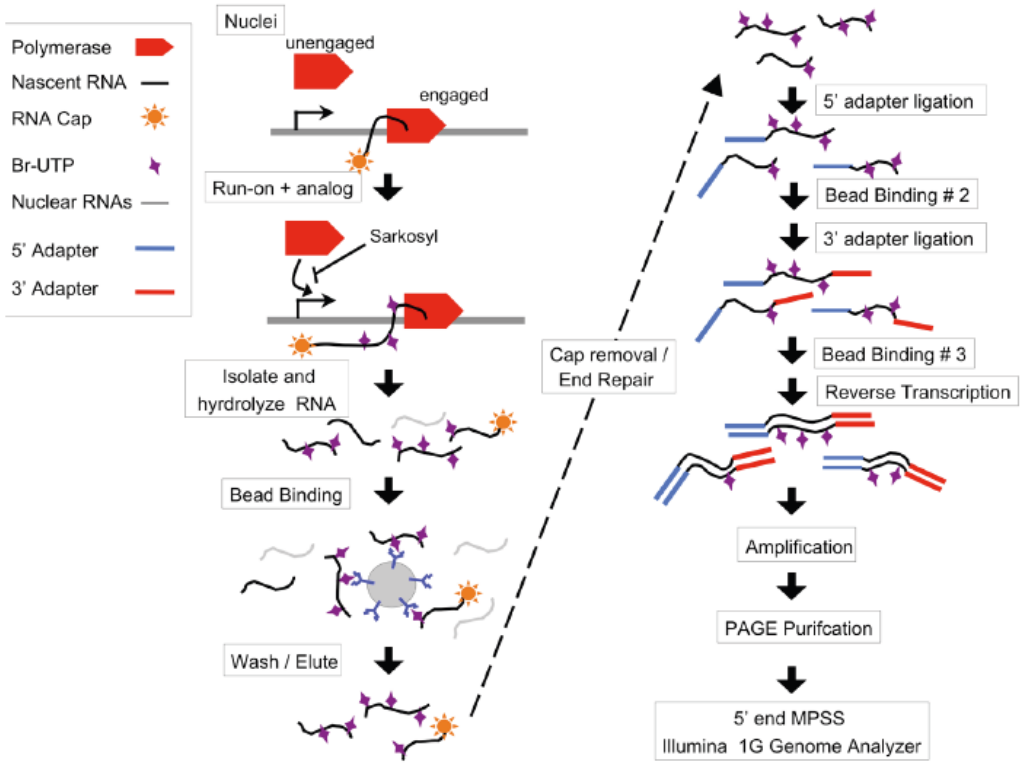


Figure 8. Schematic of GRO-Seq workflow. GRO-Seq involves the labeling of newly synthesized transcripts with bromouridine (Br-UTP). Nuclei are incubated with BrUTP in the presence of Sarkosyl, which prevents attachment of new RNA polymerases to the DNA. Therefore, only polymerases that are already attached to DNA before the addition of Sarkosyl will produce new transcripts labeled with BrUTP. The BrUTP labeled transcripts are captured using anti-BrUTP antibody beads. Adapters are added to the captured transcripts to convert them to cDNAs. cDNAs are then amplified, sequenced and aligned to the reference genome. Source: [62].

2.1.2.5. Measuring gene and enhancer expression

NGS techniques produce sequence reads that are later aligned to a reference genome. The aligned reads' positions along the genome are then used for quantifying expression of genomic regions of interest. Here we shall describe how to quantify aligned reads in genomic regions. The number of reads in each genomic region reflects its expression in a single experiment. These expression levels are used in downstream analyses to infer biologically meaningful conclusions, e.g., the functional role of the genomic region within the cell.

There are many tools that perform read quantification in genomic regions. All of them are given as input the aligned reads and genomic regions annotation data. The genomic region annotation data include the chromosome and the genomic start and end positions of each genomic region. Expression is manifested in many types of genomic regions such as genes and enhancers. The tools quantify in these genomic regions according to user supplied parameters,

e.g., minimum overlap between the read and the genomic region and whether to discard reads aligned to multiple genomic regions.

Bedtools [73] is a commonly used set of utilities for comparing genomic regions. Bedtools utilities can be run only under Linux operating system. One of its utilities is 'intersect', which performs intersection between aligned reads and prescribed genomic regions. It outputs the number of reads overlapped with each genomic region under user supplied parameters.

Genomic annotation data can be obtained either from a known database or from de-novo identified regions from aligned data. Human gene annotation data are usually taken from hg19 refgene table (RefSeq genes [74]) in the UCSC genome browser [75]. Enhancer annotation data are obtained de-novo from GRO-Seq aligned reads (for more details see **Section 2.2.2**).

Gene quantification with GRO-Seq reads is a challenging task. GRO-Seq measures transcription rates directly from the genes' regions, resulting with pre-mature RNAs captured before the splicing step. Therefore, the pre-mature RNAs may contain both exonic and intronic fragments. The quantification should be done on the pre-mature RNAs (single pre-mature RNA per gene) as annotations. Gene products may originate from different gene's isoforms, each is a transcript with different combination of exons. The standard gene annotation data contain multiple isoforms per gene. Hence, it is not clear which isoform is the origin of the RNA. A compromise for this issue is to take for each gene its consensus transcript (also known as canonical transcript), which is a single transcript per gene. There is no agreement of what is the canonical transcript and different databases use different definition. For example, UCSC defines the canonical transcript as the longest isoform. The groHMM [68] has a built-in utility that creates canonical transcripts given the UCSC hg19 refgene table. The utility, named makeConsensusAnnotations, creates canonical transcripts by reducing redundant overlapping regions from multiple isoforms of the same gene. In addition to the annotation data issue, some of the resulting RNAs are short fragments, perhaps due to the fact that POL2 loads more rapidly at the TSS in response to treatment than it escapes into the gene's body [63]. Therefore, previous works on GRO-Seq [63,76] took the canonical transcripts and ignored read segments falling a few hundred bases downstream from the TSS in order to avoid over-counting of short RNAs resulting from POL2 extensive load at the TSSs.

Enhancer quantification, on the other hand, is much easier. Enhancers are known to have a bi-directional transcription relative to their center, although there are some enhancers that have unidirectional transcription [14]. Identification of putative enhancer regions can be done using the dREG [32] tool, which identifies bi-directional transcribed enhancers. Bi-directional transcribed enhancers are quantified from both strands. However, enhancers that fall within intronic regions should be quantified only from the antisense strand of the gene in order to prevent double quantification of reads that fall in both gene and enhancer regions.

The expression data generated for a single experiment is called a *profile*. Gene/enhancer expression matrices are a combination of multiple profiles. Comparison between multiple profiles requires normalization. The normalization uses the genomic regions widths and the profile library size, i.e., the total number of aligned reads (for more details see **Section 2.2.1.1**).

The normalized expression matrices can be used for downstream analyses and for comparing between samples and experiments.

2.2. Computational background

In this chapter we lay out the computational background of this thesis. Each section deals with a different type of computational problems. More details on the computational problems addressed are given in the references in each section.

2.2.1. Data representation

In this section we describe the data structure used in this thesis.

2.2.1.1. Expression data

The expression data of genomic features (e.g., genes or enhancers) can be 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. Each row in the matrix contains the expression pattern of a specific genomic feature across all samples, and each column represents the expression profile of a sample. Thus, columns can represent different experiments, conditions, cells, or individuals.

The entries $D_{i,j}$ can represent counts or normalized values. In our analyses we consider both types depending on the computational method used. Normalized expression values are computed by dividing each entry $D_{i,j}$ count number by the library size of sample j and by the region width of genomic feature i. Normalized expression values are often defined as Read per Kilobase exon per Million mapped reads. When measuring expression values from GRO-Seq data, which produces unspliced transcripts, the normalized expression data are defined as *Read per Kilobase **transcript** per Million mapped reads* (RPKM).

Useful auxiliary information is a mapping of each sample to one or more labels that represent a treatment, disease or cell-type. That is, each sample j is given a label $l \in l_1, l_2, \dots, l_K$, where K is the number of labels in the data. In our analyses we map each sample to its cell-type.

2.2.1.2. Genomic position data

Enhancer-promoter (E-P) linking requires taking into consideration the genomic positions of the enhancers and their target genes. Each genomic feature (e.g., genes or enhancers) is defined by four fields: chromosome, start position, end position, and the strand. The start position is always smaller than the end position. The strand may be either positive ('+'), negative ('-'), or un-stranded ('*'). In terms of transcription, a positive strand denotes transcription from the start position to the end position whereas the negative strand denotes the reverse. Un-stranded genomic features do not show preference to any direction (e.g., TF binding to the DNA).

Genes (coding and non-coding) are always stranded, whereas enhancers that are transcribed bi-directionally and binding site positions of double stranded DNA (dsDNA) sequence specific TFs are un-stranded.

The common file format for storing genomic positions is the Browser Extensible Data (BED) format. BED format contains, in addition to the previously mentioned position fields and the strand, other informative fields such as the name and BS scores, e.g., intensity and P-value computed by peak calling tool, of the genomic feature and other parameters for controlling the visualization in UCSC genome browser. Additional formats used for representing BSs are narrowPeak (TFs) and broadPeak (histone modifications) formats used in ENCODE consortium [35].

All the above mentioned formats are used and analyzed in this thesis. We used rtracklayer package [77] implemented in R programming language for parsing these formats.

2.2.2. Analysis of GRO-Seq data

GRO-Seq is the main NGS technique used in this work to measure expression and positions of enhancers and genes. This section describes two different approaches for identifying genomic positions of enhancers based on GRO-Seq data. In each approach we shall describe the computational method used and how we used it in our work.

2.2.2.1. De-Novo assembly

The term *de-novo* refers for predicting genomic positions from GRO-Seq aligned reads without comparing to any reference genome. As previously presented in (see **Section 2.1.2.4**), there are two methods that perform de-novo transcript identification based on GRO-Seq aligned reads: groHMM [68] and HOMER [69]. Both methods perform segment-based analysis to classify regions into transcribed and non-transcribed regions. We shall describe only the groHMM method, as it is used in this thesis.

groHMM is a complete pipeline for the (1) accurate identification of the boundaries of transcriptional activity across the genome using GRO-Seq data and (2) classification of these transcription units using a database of available annotations. It is provided as an R package in Bioconductor [68]. groHMM takes as an input read counts from GRO-Seq data in 50 bp windows mapped to the plus and minus strands separately, and then partitions each strand into states representing “transcribed” and “non-transcribed” regions.

groHMM uses a two-states hidden markov model (HMM, see **Fig. 9**). Model parameters include the probability distributions representing the number of GRO-Seq reads each hidden state emits and by a 2x2 matrix of transition probabilities between the hidden states. Gamma distribution was used to model the GRO-Seq read counts due to its flexibility for representing a variety of probability distributions depending on the values of its parameters, shape (k) and scale (θ).

The model parameters are unknown and need to be estimated. The gamma distribution parameters representing read counts in the transcribed state (k_T, θ_T) and the transition probability from the non-transcribed to the transcribed state (N) are trained using the Baum-Welch expectation maximization (EM) algorithm [78]. The tuning parameters, the transition probability of the transcribed state to the non-transcribed state (T) and the variance of the non-transcribed state (σ^2), were estimated using known gene annotations (more details in groHMM paper [68]). T was chosen as a penalty controlling the length of the called transcription units and σ^2 was chosen to control the variance of the GRO-Seq background signal.

After estimating the parameters, the Viterbi algorithm [79] is used to identify the most probable HMM state paths generating the set of primary transcription units given the strand-specific 50bp windows of GRO-Seq read counts.

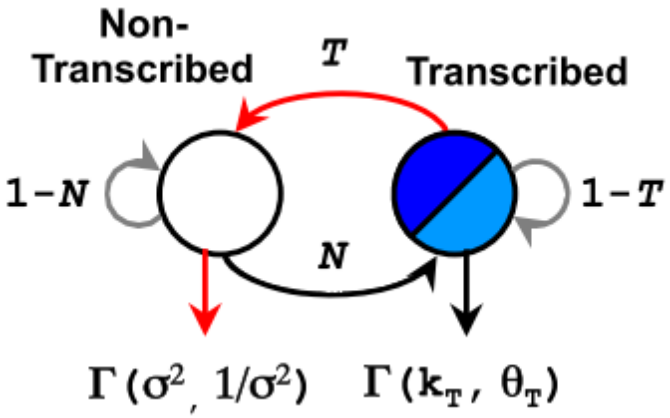


Figure 9. Schematic representation of the groHMM hidden-Markov model approach. The emission probabilities of each state (i.e., transcribed and non-transcribed) were modeled with gamma distributions. Red arrows represent two reserved tuning parameters for the model; T, the transition probability from the transcribed state to the non-transcribed state, and σ^2, the variance of the non-transcribed state in a constrained gamma distribution $\Gamma(\sigma^2,1/\sigma^2)$; $\Gamma(k_T, \theta_T)$, gamma distribution of the transcribed state; N, the transition probability of the non-transcribed to the transcribed state. Gray arrows, self-transition probabilities (i.e., transcribed to transcribed or non-transcribed to non-transcribed), which, by definition, have probabilities 1-T and 1-N, respectively. Source: [68].

2.2.2.2. Transcriptional regulation elements (TREs)

The dREG tool does not use a reference genome. It uses the GRO-Seq expression data as well as regulatory features to train a classifier for identifying transcriptional regulation elements (TREs). dREG (*Discriminative regulatory-element detection from GRO-Seq*) [80] is a sensitive machine learning method that uses support vector regression [81] (SVR) to identify active TREs from data of GRO-Seq, or of a similar technique called precision run-on sequencing (PRO-Seq). dREG detects characteristic patterns of divergent transcription at TREs (see **Fig. 10**).

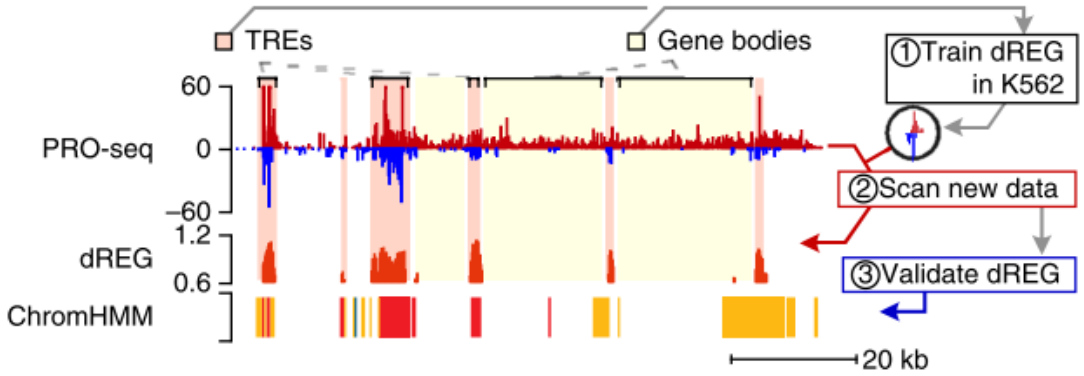


Figure 10. dREG outline. High PRO-seq (or GRO-Seq) signal intensity marks TREs and gene bodies. dREG is a shape detector trained to recognize the characteristic pattern of divergent transcription on the positive (red) and negative (blue) strand near TREs in PRO-seq data (1). After training, dREG can be used to identify TREs using a new PRO-seq data set (red peaks) (2). For comparison, ChromHMM-predicted promoters (red), enhancers (yellow) and insulators (dark blue) are also shown (3). Source: [80].

dREG uses a feature vector that summarizes the patterns of aligned GRO-seq reads near each candidate element at multiple scales (**Fig. 10**). This feature vector consists of read counts for windows in various sizes, standardized using the logistic function, $F(t)$, with parameters α and β, as follows:

$$F(t) = \frac{1}{1 + e^{-\alpha(t-\beta)}}$$

Where t denotes the read counts in each window. The tuning parameters α and β are defined using transformed pair of parameters, x and y, such that x represents the fractional portion of the maximum read count depth at which the logistic function reaches 1 and y represents the value of the logistic function at read count of 0. The relationship of (α, β) to (x, y) is given by the following equations: $\beta = x \max(t)$ and $\alpha = \frac{1}{\beta}\log\left(\frac{1}{y} - 1\right)$ where $\max(t)$ denotes the maximum read depth, as computed separately for each window size and strand in the feature vector. x and y were fixed to 0.05 and 0.01 respectively after selecting different x values and checking the classification accuracy (for more details see [80]). Using this function in its optimized form tends to assign each position in the genome a value near 0 or 1, and, consequently, most of the signal for dREG is dependent on where reads are located rather than on the relative read depths. The feature vector is passed to a SVR, which scores sites with high GRO-seq signal for similarity to a training set of TREs. The goal of the trained classifier is to separate regions of high GRO-Seq signal intensity into a class in which RNA polymerase originates by initiation (i.e., from the TSS) and rapidly transitions to elongation (positive set, composed of TSSs) and a class in which polymerase elongates (negative set, largely composed of gene bodies).

To train the classifier, TREs were identified from GRO-cap data [70] (see **Section 2.1.2.4**). This technique identifies TSSs in promoters and enhancers that were used as positive examples. Regions of matched GRO-seq signal intensity lacking additional marks associated with TREs were used as negative examples. To increase sensitivity, in the training set, the positive set included only positions (labeled as 1) from GRO-cap identified TSSs intersected

with high-confidence DHS peaks. Sites that intersected with any functional mark indicative of regulatory elements (e.g., from known database) but not from GRO-cap peak were removed from the training set. The negative set included all other positions (labeled as 0) in the genome meeting the GRO-Seq high signal requirements and not from a GRO-cap peak.

The training of dREG was done using data from K562 cells. When applying the classifier on new GRO-Seq data, it produces scores between 0-1 (the higher the score the more likely that the position is in a TRE) at each position meeting the GRO-Seq signal intensity thresholds that were a priori defined to remove positions with low signal levels (for more details see [80]). To call dREG peaks, the threshold score was defined as 0.8 (default value in the tool) and adjacent positions meeting this threshold were concatenated.

2.2.3. Genomic data analysis

In this section we describe common types of downstream analysis on expression data used in this thesis.

2.2.3.1. Enrichment analysis

Enrichment analysis methods [2,3,82,4] are used to identify classes of genes that are over-represented in a set of genes produced by a certain analysis. The classes originate from known biology, such as metabolic and signaling pathway, biological processes, and molecular functions.

In this type of analysis, we ask if a group of genes is likely to be related to some biological function or process, or is under common regulation. Enrichment analysis is usually applied to a group of genes detected by differential expression analysis, or a group of co-expressed genes identified by clustering. In this section, we outline the common statistic test used for enrichment analysis, and present two enrichment analysis tools: The TANGO algorithm developed for functional enrichment analysis using gene ontology (GO) and the AMADEUS algorithm for de-novo motif enrichment analyses.

2.2.3.2. The hypergeometric test

This test is the most popular in enrichment analysis. Formally, let G be the underlying set of genes (the *background gene set*), let T be a gene group (the *target gene set*) and let A be an a-priori defined set of genes (the *annotation gene set*). A can denote a set of genes defining a biological process, a pathway, or a target of some regulatory factor. We test the significance of the intersection between T and A.

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

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

When performing multiple tests with different annotation sets $A_i, 1\leq i\leq k$ with the same gene group T, one needs to take also into account the inner dependencies between the tested gene sets. In addition, if N gene groups are analyzed vis-à-vis M a priori gene sets, the number of statistical tests amounts to $N\cdot M$. Hence, multiple testing correction (e.g., Bonferroni [83] or FDR [84]) is mandatory to control the number of false positives (FPs).

2.2.3.3. The TANGO algorithm for GO enrichment analysis

TANGO (Tool for analysis of GO enrichments) [85,86] is a computational tool for enrichment analysis that accounts for the relationships among tested classes. As discussed above, one needs to account for multiple testing when comparing the gene group T with multiple gene sets (classes) $A_i, 1\leq 1\leq k$. However, when the sets A_i are highly dependent (e.g., nested terms in the GO hierarchy, like 'cell cycle' and 'cell cycle regulation'), such correction may be too stringent. To cope with this problem, TANGO computes the empirical distribution of the minimal enrichment P-value by sampling large number of random gene sets and computing their P-values versus each of the GO groups. When annotating several gene sets T_j in a single analysis (e.g., all clusters identified in a data set), TANGO also corrects for the additional multiple testing that takes place.

In practice, TANGO computes corrected P-values for multiple gene sets T_j and classes A_i by estimating the distribution of enrichment P-values in permuted gene sets T_j' of the same size.

TANGO also performs redundancy filtering. When multiple GO terms reflecting a similar function or process have significant corrected P-values, in order to avoid reporting related terms, TANGO performs greedy redundancy filtering based on approximated conditional HG test. Formally, given a target gene set T that is enriched with genes from set A', we test if T is enriched with genes from another set A, assuming we already know the size of the intersection between A' and T and between A' and A:

$$\mathrm{CondP(T,A|A')}=\sum_{\mathrm{x}\geq|\mathrm{T}\cap\mathrm{A}\cap\mathrm{A'}|}\mathrm{P_{hg}}(|\mathrm{A'}|,|\mathrm{A}\cap\mathrm{A'}|,|\mathrm{T}\cap\mathrm{A'}|,\mathrm{x})\times\sum_{\mathrm{k}\geq|(\mathrm{T}-\mathrm{A'})\cap\mathrm{A}|}\mathrm{P_{hg}}(|\mathrm{G}|-|\mathrm{A'}|,|\mathrm{A}-\mathrm{A'}|,|\mathrm{T}-\mathrm{A'}|,\mathrm{k})$$

Where P_{hg} is the HG test (see **Section 2.2.3.2**). TANGO starts with all GO terms that got a significant enrichment for a certain target set T, sorted by their P-values $A_{i_1}\leq A_{i_2}\leq\cdots\leq A_{i_k}$.

Then, the list is traversed and only sets A_i for which $CondP(T, A_i|A_j) < p_{\min}$ for all $j < i$ are reported, where p_{min} is a pre-defined cutoff (usually 0.05).

2.2.3.4. AMADEUS: de-novo motif discovery

Amadeus (A motif algorithm for detecting enrichment in multiple species) [5] is a software suite for efficient genome-scale de-novo detection of enriched sequence motifs. The motifs are assumed to be short sequence patterns that are overrepresented in the promoters or the 3' UTRs of a given set of genes compared to their occurrences in a large background set. Amadeus evaluates the discovered motifs using several statistical tests.

The general architecture of Amadeus is a pipeline of filters, or refinement phases, where each phase receives as input a list of motif candidates and applies an algorithm for refining the list and producing a set of improved candidates, which serves as a starting point for the next phase. The first phases typically work on a very large number of possible motif candidates, such as all possible k-mers (i.e., all sequences of length k over the {A,C,G,T} alphabet), and execute simple procedures for choosing the most promising motifs. More advanced phases merge and shift promising k-mers and then perform EM-like optimization of position-weight matrix (PWM) motifs. The default score for evaluating each candidate motif is the HG P-value (see **Section 2.2.3.2**). Random permutations are also used for multiple testing correction when comparing a single target set of genes T with multiple motif candidates.

2.2.3.5. Unsupervised analysis

Unsupervised analysis of gene or enhancer expression data over a large number of samples calls for inferring hidden structure without using the sample labels. Unsupervised analysis methods rely only on the expression data without any additional information on the samples.

Clustering methods [87–90] aim to discover a subset of genes that manifest a similar expression pattern across the samples, or that are highly correlated across samples. In addition, when the number of samples is large it is common to seek also a subset of genes and a subset of samples that have a similar pattern across genes or across samples, which is a method called *bi-clustering* [91,92]. There are many formulations for the clustering and biclustering problem, and most of them are NP-hard. Therefore, approximations and heuristics are used.

Examples for heuristic clustering methods are: (1) hierarchical methods [93–95] that construct a tree-like structure to represent the relations among genes/samples. For example, average linkage hierarchical clustering builds the tree structure by iteratively selecting the pair of genes or gene groups with the maximum similarity and uniting them to form a new gene cluster. The similarity of this cluster with other objects (genes or clusters) is defined as the average similarity of the cluster's components to the components of the other object. (2) Clustering methods for finding homogenous gene groups that not necessarily cover all genes. For example, Click [90] partitions some (but not necessarily all) genes into clusters using graph-based algorithms. A full graph is built with genes as vertices, and edges with weights that reflect gene similarity (e.g., Pearson correlation between patterns of genes or a probabilistic

score). The goal is to split the graph into modules of similar and high homogeneity scores by applying the minimum weighted cut (i.e., edges in the cut separate the graph into two sub graphs such that their sum of weights across the cut is minimal). SAMBA bi-clustering [91] seeks homogenous biclusters that exhibit similar pattern across two subsets of genes and samples. SAMBA models the input expression data as a bipartite graph whose two parts correspond to samples and genes, respectively, with edges for significant expression similarity. SAMBA outputs heavy sub graphs that correspond to significant biclusters. Click and SAMBA methods hold the potential to remove outlier genes, and therefore are more robust.

2.2.4. Regression analysis

In this chapter we describe the regression methods used in our work. First we introduce the input: Let X_g be a matrix of size $n\ x\ (k+1)$, where n is the number of samples and $k+1$ is the number of independent variables (including a first column of unit vector for the intercept). Let y_g be the n-sample response vector. In our analysis y_g is the expression pattern of gene g, and X_g contains the expression patterns of the k enhancers closest to gene g along the genome.

We want to build a linear model:

$$(1)\ y_g = X_g\beta_g + \varepsilon_g$$

Where ε_g is a vector that denotes the errors of the model and β_g is the $(k+1)\ x\ 1$ vector of coefficients (including the intercept) to be estimated. The goal is to find values $\beta_{g_0}, \beta_{g_1}, \dots, \beta_{g_k}$ such that ε_g is as small as possible. Different methods quantify the error ε_g in different ways.

2.2.4.1. Ordinary least squares (OLS)

OLS is a method for estimating the unknown parameters β_g in a linear regression model, with the goal of minimizing the sum of squares of the differences between the observed responses y_g in the given data set and those predicted by a linear function of the independent variables X_g.

OLS coefficient estimation is addressed by minimizing the error sum of squares (**SSE**) defined as:

$$(2)\ S(b_g) = (y_g - X_g b_g)^T (y_g - X_g b_g)$$

The values of b_g that minimize this sum are called the *OLS estimator* for β_g. The function $S(b_g)$ is quadratic in b_g with positive-definite Hessian (a square matrix of the second derivatives of S function whose determinant is > 0), and therefore this function has a unique global minimum at $b_g = \hat{\beta}_g$, which can be given by the explicit formula (see [96]):

$$(3)\ \hat{\beta}_g = argmin_{b\in\mathbb{R}^k} S(b) = (X_g^T X_g)^{-1} X_g^T y_g$$

After estimating β_g, the predicted values from the regression will be: (4) $\hat{y}_g = X_g\hat{\beta}_g = Py_g$, where (5) $P = X_g\left(X_g^T X_g\right)^{-1}X_g^T$ is the projection matrix onto the space V spanned by the columns of X_g. Another matrix is the *annihilator matrix* (6) $M = I_n - P$, which is the projection matrix onto the space orthogonal to V. Both P and M are symmetric and idempotent (i.e., $P^2 = P$) and follow the identities with respect to X_g: (7) $PX_g = X_g$ and (8) $MX_g = 0$ [96]. The M matrix creates the residuals from the regression:

$$(9)\ \hat{\varepsilon} = y_g - \hat{y}_g = y_g - X_g\hat{\beta}_g \underbrace{=}_{3,6} My_g \underbrace{=}_{1} M\left(\mathrm{X_g}\beta_\mathrm{g} + \varepsilon_\mathrm{g}\right) = \left(\mathrm{MX_g}\right)\beta_\mathrm{g} + \mathrm{M}\varepsilon_\mathrm{g} \underbrace{=}_{8} \mathrm{M}\varepsilon_\mathrm{g}$$

Under these residuals we can estimate the value of σ^2, which is the variance of the random error variable ε_g:

$$(10)\ s^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-k} = \frac{\left(My_g\right)^T My_g}{n-k} = \frac{y_g^T M^T My_g}{n-k} \underbrace{=}_{M^TM=M^2=M} \frac{y_g^T My_g}{n-k} = \frac{S\left(\hat{\beta}_g\right)}{n-k},\quad (11)\ \hat{\sigma}^2 = \frac{n-k}{n}s^2$$

The numerator $n-k$ is the degrees of freedom, s^2 is the OLS estimate for σ^2, and $\hat{\sigma}^2$ is the maximum likelihood estimator (MLE) for σ^2 computed under the assumption that the residuals (or errors) are normally distributed $\varepsilon_\mathrm{g} \sim N(0, I_n\sigma)$. This normality distribution of the residuals also means that the distribution of y_g conditionally on X_g is $y_g|X_g \sim N\left(X_g\beta_g, I_n\sigma\right)$.

The coefficient of determination R^2 (also known as r.squared) is commonly used to access the goodness-of-fit of the OLS regression. R^2 is defined as the ratio of predicted (or explained) variance to the total variance of the dependent variable y_g:

$$(12)\ R^2 = \frac{\sum\left(\hat{y}_{g_i} - \bar{y}_g\right)^2}{\sum\left(y_{g_i} - \bar{y}_g\right)^2} = \frac{SSR}{SSTO} \underbrace{=}_{SSTO=SSR+SSE} 1 - \frac{SSE}{SSTO}$$

Where SSTO is the "total sum of squares" and SSR is the "regression sum of squares". Note that the R^2 depends on the SSE, and therefore is also normally distributed.

2.2.4.2. Regularized regression

Regularization in machine learning refers to a process of introducing additional constraints on the model in order to prevent model over-fitting. Over-fitting results in a model with too many variables with non-zero β_g coefficients relative to the number of samples, making the model too complex. Such a model may have poor predictive performance on new samples.

The OLS regression method will output a model that uses all k independent variables for predicting gene's expression. Therefore, the trained model may be over-fitted to the trained samples and as a consequence may be poorly predictive on new samples. Regularization

methods can help in reducing the number of variables at the cost of increasing the mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_{g_i} - y_{g_i}\right)^2 = \frac{1}{n}S(\hat{\beta}_g)$$

Many regularization methods were suggested to perform a model shrinkage (i.e., reducing the number of variables).

LASSO (Least absolute shrinkage and selection operator) regression, introduced by Robert Tibshirani in 1996 [97], in general forces the sum of absolute values (l_1-norm) of the regression coefficients β_g to be less than a fixed value, which forces certain coefficients to be set to zero, effectively choosing a simpler model that does not include those coefficients. Formally the new objective function will be:

$$(13)\ \hat{\beta}_g = argmin_{b\in\mathbb{R}^k}\{S(b) + \lambda\|b\|_1\},\ \ \|b\|_1 = \sum_{i=1}^{k}|b_i|$$

Where λ is the regularization (or tuning) parameter. If $\lambda = 0$ then equation (13) reduces to the simple OLS regression.

Ridge regression (also known as Tikhonov regularization) [98] is another shrinkage method that was first introduced by Andrey Tikhonov. Ridge differs from Lasso by including in the minimization a term for the sum of squares of the coefficients (l_2-norm), formally:

$$(14)\ \hat{\beta}_g = argmin_{b\in\mathbb{R}^k}\{S(b) + \lambda\|b\|_2^2\},\ \ \|b\|_2^2 = \sum_{i=1}^{k}b_i^2$$

The major difference between Lasso and ridge is that ridge will typically only reduce the coefficients towards to zero but will not set any one of them to zero as Lasso does.

Elastic net (enet) regularization method [99] combines both Ridge and Lasso method together by adding terms for both $l_1 - norm$ and $l_2 - norm$ to the objective function:

$$(15)\ \hat{\beta}_g = argmin_{b\in\mathbb{R}^k}\{S(b) + \lambda_2\|b\|_2^2 + \lambda_1\|b\|_1\}$$

Enet tries to balance between limitations of Lasso and Ridge. In highly correlated group of variables, Lasso will tend to select a single variable from the group and ignoring the others while Ridge will tend to give equal coefficient size to all variables. The R package 'glmnet' [100] provides implementation of elastic-net and uses a single regularization parameter λ such that $\lambda_1 = \alpha\lambda$ and $\lambda_2 = (1-\alpha)\lambda$ where α is a mixing parameter.

2.2.4.3. Generalized linear model (GLM)

GLM [101] is a flexible generalization of the OLS regression that allows the response variables y_g to have error distribution models other than a normal distribution. The GLM generalizes the linear regression by allowing the linear model to be related to the response variable via a *link function*, which models the relationship between the linear predictor and the mean of the assumed distribution.

The simple OLS regression model assumes that the errors ε_g are normally distributed. The normality assumption is appropriate when the original values in the response variable can vary essentially indefinitely in either direction with no upper or lower bound, or more generally it is suitable for any quantity that varies by a relatively small amount, e.g., human heights. However, the normality assumption does not always hold. For example, in cases where the response variable is expected to be always positive and varying over a wide range, e.g., enhancers have typically low expression than genes, and a gene model prediction for a little change in enhancer expression might lead to higher change in gene expression.

2.2.4.4. The Poisson and Negative Binomial model (NB2)

In this section we describe GLM methods whose link function differs from the normal distribution. First we start with the Poisson distribution. In the *Poisson distribution*, the conditional means and the conditional variances are equal, formally:

$$(16)\ \mathrm{E}[y_g|\mathrm{X_g}] = \mathrm{Var}[y_g|\mathrm{X_g}] = \mu_g(\mathrm{X_g})$$

Where μ_g represents the n-sample expected mean values of gene g. However, eq. (16) often breaks in counts data and the conditional variance exceeds the conditional mean, a situation called over-dispersion. This may be due to unknown sub-populations of samples (replicates) that have different variability from others (this phenomenon is also known as Heteroscedasticiy). To address this problem, the *negative binomial* (NB) distribution was proposed by Jerald Lawless in 1987 as a generalization of the Poisson distribution [102].

The GLM.NB method allows estimating the dispersion parameter, α_g for gene g, that measures the mean-variance relationship. Lawless purposed to add an additional variable to the variance, as follows:

$$(17)\ \mathrm{E}[y_g|\mathrm{X_g}] = \mu_g(\mathrm{X_g}), \qquad \mathrm{Var}[y_g|\mathrm{X_g}] = \mu_g(\mathrm{X_g}) + \alpha_g\mu_g^2(\mathrm{X_g})$$

Where $y_g \sim \mathrm{NB}(\mu_g(\mathrm{X_g}), \alpha_g)$. Note that in the limit $\alpha_g \to 0$ we get the Poisson model $y_g \sim \mathrm{Poisson}\left(\mu_g(\mathrm{X_g})\right)$. Instead of fitting a linear model to y_g we fit a linear model to $\mu_g(\mathrm{X_g})$ as follows:

$$(18)\ \ln\left(\mu_g(X_g)\right) = X_g^T\beta_g + \ln N$$

Or equivalently:

$$(19)\ \mu_g(X_g) = N \cdot \exp\{X_g^T\beta_g\}$$

Where N is the n-sample vector of library sizes (i.e., total mapped reads). The formulation of $\mu_g(X_g)$ using the exponent was done since $\mu_g(X_g)$ should be positive for count data. The NB density function in the exponential family is:

$$(20)\ f(y_g|\mu_g,\alpha_g) = exp\left\{-\alpha_g^{-1}\ln(1+\alpha_g\mu_g) + \ln\left(\frac{\Gamma(y_g+\alpha_g^{-1})}{\Gamma(y_g+1)\Gamma(\alpha_g^{-1})}\right) + y_g\ln\left(\frac{\alpha_g\mu_g}{1+\alpha_g\mu_g}\right)\right\}$$

Where $\Gamma(\cdot)$ is the gamma function. We estimate the log-likelihood function:

$$(21)\ \ln L(\beta_g,\alpha_g) = \sum_{i=1}^{n}\left\{-\alpha_g^{-1}\ln(1+\alpha_g\mu_{g_i}) + y_{g_i}\ln\left(\frac{\alpha_g\mu_{g_i}}{1+\alpha_g\mu_{g_i}}\right) + b(y_{g_i},\alpha_g)\right\}$$

Where $b(y_{g_i},\alpha_g) = \ln\left(\frac{\Gamma(y_{g_i}+\alpha_g^{-1})}{\Gamma(y_{g_i}+1)\Gamma(\alpha_g^{-1})}\right)$.

The first order conditions of Eq. 21 yield unsolvable formula. To cope with this problem, an iterative process is used to maximize Eq. 21. In R 'MASS' package [103], the iterative reweighted least squares (IRLS) method is used. We used the glm.nb function in 'MASS' package to build the GLM.NB model.

2.2.4.5. The zero-inflated negative binomial model (ZINB)

Large-scale count data often suffers from a problem of frequent zero-valued observations. This phenomenon is called *zero inflation* (ZI). For example, the number of insurance claims within a population for certain type of risk would be zero-inflated by those people who have not taken out insurance against the risk and thus are unable to claim. The number of times that a gene is expressed within a subset of samples from the same cell-type will be zero-inflated by the other samples not from the same cell-type that do not express the same gene due to different gene regulation. GLM.NB is unable to solve the problem, hence, we shall introduce an extension of the approach that also models the issue.

Zero-Inflated negative binomial (ZINB) and *zero-inflated Poisson* (ZIP) regression analyses [104,105] were formulated to handle the ZI issue. We shall introduce the ZINB GLM approach. Suppose we have a degenerate probability mass function $f_1(y_{g_i})$ with $f_1(j) = 1$ if $j = 0$ and $f_1(j) = 0$ if $j > 0$, and the NB base count density is $f_2(y_{g_i})$ with the support $y_{g_i} \in \{0,1,2,\dots\}$ for

sample $1 \leq i \leq n$. In ZI models we add a separate component, π_g, that inflates the probability of a zero. Then, the ZI model will be:

$$(22)\ \Pr[y_{g_i} = j] = \begin{cases} \pi_{g_i} f_1(0) + (1-\pi_{g_i}) f_2(0) & \text{if } j = 0 \\ \pi_{g_i} f_1(j) + (1-\pi_{g_i}) f_2(j) & \text{if } j > 0 \end{cases} = \begin{cases} \pi_{g_i} + (1-\pi_{g_i}) f_2(0) & \text{if } j = 0 \\ (1-\pi_{g_i}) f_2(j) & \text{if } j > 0 \end{cases}$$

In (22) the proportion of zeros, π_{g_i}, is added to the baseline distribution, and the distributions from the base model, $f_2(y_{g_i})$, are decreased by a proportion of $(1-\pi_{g_i})$. The probability, π_{g_i}, may be set as constant or depend on x_{g_i}, β_g via a binary outcome model such as the logit model. The logit model is the canonical link function of the Bernoulli distribution in the GLM. Suppose we model π_{g_i} as the probability of success, which is $y_{g_i} = 0$, and $1-\pi_{g_i}$ as the probability of failure, which is $y_{g_i} > 0$, then from the inverse-logit function (i.e., the logistic function) we get:

$$(23)\ \begin{aligned} \pi_{g_i}(x_{g_i}, \beta_g) &= N_i\left(1 + \exp\{-x'_{g_i}\beta_g\}\right)^{-1} \\ 1 - \pi_{g_i}(x_{g_i}, \beta_g) &= 1 - N_i\left(1 + \exp\{-x'_{g_i}\beta_g\}\right)^{-1} \end{aligned}$$

Where N_i is the sample i library size.

Maximum Likelihood estimation

We define a binary censoring indicator for gene g and sample i:

$$(24)\ d_{g_i} = \begin{cases} 1 & \text{if } y_{g_i} > 0 \\ 0 & \text{if } y_{g_i} = 0 \end{cases}$$

Where $d_{g_i} = 0$ with probability π_{g_i} and $d_{g_i} = 1$ with probability $1-\pi_{g_i}$. Then, the density for single sample observation is:

$$(25)\ f(y_{g_i}) = \left[\pi_{g_i} + (1-\pi_{g_i}) f_2(0)\right]^{1-d_{g_i}} \times \left[(1-\pi_{g_i}) f_2(y_{g_i})\right]^{d_{g_i}}$$

The log-likelihood function, using (23) and the base density $f_2(y_{g_i}|x_{g_i}, \beta_g, \alpha_g)$, is:

$$(26)\ \begin{aligned} \mathcal{L}(\beta_g, \alpha_g) &= \sum_{i=1}^{n} (1-d_{g_i}) \ln\left[\pi_{g_i}(x_{g_i}, \beta_g) + \left(1-\pi_{g_i}(x_{g_i}, \beta_g)\right) f_2(0|x_{g_i}, \beta_g, \alpha_g)\right] \\ &+ \sum_{i=1}^{n} d_{g_i} \ln\left[\left(1-\pi_{g_i}(x_{g_i}, \beta_g)\right) f_2(y_{g_i}|x_{g_i}, \beta_g, \alpha_g)\right] \end{aligned}$$

Where from the NB probability mass function:

$$(27)\ \begin{aligned} f_2(0|x_{g_i}, \beta_g, \alpha_g) &= \left(1 + \alpha_g \mu_{g_i}\right)^{\alpha_g^{-1}} \\ f_2(y_{g_i}|x_{g_i}, \beta_g, \alpha_g) &= \frac{\Gamma(y_{g_i} + \alpha_g^{-1})}{\Gamma(\alpha_g^{-1})\Gamma(y_{g_i}+1)} \left(\frac{\alpha_g^{-1}}{\alpha_g^{-1} + \mu_{g_i}}\right)^{\alpha_g^{-1}} \left(\frac{\mu_{g_i}}{\alpha_g^{-1} + \mu_{g_i}}\right)^{y_{g_i}} \end{aligned}$$

Starting parameter values for Eq. 26 are computed by first maximizing NB (f_2) and ZI (f_1) models separately using the IRLS method as done in **Section 2.2.4.4**. Eq. 26 is then maximized for both NB and ZI models together given the starting parameter values using the quasi-Newton method (also known as a variable metric algorithm) in an iterative-like algorithm suggested by Broyden-Fletcher-Goldfarb-Shanno (BFGS) in 1970 [106]. We used the zeroinfl function in R package 'pscl' to build the ZINB model [107].

3. Computational Procedures

This chapter describes the computational procedures performed and the data used in this thesis. First we describe in detail the GRO-Seq preprocessing. Then we explain how we compared two methods for enhancer identification, and show how to use regression-based analysis to link proximal enhancers to their target genes. Finally we briefly show as a proof of concept how we can utilize E-P links in enrichment analyses.

3.1. A new workflow for preprocessing GRO-Seq data

GRO-Seq raw data was downloaded from the Gene Expression Omnibus (GEO) database. Specifically, the raw data of each sample is stored in a file containing the sequenced reads and their base call quality score in FASTQ format. Files were automatically downloaded from sequence read archive (SRA) DB linked to the GEO DB using the SRA toolkit. Preprocessing is done in two steps: (1) Single study preprocessing and (2) Study collation. **Figure 11** outlines the preprocessing steps.

We downloaded 366 profiles (i.e., files) from the SRA DB and 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 246 samples from 23 cell lines, each examined under control and stress conditions (see **Supplementary Table S.1**).

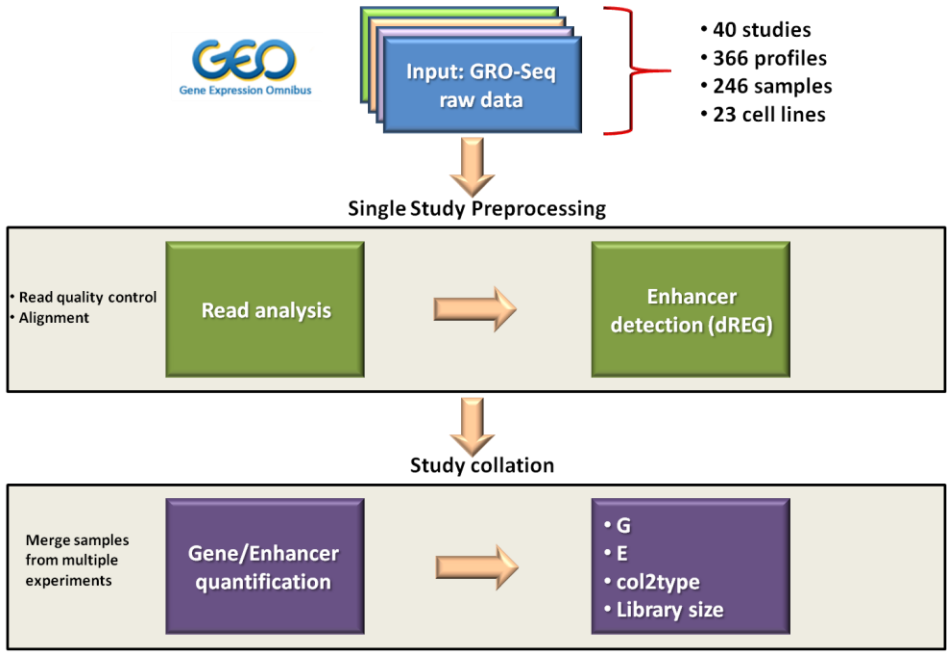


Figure 11. The preprocessing workflow. GRO-seq raw data were downloaded from GEO. Single study preprocessing: Read analysis, included read quality control (QC) and alignment to hg19 reference genome. Detecting of enhancers was done using dREG. Study collation: Quantification and normalization of read alignments to RefSeq genes and detected enhancers produced four data types: G/E are the genes and enhancers expression matrices, col2type maps each sample (column) in G/E to its cell line type and library size is n-sample vector of total mapped reads to the genome.

3.2. Single sample preprocessing and analysis

3.2.1. Read quality control (QC)

Read quality control (QC) is a crucial step that should be applied prior to aligning reads to the reference genome. Failing to perform a proper read QC may result with either low number of aligned reads and/or with misaligned reads. In addition, the same QC should be applied on each profile in order to prevent read biases when comparing multiple samples in downstream analyses.

We used Trimmomatic, a tool for read quality control and trimming [108]. This tool analyzes each read and removes (1) bases from Illumina Tru-seq adapters, and (2) bases with low base quality scores (Phred33+ scores) from both ends. In addition, reads with high proportion of low quality scores are excluded. We adapted Trimmomatic to our GRO-seq data as follows. For (1) we selected all Illumina Tru-Seq adapters. For (2) we trimmed bases with quality score below 5 from the 5' and 3' ends. As an additional filter, whenever a read had four consecutive bases with an average score < 15, we trimmed the read from these bases till the 3' end. Reads with length < 30 bases were excluded. Finally, we cropped all surviving reads to the first 30 bases from the 5' end. By performing trimming to 30 bases we reduce biases caused by different read length between experiments.

3.2.2. Read alignment and filtering

The majority of extracted RNA from transcriptomic experiments like GRO-seq is ribosomal RNA (rRNA), sometimes constituting up to 90% of the reads [109]. The observed portion of rRNA is still high even when various treatment procedures are applied prior to sequencing [110]. For example, in our analyzed data, two HCT116 cell line replicates had 19% and 33% rRNA reads. Thus, even controlled replicated samples may have a great difference in rRNA abundance. Failing to remove most of the rRNA reads might lead to erroneous conclusions in downstream analyses [111].

We detected putative rRNA reads by aligning all reads to a set of known rRNA genes taken from NCBI (rRNA gene symbol ids: RN18S1, RN28S1, RN5, and RN5S17) using bowtie2 (default parameters). All the reads that were mapped to rRNA genes were discarded, and the rest were aligned (using bowtie2 with default parameters) to the reference genome (hg19). For subsequent analyses we used only reads that had a MAPQ score greater than 10.

3.2.3. Single sample analysis

We detected high quality reads and aligned them to the reference genome using Bowtie2 [112] (for more details see **Sections 3.2.1-3.2.2**). We applied dREG [32] on the aligned reads to detect transcriptional regulation elements (TREs). dREG assumes that TREs have a symmetric forward and reverse read coverage relative to their center position. This symmetry is a known marker of short putative enhancers [65]. In our analyses we show that this approach outperforms

groHMM [113] in detection of active enhancers (for more details see **Section 3.5**). Finally, we quantify gene expression by counting the number of reads mapped into each (unspliced) gene. As gene models we used a single transcript per gene derived using groHMM makeConsensusAnnotations R function on hg19 UCSC refGene table, producing 22,891 consensus genes. We only used reads mapped to the gene's transcript body (0.5kb to 20kb downstream of the TSS). If the transcript's length was less than 20kb then we used only the region until the transcript termination site (TTS).

3.3. Joint analysis of multiple samples

Using the results of the single-sample analysis, we merged overlapping TREs detected in different samples to create *merged TREs* (mTREs, see **Fig. 12**). For each mTRE we calculated its read count in each sample using bedtools [73]. We partitioned the mTREs into four types using GenomicRanges [114]: (1) Intergenic: mTREs whose center is located at least 5kb from the closest gene's transcript start site (TSS) and does not overlap any gene's transcript body, (2) Promoter: mTREs whose center is located at most 5kb upstream from the closest gene's TSS and does not overlap any gene's transcript body, (3) Exonic: mTREs that have some overlap with an exon, and (4) Intronic: mTREs that are not exonic and have overlap with an intron of a gene. For the next steps, we define an enhancer as an mTRE that is either intergenic or intronic. Note that these are actually putative enhancers.

The gene expression count profiles were summarized as a genes x samples matrix G. Similarly, the enhancer expression count profiles were summarized as an enhancers x samples matrix E.

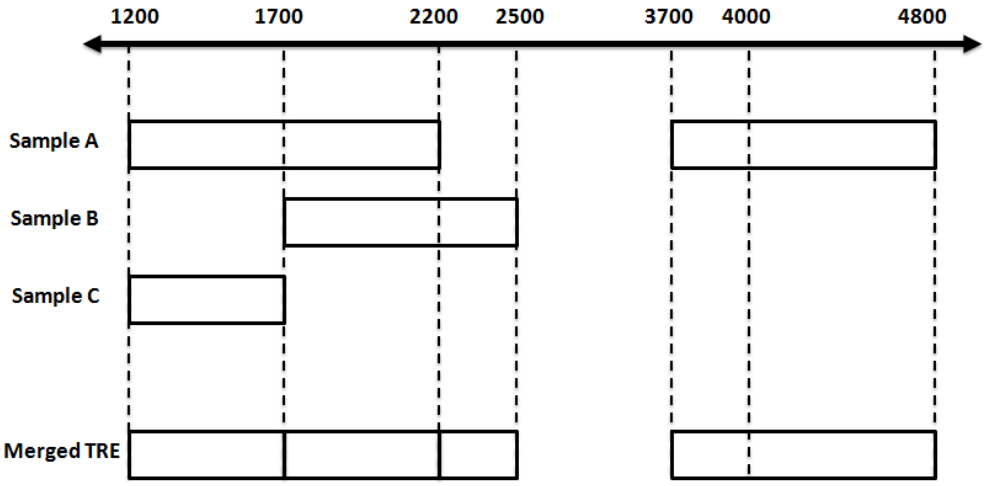


Figure 12. Merging TREs. The samples A,B,C have overlapping TREs spanning the region 1200-2500. These TREs are merged into a single merged TRE between 1200-2500. Sample A has a distinct TRE in the region 3700-4800 with no overlaps in other samples. Therefore, the second merged TRE constitutes a single TRE.

3.4. Gene and Enhancer quantification and normalization

We normalized the expression of each gene/enhancer in each sample by computing its reads per kilobase transcript per million mapped reads (RPKM). The corresponding normalized gene and enhancer matrices are G^n and E^n, respectively. The gene and enhancer expression matrices were further filtered to include only genes/enhancers (rows) with at least one sample (columns) with RPKM ≥ 1 to preserve only expressed genes/enhancers. Next, a common practice is to focus on the k% genes with the highest variance. However, the variance is positively related to expression level (see **Supplementary Fig. 1a**), and therefore, selection by variance gives preference to highly or lowly expressed genes. Instead, we considered using the coefficient of variation (CV), which is the ratio between the gene's standard deviation σ to the mean μ, rather than the variance. We still found a strong relationship between the CV and the lowly expressed genes (See **Supplementary Fig. 1b**). These genes had high CV values since they were noise sensitive. In order to reduce preference to highly or lowly expressed genes, we applied the following procedure: (1) we partitioned the genes according to their mean RPKM expression into 20 bins. (2) In each bin we retained the genes with CV above the median CV of the bin. **Supplementary Figure 1b** shows that the binning procedure filtered out genes that were not variable across samples and had low expressions. The final gene matrices contained 8,360 genes, and the final enhancer matrices contained 255,925 enhancers.

3.5. Enhancer detection and quantification

We compared two methods for finding putative enhancers. The first, dREG [32], uses a support vector machine (SVM) classifier that was trained on K562 nuclear run-on sequencing (NRO-seq) data. This classifier combines multiple window sizes at different symmetric read coverage scales on the +/- strands in order to distinguish between TREs and non-TREs. dREG applies this classifier to a new GRO-seq read data and outputs a set of un-stranded TREs. The second tool, groHMM [68], uses a two-state hidden Markov model (HMM) to distinguish between transcribed and non-transcribed regions in the genome given GRO-seq coverage data and outputs stranded de-novo transcripts (DNTs).

We identified intergenic DNTs using groHMM as follows: (1) As an input to groHMM, we provided a BAM file of a single sample (control or treatment). (2) We identified intergenic DNTs whose TSS is located at least 5kb from a gene's TSS/TTS. (3) We identified two divergent intergenic DNTs on different strands and the distance between their TSSs is at most 2kb. Such pair is considered a single divergent DNT.

To evaluate the quality of each tool's predictions of enhancers we used open chromatin DNase-seq data as an external validation source. Open chromatin regions are more likely to harbor active enhancers [14] and therefore high overlap of predicted enhancers with open chromatin regions is indicative of good prediction. We also used ChIP-seq epigenetic signals of p300/H3K4me1/H3K27ac (see **Supplementary Table S.2**) available for MCF7 and HCT116 cell lines. We denote the DNAse hypersensitivity (DHS) peaks that overlap with at least one intergenic TRE of dREG as *DHS-TREs*, and DHS peaks that overlap with at least one intergenic

divergent DNT as *DHS-DNTs*, and DHS peaks that overlap both intergenic TRE of dREG and intergenic divergent DNT of groHMM as *DHS-TRE-DNT*.

Expression of putative enhancers was quantified using Bedtools. In intergenic mTREs, we counted reads falling on both genomics strands. In intronic mTREs, we counted only reads falling on the gene's antisense strand. Reads that fall within the gene's sense strand were counted for quantifying the gene's expression.

3.6. Enhancer-Promoter mapping via regression analysis

In this section we describe methods we used for inference of enhancer-promoter (E-P) mapping. Only intergenic mTREs or intronic mTREs were used in the analyses below.

We used regression analysis to learn predictive models for a gene's expression pattern given the expression of its proximal enhancers. The input for analyzing a gene g is a vector $y_g \in \mathbb{R}^n$ with the expression levels of g in each of the n samples, and a matrix $X_g \in \mathbb{R}^{nxk}$ containing the expression values of the k closest enhancers in the same samples within a window of 1 Mb (± 500 kb) around the gene's TSS (we used $k = 10$ throughout). We tested three regression models: ordinary least squares (OLS), generalized linear model with the negative binomial distribution (GLM.NB) and zero-inflated negative binomial (ZINB). For OLS, the values in y_g and X_g were normalized using the sample library sizes. ZINB and GLM.NB, on the other hand, directly model the read counts and use the library size as an "offset". 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 expression in the gene. See **Section 2.2.4** for details on each method.

We used *leave-cell-type-out cross validation* (LCTOCV) per gene and per regression method when evaluating the prediction power of each model, in order to avoid overfitting.

3.6.1. Validation

We used two non-parametric measures to evaluate the agreement between the predicted values of g's expression y_g^p and its normalized expression profile y_g^o on left-out samples.

The ***binarized expression validation*** tested whether y_g^p discriminates between the samples in which g was expressed (≥ 1 RPKM, denoted as positives) and the samples in which g was unexpressed (< 1 RPKM, negatives). We quantified this difference using a two-sided Wilcoxon rank sum test.

The ***expression level validation*** calculated for the positive samples the significance of the Spearman correlation between y_g^p and $y_{g_i}^o$. Spearman correlation compares between the ranks of the original and predicted expressions.

A good gene model should discriminate well between positive and negative samples and preserve the original expression rank of the positive samples. The P-values obtained above were corrected for multiple testing using the Benjamini and Yekutieli (BY) FDR procedure [115].

3.6.2. Feature selection

Our next goal was to select informative enhancers for each gene 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. We applied BY correction on the P-values produced by all models together and selected enhancers with FDR ≤ 0.01. To identify the most relevant enhancers for each gene we applied elastic-net model shrinkage (**enet**) with a regularization parameter λ. We used the glmnet function [100] with elastic mixing parameter $\alpha = 0.5$, giving equal weight to the Lasso and Ridge regularizations. We required that all the enhancers that survived eBY filtering will be included in the shrunken model. To achieve this we took 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 gene that had a model after the two validation tests (**Section 3.6.1**) also had a model following the enet step.

We call the complete method, which combines OLS, leave-cell-type-out cross validation, elastic net and significance correction for multiple testing FOCS (FDR-corrected OLS with Cross-validation and Shrinkage).

3.7. Downstream analysis

This section describes the methods used for analyzing our inferred E-P links.

3.7.1. External validation

In order to assess the performance of our method compared to other methods for E-P linking, an external validation is needed. We used two external data types to determine the percentage of E-P links supported by them: (1) ChIA–PET interactions and (2) 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]. We used the liftOver tool (from Kent utils package provided by UCSC) to transform the genomic positions of the interactions from hg38 to hg19.

2,283,827 unique eQTL SNPs covering 44 different tissues were downloaded from GTEx portal [1]. We used the significant SNP-gene pairs from GTEx analysis V6 and V6p builds.

We used TSS intervals (±500 bp upstream/downstream) for the promoters and supplied the enhancers' genomic intervals (±500 bp from the enhancer center) as the regulatory regions. An E-P pair is supported by a particular ChIA-PET interaction if and only if the promoter and enhancer intervals overlap different anchors of an interaction. An E-P pair is supported by eQTL SNP if and only if the SNP occurs in the enhancer's interval and is associated with the promoter's gene. For each considered pair of promoter and enhancer we checked if their intervals are supported by ChIA-PET and eQTL data. We then measured the fraction of E-P pairs supported by these external data.

To get an empirical p-value for the significance of the fraction, we performed 100 permutations on the data. In each permutation, for each promoter independently, if it had k E-P links, then k enhancers with similar distances from the gene's TSS as the true k enhancers were selected randomly on the same chromosome as its E-P links. For this purpose we used the R 'Matching' package [117]. The fraction of overlap with the external data was computed on each permuted data.

3.7.2. Functional genomics

An advanced research of E-P links is how we can utilize them for downstream analyses such as motif and GO enrichment analyses. It is still not well understood how to use E-P links for inferring deeper biological insights given expression data of genes and/or enhancers. Here we shell describe in short as proof of concept how to start such research that will be further investigated during the author PhD degree.

3.7.2.1. Preprocessing and clustering analysis

We assume that we are given a GRO-Seq RPKM normalized gene expression data, $M^n_{g\,x\,p}$ (n denotes normalized) with g genes and p samples, of a single cell-type dataset.

The preprocessing step is done as follows: first, we apply quantile normalization on $M^n_{g\,x\,p}$ in order to compare between samples with distributions identical in the statistical properties [118]. Second, we retained genes manifesting at least two-fold change (FC) in expression across all samples relative to a selected control sample. This will allow us to focus our analyses on informative genes that their expression was changed due to treatment.

Next, we used gene clustering analysis to find clusters of similar gene expression pattern across all samples. Each cluster may contain unique biological regulation that governs genes in the cluster leading to their similar expression pattern across samples. Such biological regulation could be a set of TFs regulating the gene expression, and/or a common function of the genes in some process (e.g., cell cycle). We used the Click clustering algorithm (default parameters; see **Section 2.2.3.5**) implemented within Expander tool [85].

3.7.2.2. Downstream enrichment analysis

We performed de-novo motif finding and GO enrichment analysis on the gene clusters from **Section 3.7.2.1**. In addition, for each gene cluster we created the set of enhancers that have link to the genes in the cluster according to the predicted E-P links. Den-novo motif finding was also applied on each cluster of enhancers.

We used AMADEUS (see **Section 2.2.3.4**) for motif finding. Motif lengths allowed were 9/10 bps for enhancers/promoters, respectively. The JASPAR 2016 database [119] was used to find associated TFs to enriched motifs. In addition, we performed GO functional enrichment analysis using TANGO (default parameters; see **Section 2.2.3.3**) implemented in Expander batch mode tool [85,86,120].

We selected as promoters a region of -300/100 upstream/downstream of the gene's TSS and for enhancers we selected a region of 400 bp around their center. We selected equal widths (400 bp) for both promoters and enhancers in order to find motifs that are appear or do not appear in both promoters and enhancers. Background set for promoters was all genes that passed the FC criteria from **Section 3.7.2.1**. Background set for enhancers was the four adjacent regions of 400 bp to each enhancer target (300 bp separation between each start and end of different regions). For TANGO, the target set was a gene cluster and the background set included all genes that passed the FC criteria from **Section 3.7.2.1.**

4. Results

This chapter describes the results of our analysis of GRO-Seq data. **Sections 4.1-4.3** describe data preprocessing and enhancer detection methods. In **Section 4.4** we show how to filter false positively detected enhancers. **Section 4.5** compares the performance of enhancer-gene mapping methods. In **Section 4.6** we describe as a proof of concept the results achieved from analyzing a single GRO-Seq dataset of the same cell-type. The methods described in **Sections 4.4** and **4.6** are still under development.

4.1. GRO-Seq data

We collected 366 GRO-Seq profiles covering 23 different cell-lines from the Gene Expression Omnibus (GEO) database [121] and applied the preprocessing pipeline described in **Section 3.1** on each profile. After pooling together aligned reads from profiles with the same profile ID (via GEO GSM identifiers) we had 246 samples. Using the pipeline described in **Section 3.2** on each sample, we identified enhancer regions and quantified enhancer and gene expressions. The final G and E matrices contained 8,360 genes and 255,925 enhancers, respectively, across 246 samples.

4.2. Trimming reads improves mapping

We used Trimmomatic for read quality control and trimming [108] after adapting the tool to our GRO-seq data (see **Section 3.2.1**). On most samples (292 out of 356), the approach increased the number of aligned reads (**Fig. 13**). For example, for the GM12878 cell line, out of ~205M raw reads, only ~20M were initially mapped, but the number increased to ~114M after applying our procedure. The improvement is mainly due to removal of adapter sequences, which confuses the alignment to the genome. In some rare examples (e.g., A549 cell line samples) the method reduced the number of aligned reads.

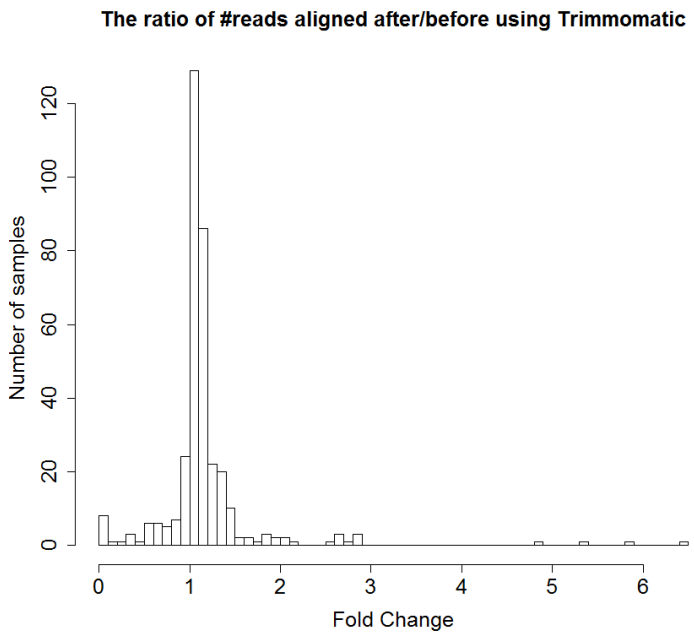


Figure 13. Read mapping. Fold change (FC) was computed by dividing the number of aligned reads after and before using Trimmomatic. For most of the samples the process increased the number of aligned reads (FC>1 for 292 out of 356 replicates). Samples from A549 cell line showed the most extreme decrease in the number of aligned reads (FC<0.03).

4.3. Enhancer detection method

We tested which method for identifying genomic positions of putative enhancers from GRO-Seq expression data should be used in our analysis. We compared two methods for enhancer detection: dREG [32] and groHMM [68] (see **Section 3.5**). To evaluate the quality of each tool's predictions of eRNA expressed regions we used two data types for external validation: open chromatin data, which is measured using DNase-Seq, and epigenetic marks of active enhancers (ChIP-Seq of p300/H3K4me1/H3K27ac) available for MCF7 and HCT116 cell lines (see **Supplementary Table S.2**). Open chromatin regions are more likely to harbor active enhancers [14] and therefore high overlap of enhancer predictions with open chromatin regions is indicative of good prediction. We denote the DNAse hypersensitivity (DHS) peaks that overlap with at least one intergenic TRE of dREG as *DHS-TREs*, DHS peaks that overlap with at least one intergenic divergent DNT as *DHS-DNTs*, and DHS peaks that overlap with both as *DHS-TRE-DNT*.

The median epigenetic signals of p300, HEK4me1 and H3K27ac were much higher around the DHS peak center of DHS-TREs compared to DHS-DNTs (see **Fig. 14**, **Supplementary Figure S.2** and **Supplementary Table S.3**). The number of DHS-TREs was almost twice the number of DHS-DNTs (see **Table 1**). Moreover, the percentage of dREG TREs that overlap DHS peaks was higher than that of groHMM DNTs that overlap DHS peaks (46%-48% vs. 35%-36%, see **Table 1**). Hence, dREG predictions achieve higher agreement with epigenetic markers than groHMM predictions, both in quality and quantity, suggesting a higher performance in detecting true active enhancers. The DHS-TRE and the DHS-DNT sets overlapped by ~50% (control) and

~51% (treatment), suggesting that both methods capture some enhancers with epigenetic signals. As our results suggest that using the enhancers detected by groHMM DNTs may result in higher error rate, we chose to use dREG for all subsequent analyses.

Table 1. Comparison of dREG and groHMM: results of the MCF7 cell line			
Sample type	**Group**	**Set size**	**Covered enhancers (%)***
Control	DHS-TRE	4,159	48.5
Control	DHS-DNT	2,708	35.1
Control	DHS-TRE-DNT	979	49.7
Treatment	DHS-TRE	4,262	46.3
Treatment	DHS-DNT	2,347	36.1
Treatment	DHS-TRE-DNT	790	50.7
* The percent of reported enhancers that had any overlap with some DHS peak. For DHS-TRE-DNT we computed the percentage as follows: (1) we found the set E of dREG enhancers that overlap groHMM enhancers, (2) we found the set $\boldsymbol{E}' \subseteq \boldsymbol{E}$ of enhancers that overlap some DHS peak in DHS-TRE-DNT group, and (3) we computed the percentage by dividing $\lvert\boldsymbol{E}'\rvert$ with $\lvert\boldsymbol{E}\rvert$.			

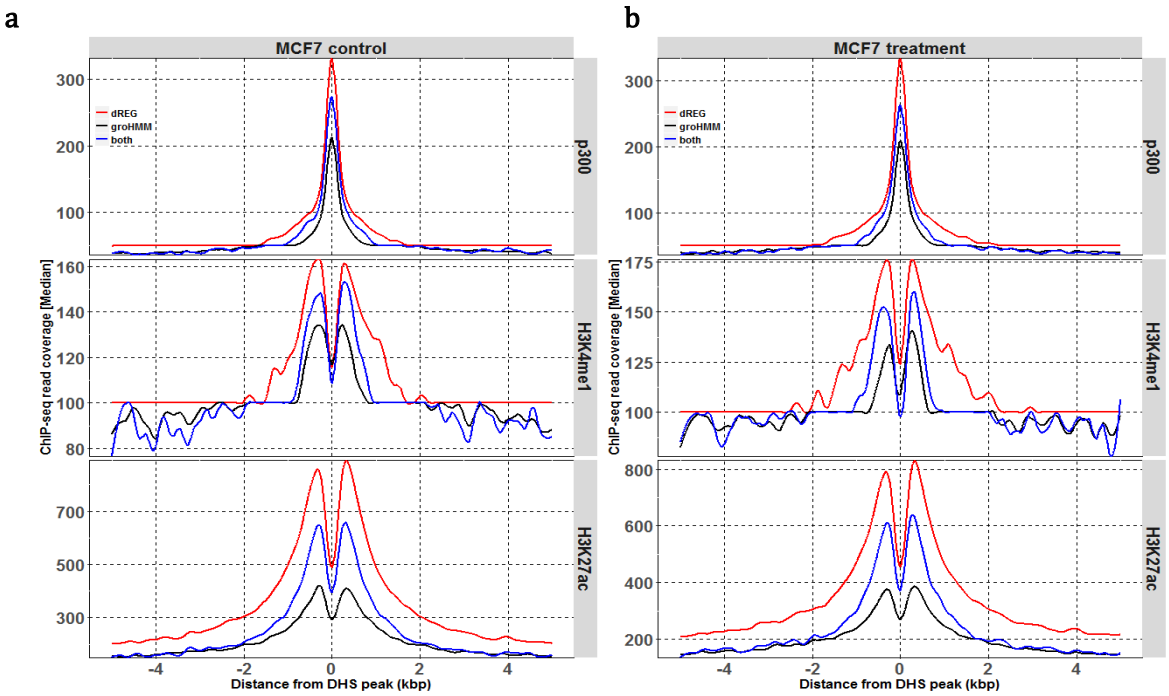


Figure 14. Epigenetic marks: ChIP-seq median read coverage across DHS peaks. a/b MCF7 control/treatment plots. Red, black, and blue curves show median read coverage for enhancers predicted by dREG, groHMM, and both dREG and groHMM, respectively, that overlap DHS peaks. Rows correspond to the epigenetic markers p300, H3K4me1 and H3K27ac. Negative/Positive distances from DHS center denote upstream/downstream distances, respectively. The results show that the regions detected by dREG manifest much higher ChIP-seq signals than the regions detected by groHMM.

4.4. Filtering false positive putative enhancers

We aimed to clean from dREG output false positive TREs. We used DHS data to partition the TREs in each sample into two classes: (1) TREs that overlap DHS peaks and (2) TREs that do not overlap DHS peaks. We used data of five cell lines for which DHS data were available (see **Supplementary Table S.2**). As a preliminary test we checked whether class 1 manifests a different GRO-seq read coverage behavior compared to class 2. Inspection of coverage plots around the TRE center showed a sizeable difference in the mean and a slighter difference in the median and in the 75 quantile coverage between the classes (see **Fig. 15**, and **Supplementary Figure S.3**). Moreover, we observed consistent differences in the intensity of the tails, which were lower in class 1 compared to class 2 (**Fig. 15b**). This suggests that the TRE length is an important property that can be further exploited.

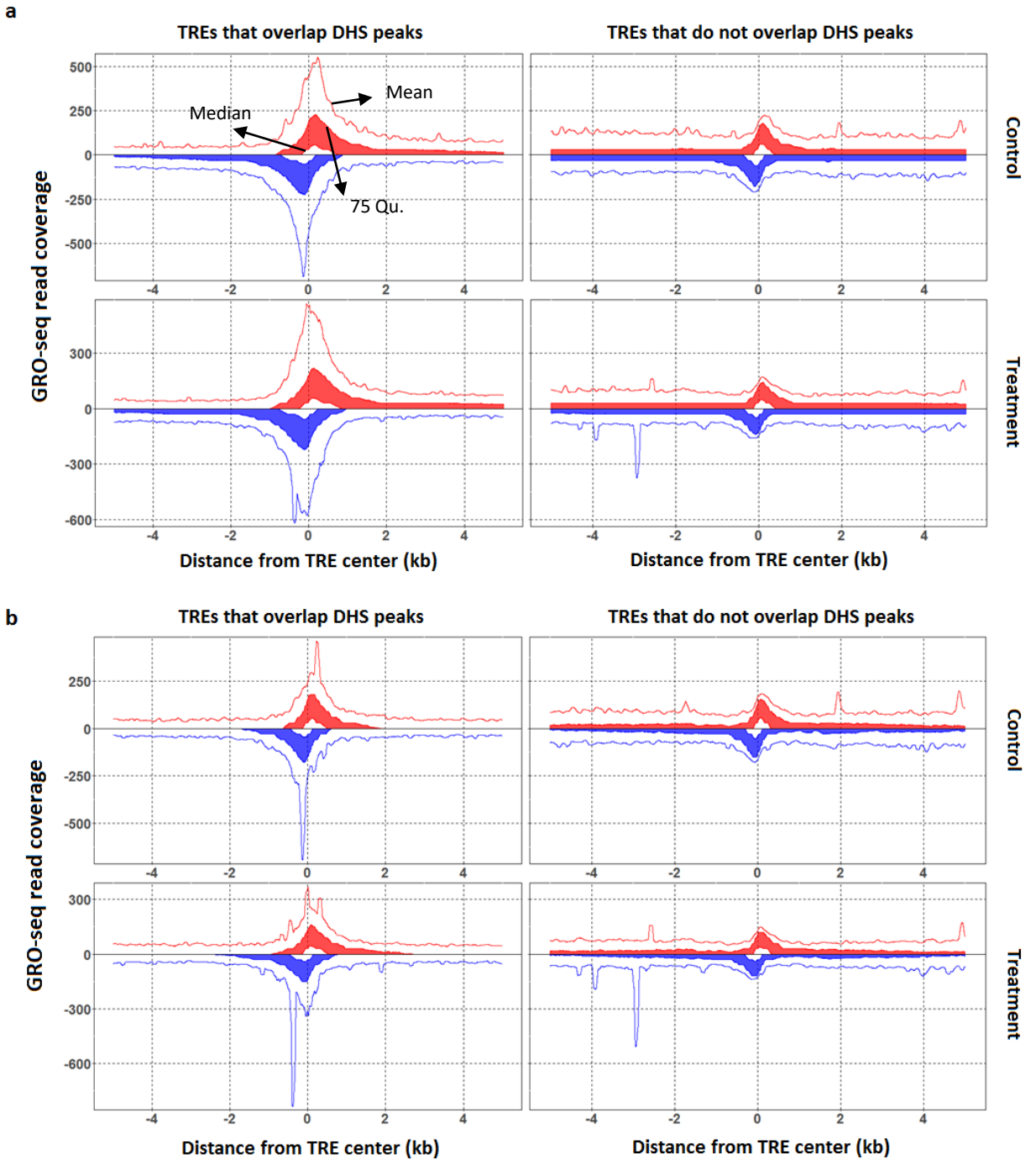
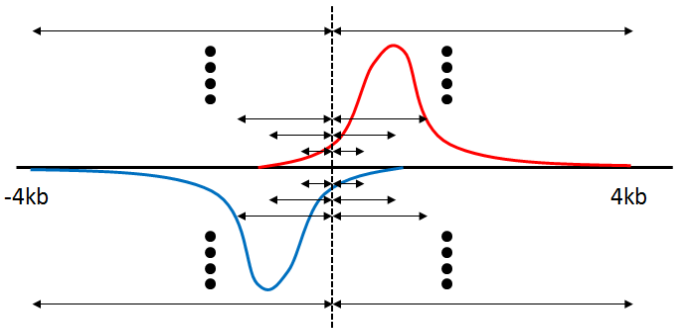


Figure 15. GRO-seq read coverage of TREs. (a) All TREs and (b) Intergenic/Intronic TREs from MCF7 control and treatment samples were divided into those that overlap (left column) or do not overlap (right column) MCF7 DHS peaks. Red/blue denotes the coverage of the forward/reverse strands respectively. The mean, 75 quantile and median are marked on each strand (see top left figure for details). Positive/Negative distances from TRE center denote downstream/upstream of the TRE center, respectively. TRE regions that overlap DHS peaks manifest different coverage and tail behavior compared to TRE regions that do not overlap DHS peaks.

In order to systematically assign a TRE to one of the classes above we took a machine learning approach. First, we defined a set of 41 features for each TRE as follows: (1) the average GRO-Seq coverage in 40 different windows around the TRE center: $(0,400), (0,800), \dots, (0,4000)$ and symmetrically: $(-4000,0), \dots, (-800,0), (-400,0)$ for the positive strand, and the same windows for the negative strand (**Fig. 16a**). We favored shorter overlapping windows towards the center since coverage tends to be high around the TRE's center. In this way the shorter windows will capture the coverage shape along the TREs. (2) TRE width. For the learning process, we used binary labels to partition the TREs into two groups: TREs that overlap some DHS peak (labeled as 1), and TREs that do not overlap any DHS peak (labeled as 0). Using the set of TREs in each sample and their labels we trained a linear SVM classifier. Here we used samples from five cell lines (MCF7, HCT116, IMR90, K562 and LNCaP) that had GRO-Seq and DHS data for control and treatment samples. We used the same DHS data for control and treatment samples. Validation was done using pairs of samples from different cell lines: for each pair of samples we generated a classifier using the TREs of one sample, and tested its performance on TREs from the other sample, and vice versa.

Most of the classifiers obtained AUC≥ 0.8 (**Fig. 16b**). Interestingly, training on most samples produced predictions of AUC> 0.88 on both HCT116 samples. The reason for this particular behavior remains an open question.

a



b

Training \ Test	MCF7.crt	MCF7.trt	HCT116.crt	HCT116.trt	IMR90.crt	IMR90.trt	K562.crt	K562.trt	LNCaP.crt	LNCaP.trt
MCF7.crt	0.82	0.82	0.89	0.88	0.79	0.8	0.86	0.84	0.81	0.8
MCF7.trt	0.82	0.82	0.89	0.88	0.79	0.8	0.86	0.84	0.81	0.8
HCT116.crt	0.8	0.8	0.89	0.88	0.79	0.8	0.85	0.84	0.8	0.79
HCT116.trt	0.8	0.8	0.89	0.89	0.79	0.8	0.84	0.84	0.8	0.79
IMR90.crt	0.81	0.81	0.89	0.88	0.78	0.79	0.85	0.82	0.81	0.8
IMR90.trt	0.8	0.81	0.89	0.88	0.79	0.8	0.85	0.83	0.81	0.81
K562.crt	0.81	0.81	0.89	0.88	0.79	0.79	0.87	0.83	0.81	0.81
K562.trt	0.81	0.82	0.89	0.88	0.79	0.8	0.85	0.84	0.81	0.8
LNCaP.crt	0.81	0.81	0.87	0.86	0.76	0.77	0.85	0.79	0.81	0.8
LNCaP.trt	0.81	0.81	0.87	0.86	0.77	0.78	0.85	0.8	0.81	0.81

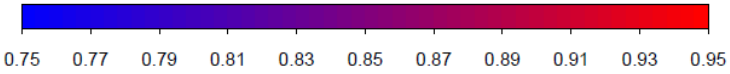


Figure 16. Predicting TREs that overlap with DHS peaks. (a) Windows upstream/downstream from the TRE's center (dashed line) used for computing features. Arrows denote each window span. The average coverage in each direction was used as one feature. (b) Cross-dataset validation – The rows show the training sample and the columns denote the tested sample. Entries are the AUC values for each train-test pair. Using MCF7 control sample as the training set gave the best prediction (mean AUC across predicted samples).

4.5. Enhancer-promoter mapping

In order to map between enhancer and their target genes' promoters we applied our algorithm (see **Section 3.6**) on the enhancer and gene expression matrices. We tested three regression methods: OLS, GLM.NB and ZINB (see **Sections 2.2.4.1, 2.2.4.4-2.2.4.5**), and evaluated the performance of each. Two non-parametric criteria were used to evaluate the results: *binarized expression validation* and *expression level validation* (see **Section 3.6.1**). For each result significance Q-values were computed.

The number of models obtained by each method is summarized in **Supplementary Tables S.4-S.5**. At FDR<0.1 OLS produced 6,323 models, GLM.NB gave 3,642 and ZINB 4,835. **Figure 17a-b** summarizes the performance of each method in terms of the two validation criteria. OLS regression outperformed GLM-NB and ZINB in both validation tests. Since OLS produced more models and performed better in terms of model quality validation, we used the OLS-based models henceforth.

When using OLS, 3,507 genes passed both validation tests at FDR<0.1 (**Fig. 17c**), 2,580 passed only the binary test, 236 passed only the expression level test, and 2,037 genes failed in both tests. Genes in the 'binary only' group had very few samples with positive expression (48 samples on average) compared to 'level only' group (153 samples on average). Thus, in these cases ('binary only'), the power of the Spearman test was very limited. We therefore chose to exclude the genes that failed both tests, and used the remaining 6,323 models for further analysis.

The FANTOM5 study calculated enhancer-gene interactions by constructing a model for each gene using OLS (without cross validation) and reported all models that had $R^2 \geq 0.5$ [12]. For each gene we compared its FANTOM5 R^2 (calculated over all samples, without leave-out CV) to the Spearman correlation value calculated for our OLS model (comparing levels predicted by our model on left-out samples and observed levels in these samples) for that gene (**Fig. 17d**). More than 37% of the genes with $R^2 \geq 0.5$ had Spearman $\rho < 0.5$, suggesting a high FDR in the original FANTOM5 results, probably due to overfitting. Some examples for high and low discrepancy between R^2 and ρ values are shown in **Fig. 18**.

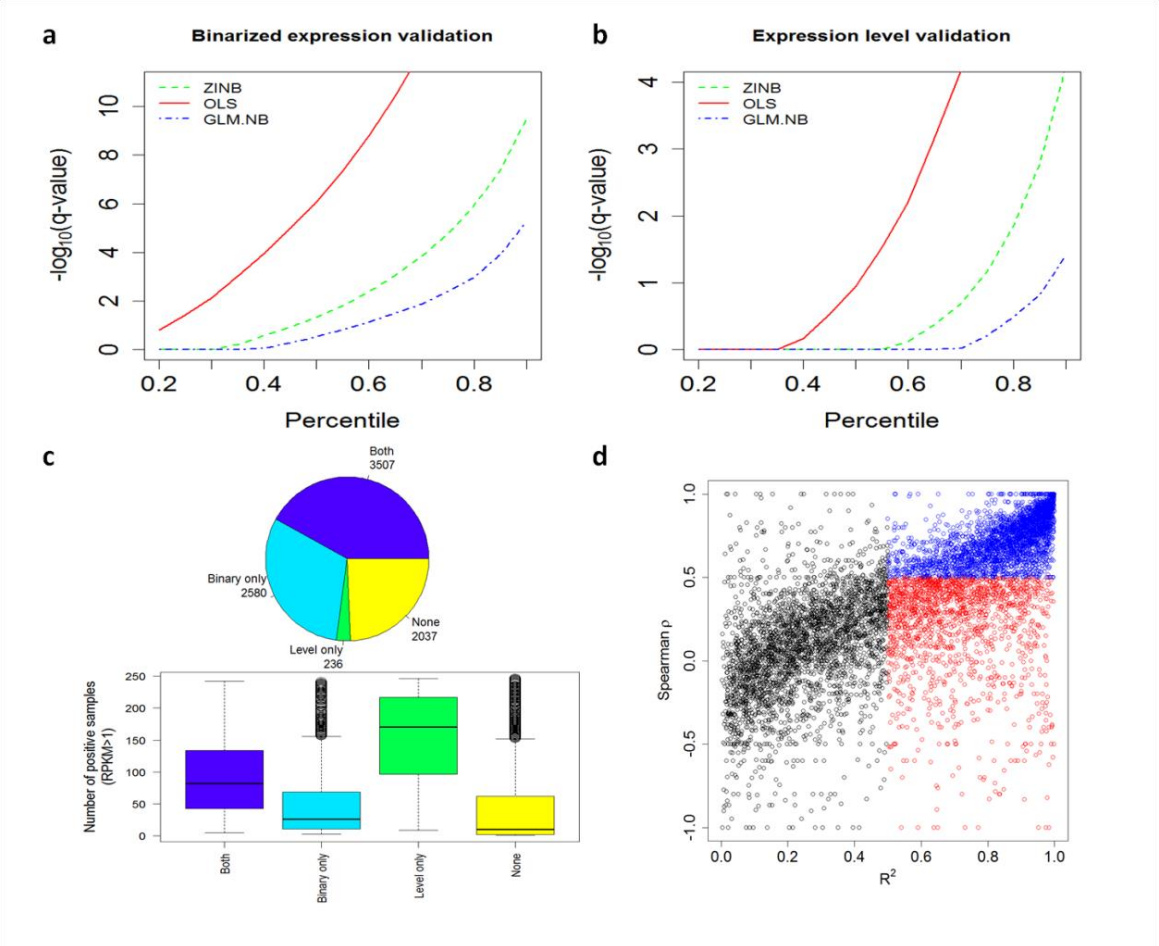


Figure 17. Performance of methods for constructing enhancer-promoter models. (a) Binarized expression validation scores. The x-axis is the percentiles of the $-\log_{10}$[q-values] computed by Wilcoxon rank sum test. (b) Expression level validation scores. The x-axis is the percentiles of the $-\log_{10}$[q-values] computed by the Spearman correlation test. Both plots show advantage of OLS over the other methods. (c) Top: Breakdown of the genes whose OLS models passed each of the validations. Binary/Level only: genes that passed only binary/level validation ($q<0.1$). The number of genes in each category is shown next to each pie slice. Bottom: The distribution of the number of samples that showed positive expression (RPKM≥ 1) for the genes in each category. The 'Level only' and 'Both' categories capture the majority of the genes that had many samples with positive expression. (d) Comparison between Spearman ρ correlation obtained by the OLS models and gene model R^2 values as computed by FANTOM5 (without cross validation). Blue dots: genes with $R^2 \geq 0.5$ and $\rho \geq 0.5$; red dots: genes with $R^2 \geq 0.5$ and $\rho < 0.5$. Gene model selection based on R^2 might produce many over-fitted models (red dots).

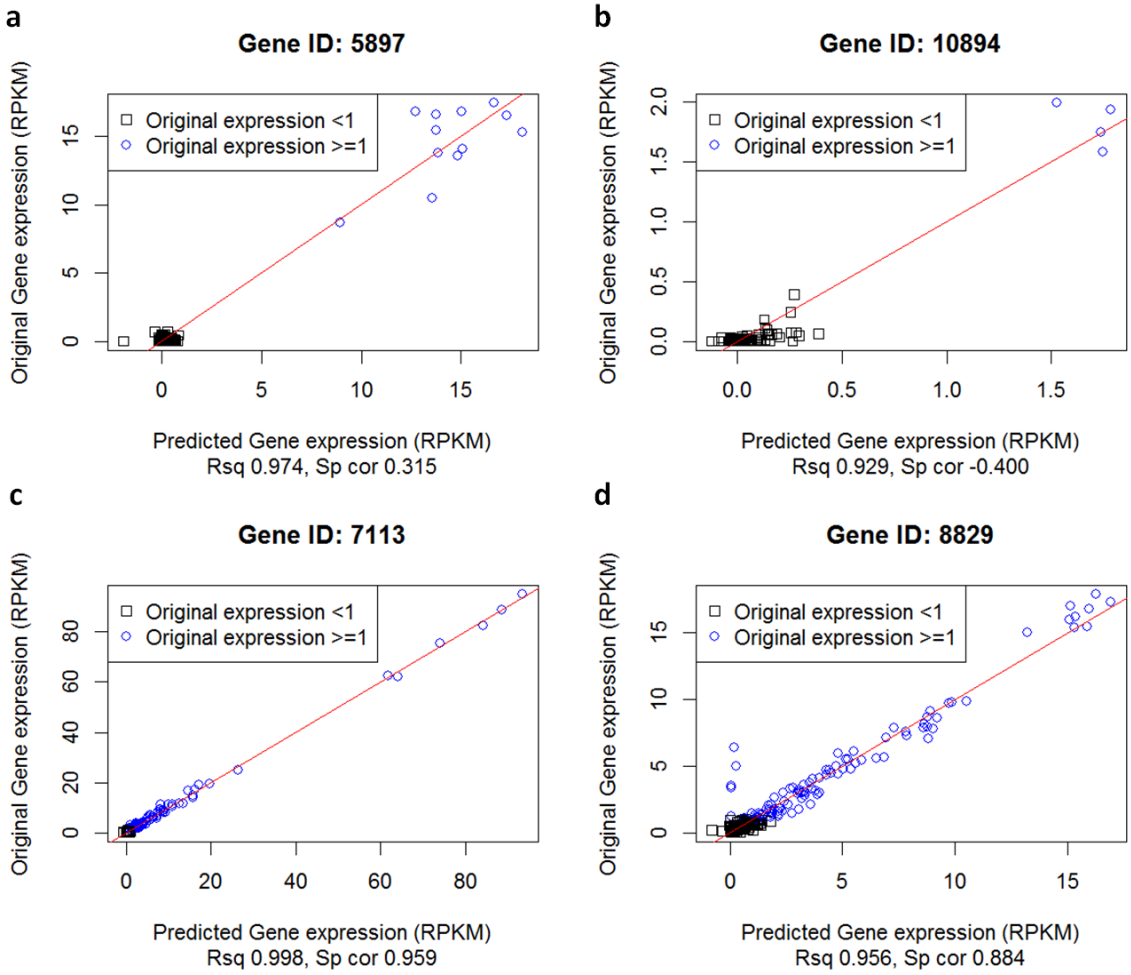


Figure 18. The effect of cross validation. The plots show for each sample its OLS-predicted gene expression value (x-axis) vs. the true gene expression value (y-axis). R^2 (Rsq) of the model, computed on all samples, and Spearman correlation (Sp cor) values, computed only on positive samples (RPKM≥1), are listed below each figure. Blue circles: samples in which the gene's expression was RPKM ≥1. These are the only samples used in the expression level validation step. Black squares: samples with RPKM<1. The binarized expression validation compares to what extent these samples match those samples predicted to have RPKM<1. (a-b) Two gene models that did not pass our two validation tests. These models manifest high R^2 but have low Spearman correlation suggesting overfitted models. (c-d) Two gene models that passed our two validation tests. These models manifest both high R^2 and high Spearman correlation and thus are less likely to overfit the data.

Next, we examined the properties of the full models (based on the 10 closest enhancers for each gene) and of the shrunken models. First, we computed the proportional contribution of each enhancer in the full model: The proportional contribution, as defined in [12], is r^2/R^2 where r is the pairwise Pearson correlation coefficient between the enhancer and the gene expression and R^2 is the coefficient of determination of the full model. The proportional contribution tended to decrease with the enhancer's distance from the gene (**Fig. 19a**). Second, we examined the distribution of R^2 values of the models and observed that ~70% of the models (4,449 out of 6,323) had $R^2 \geq 0.5$ (**Fig. 19b**). Models with $R^2 < 0.5$ passed our 'level only' and 'binary only' tests, based on the non-parametric validation scores. Note that these models were constructed using cross-validation and not trained on all samples as in the FANTOM5 report.

84% of the 3,507 models from the 'Both' group had $R^2 \geq 0.5$, compared to 16.5% of the 236 'Level only' models and to 57% of the 2,580 "Binary only" models. Third, we computed the frequency of enhancer inclusion in the shrunken models as a function of their ranked distance from the gene (**Fig. 19c**). We observed a moderate decrease with rank, where the closest enhancer appeared in ~55% of models, and the 9th in <30% only. Most shrunken models (3,464 out of 6,323) had 1-3 enhancers only (**Fig. 19d**), an average of ~3.6 enhancers were linked to each gene, and an average of ~1.1 genes were mapped to each enhancer.

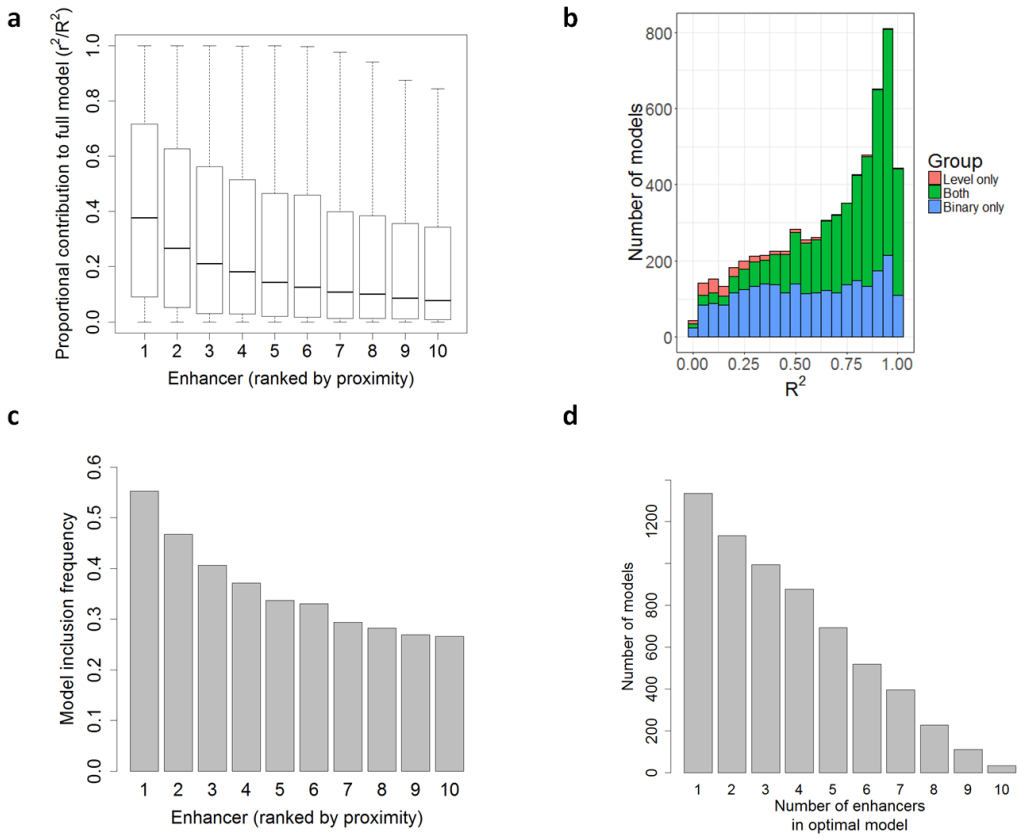


Figure 19. Enhancer contribution to full and shrunken model. In each model the ten most proximal enhancers (within ±500 kb) of the gene's TSS were considered. The full model used all ten, while the shrunken model removed some enhancers based on their scores. (a) The proportional contribution of each enhancer in the full model. Enhancers are ranked by their distance from the TSS where 1 is the closest. (b) R^2 value distribution of the full models (n=6,323, see **Table 2**) by groups (See **Figure 17c** for details). Approximately 70% of 6,323 models had $R^2 \geq 0.5$. (c) Inclusion rate of enhancers in the models after shrinkage. Enhancers are ranked by their distance from the TSS (1 is the closest). (d) Histogram of the number of enhancers included in the shrunken models.

Next, we compared the performance of FOCS with that of three previously used methods for linking enhancer activity to gene activity: (1) *Pairwise* comparison - computing pairwise Pearson correlation (r) between the expression patterns of each enhancer and each gene and correcting for multiple testing using Benjamini Hochberg [84] FDR $\leq 10^{-5}$. (2) *Pairwise+r* $= 0.7$ method – same as (1), but requiring also Pearson coefficient r ≥ 0.7. Methods 1-2 were previously used in the FANTOM5 project [12] and in inferring E-P links based on co-

appearance of DHSs at enhancers and promoters [11]. (3) *OLS-LASSO* - building gene models using OLS regression on the 10 enhancers most proximal to the gene's TSS, selecting gene models with $R^2 \geq 0.5$, and shrinking the selected models using LASSO [97]. Method 3 was also used in FANTOM5 project [12] as an alternative way to infer E-P links. FANTOM5 performed the LASSO step using the cv.glmnet function of the glmnet package [100], with 100-fold cross validations and selected the largest value of lambda such that the mean square error was within one standard error of the minimum. **Table 2** summarizes the number of E-P links obtained by each method. To compare the effectiveness of enet and LASSO, we also constructed *OLS-ENET* models using cv.glmnet function, with the same parameters as for the OLS-LASSO method except that α was chosen to be 0.5 to account for both LASSO and Ridge regularizations. As expected, we can see that the elastic net tends to identify more enhancers per gene (4.7 vs. 3.8 on average).

Table 2. Summary of E-P links inferred using GRO-Seq compendium

Method	#Gene models	#Links to enhancers	#Unique enhancers	Links/model
Pairwise	7,825	113,817	81,040	14.5
Pairwise+$r = 0.7$	4,347	26,827	24,247	6.2
OLS-LASSO*	4,570	17,141	16,121	3.8
OLS-ENET*	4,580	21,379	19,796	4.7
FOCS	6,323	22,607	20,650	3.6

*The initial number of gene models with $R^2 \geq 0.5$ was 4,851. After performing LASSO or enet enhancer selection, some or all enhancers were removed from the models.

We used two external sources to evaluate the performance of models obtained by the three methods: ChIA–PET data and eQTL SNPs. We downloaded ChIA-PET interaction data mediated by RNAPII for the cell lines MCF7, HCT-116, K562 and HelaS3 from the chromatin–chromatin spatial interaction (CCSI) database [116]. eQTL SNPs were taken from the GTEx project [1]. We defined a 1 kbp promoter interval around each gene's TSS (±500 bp upstream/downstream) and 1 kbp enhancer intervals (±500 bp from the enhancer's center) as the candidate regulatory regions.

For each method, we computed the fraction of E-P links that were supported by ChIA-PET interactions between the corresponding intervals, and the fraction of E-P links that were supported by eQTL SNPs in enhancer interval and the affected gene with the corresponding promoter (for more details see **Section 3.7.1**). The results are shown in **Figure 20**. FOCS substantially outperformed the other methods in terms of the fraction of predicted E-P links supported by ChIA-PET or eQTL SNPs data (~57%, 12,864 out of 22,607 E-P links, and ~33%, 7,558 out of 22,607 E-P links, respectively). OLS-LASSO had similar eQTL support but made almost 24% less predictions (17,141 vs. 22,607 E-P links, **Table 2**). The two pairwise methods made 1.2-5 folds more predictions compared to FOCS (**Table 2**) but had support of only 22-32% by the external sources data, indicating much higher rate of false positives.

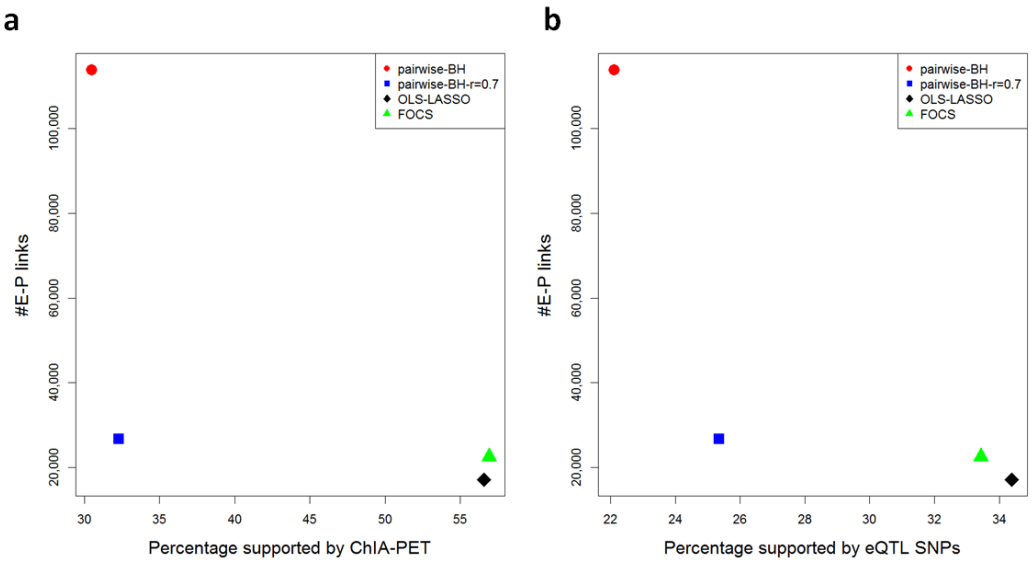


Figure 20. The performance of different E-P predictors evaluated using external sources. (a) ChIA-PET. (b) GTEx eQTL SNPs. The y-axis shows the total number of E-P links predicted and the x-axis denotes the percentage of predicted links that had support by ChIA-PET or eQTL SNPs. Our method (FOCS, green triangle) manifested high fraction of E-P links supported by ChIA-PET (~57%) and by eQTL SNPs (~33%). Empirical P-values for the significance of overlap between the links provided by each external source and the predicted ones were 0 for all methods, in both ChIA-PET and eQTLs (for more details see **Section 3.7.1**).

4.6. Downstream analysis of E-P links – Proof of concept

In this section we describe, as a proof of concept, a possible way of using the predicted E-P links. Further research is required to understand how to utilize E-P links for downstream analysis.

4.6.1. The expression data

We used GRO-Seq data of MCF7 cell line from [63]. The data consists of 8 MCF7 GRO-seq samples after treatment with 17 − β estradiol (E2) for 0,10,40,160 minutes (two replicates for each time). E2 treatment activates the estrogen receptor (ER) TF [122]. We took the relevant gene and enhancer expression profiles (i.e., the columns in the matrix that correspond to the relevant samples) from our combined matrices G^n and E^n. We applied the preprocessing and clustering as described in **Section 3.7.2.1**. The gene expression matrix contained 4,963 genes that showed 2- fold change in expression between at least two of the eight samples. Application of Click clustering [90] produced 21 gene clusters (median of 133 genes per cluster).

4.6.2. Downstream enrichment analysis results

For each cluster we performed GO enrichment analysis, and de novo motif finding on the promoters and on the enhancers linked to the cluster's genes, as described in **Section 3.7.2.2**. We selected significant GO enrichments that had an empirical P-value ≤ 0.05 and selected significant motif finding results that had hyper-geometric P-value $\leq 1e-15$. In total, ~10% of the gene clusters had significant GO enrichment, ~81% had motif enrichment in promoters, and ~81% of the enhancer clusters had significant motif enrichment.

Table 3 shows the GO enrichments found in two gene clusters. Cluster 2 had 782 genes up-regulated following 40m E2 treatment, including 650 genes linked to 2,661 enhancers. Cluster 4 contained 569 genes up-regulated following 160m E2 treatment, including 437 genes linked to 1,417 enhancers. Interestingly, the 160m cluster had 27 enriched GO terms while the 40m cluster had only a single enriched GO term. The 10-40m gene clusters contained genes that are mainly the primary target of the activated enhancers, which have influence in early responses to stimulus [123]. These results are in agreement with the results in [63] claiming that the lack of enriched terms in short time stimulus may be due to a switch from one cellular signaling program (e.g., serum response) to another (e.g., estrogen signaling); each pathway may require the same functional categories (e.g., system development) but use a distinct set of genes within each category.

Tables 4 and 5 show the motif finding results for Cluster 2 on promoters and enhancers, respectively. Interestingly, motifs found in enhancers linked to the promoters tended to have higher significance compared to those found in promoters. This may suggest that enhancer activity takes part in early responses prior to gene activation.

Tables 6 and 7 show the motif finding results for Cluster 4 on promoters and enhancers, respectively. The motifs found in promoters in this cluster had higher significance compared to cluster 2 promoter motifs. Enhancer motifs were less significant compared to Cluster 2 enhancer motifs. This also supports the hypothesis that enhancers are activated in early responses, possibly by the same TFs, and are likely to target primary genes.

In summary, we demonstrate that utilizing E-P links in downstream analysis can broaden and improve recent known annotations using functional enrichment and help discover new motifs in enhancers suggesting novel candidate TFs that regulate primary target gene expression in early stimulus. Much further work is still needed to better utilize the E-P links for downstream analysis.

Table 3. TANGO GO enrichment				
Cluster id	**GO term**	**#genes**	**Raw p-value**	**Empirical P-value***
2	system development - GO:0048731	210	3.2e-6	0.044
4	cotranslational protein targeting to membrane - GO:0006613	20	3.3e-18	0.001
4	viral transcription - GO:0019083	16	1.6e-14	0.001
4	translational termination - GO:0006415	16	1.3e-13	0.001
4	mRNA metabolic process - GO:0016071	28	1.2e-12	0.001
4	ncRNA metabolic process - GO:0034660	22	7.3e-12	0.001
4	translation - GO:0006412	41	2.7e-11	0.001
4	structural constituent of ribosome - GO:0003735	17	4.9e-11	0.001
4	ncRNA processing - GO:0034470	18	5.3e-11	0.001
4	RNA processing - GO:0006396	25	1.7e-10	0.001
4	viral infectious cycle - GO:0019058	16	2.9e-10	0.001
4	protein metabolic process - GO:0019538	136	4.8e-10	0.001
4	ribonucleoprotein complex biogenesis - GO:0022613	14	5.1e-10	0.001
4	mitotic cell cycle - GO:0000278	34	2.7e-9	0.001
4	cellular macromolecule metabolic process - GO:0044260	200	3.7e-9	0.001
4	protein targeting - GO:0006605	24	9.3e-9	0.002
4	cell cycle phase - GO:0022403	36	1.4e-8	0.002
4	macromolecule catabolic process - GO:0009057	36	1.8e-8	0.002
4	rRNA metabolic process - GO:0016072	10	2.4e-8	0.002
4	protein complex subunit organization - GO:0071822	44	4.8e-8	0.002
4	DNA metabolic process - GO:0006259	35	7.5e-8	0.002
4	mitosis - GO:0007067	19	1.6e-7	0.003
4	intracellular protein transport - GO:0006886	29	2.0e-7	0.004
4	cellular protein localization - GO:0034613	40	4.1e-7	0.007
4	cellular macromolecule localization - GO:0070727	40	4.1e-7	0.007
4	M phase - GO:0000279	22	6.3e-7	0.011
4	cell cycle - GO:0007049	46	1.9e-6	0.029
4	nitrogen compound metabolic process - GO:0006807	155	2.1e-6	0.031
4	establishment of protein localization - GO:0045184	43	2.2e-6	0.032
4	protein transport - GO:0015031	42	3.0e-6	0.042
* Empirical P-values computed using random permutation test				

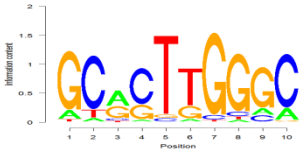
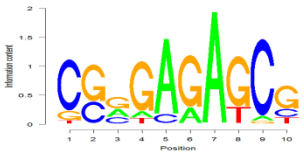
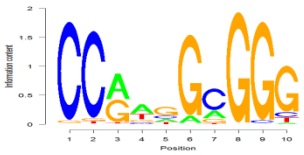
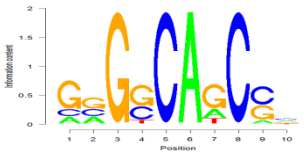
Table 4. Cluster 2 40m motif enrichment on gene promoters (top 4 motifs)		
Motif logo	**TFs**	**p-value**
		1.2e-19
		3.2e-19
	E2F6, E2F4, E2F1	1.5e-18
	TAL1::GATA1	4.1e-18

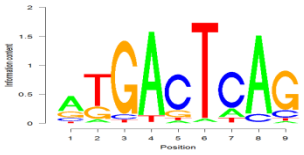
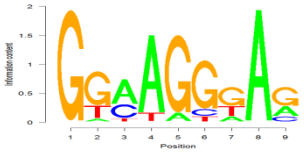
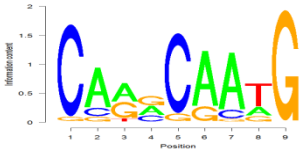
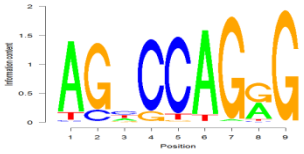
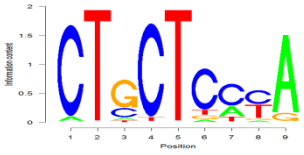
Table 5. Cluster 2 40m motif enrichment on linked enhancers (top 5 motifs)		
Motif logo	**TFs**	**p-value**
	Bach1::Mafk,NFE2::MAF,Nfe2l2, BATF::JUN,NFE2L1::MafG,FOSL2, JUNB,JUN,FOS,JUN::FOS, FOSL1,JUND,Pax2	6.7e-34
	EHF,SPI1,EWSR1-FLI1,ELF1,GABPA	3.9e-26
	SOX10, Sox2	6.1e-23
	HNF4A	1.7e-21
	ZNF263	2.8e-20

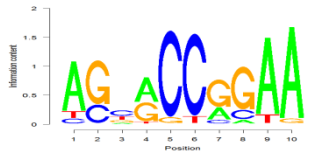
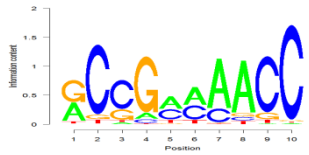
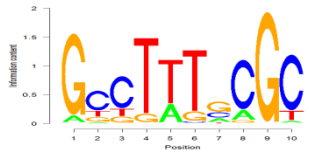
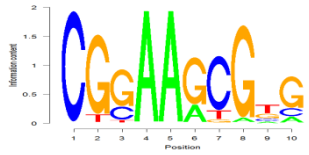
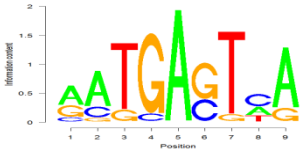
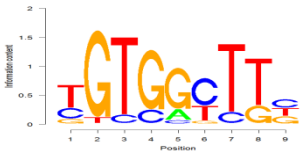
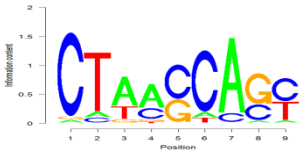
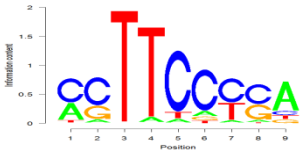
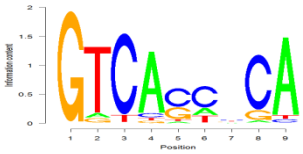
Table 6. Cluster 4 160m motif enrichment on gene promoters (top 4 motifs)		
Motif logo	**TFs**	**p-value**
	ELK4,GABPA,SPI1,ELK1,FLI1	1.5e-22
		1.8e-19
		4.5e-19
	ELK4,SPI1,SPIB,GABPA	4.4e-18

Table 7. Cluster 4 160m motif enrichment on linked enhancers (top 5 motifs)		
Motif logo	**TFs**	**p-value**
	BATF::JUN,FOS,JUNB, FOSL2,JUN,JUND,JUN::FOS, FOSL1,Nfe2l2,NFE2::MAF, NFE2L1::MafG,,Bach1::Mafk	3.5e-25
	Klf1,RUNX1,Mycn, RUNX2,USF2,Myc,USF1,MAX	2.5e-23
		1.0e-21
	SPI1,SPIB,EHF,Stat6, MZF1_1-4,STAT2::STAT1	4.1e-21
	PPARG::RXRA,JUNB, FOSL2,SREBF2,SREBF1, Pax2,Bach1::Mafk, SMAD2::SMAD3::SMAD4, ESR2,Meis1,CREB1	5.9e-20

5. Discussion

In this thesis we developed a new statistical method called FOCS for E-P network inference based on enhancer and gene expression data across the same samples. Our goal was to improve E-P mapping compared to extant methods, notably pairwise correlation and OLS regression followed by LASSO shrinkage.

First, we constructed a large compendium of 246 GRO-seq samples from 40 different studies and covering 23 different cell types, each assayed under control and stressed conditions. We developed a unified preprocessing protocol and applied it on each GRO-seq dataset. The protocol outputs enhancer and gene expression levels for each sample. Second, we applied FOCS on our compendium and predicted a total of 22,607 high confidence cross-validated E-P links. Third, we showed as a proof of concept how these predicted E-P links can be used in downstream analysis.

Enhancer regions in each sample were called using the dREG tool. We compared the performance of dREG and groHMM in calling enhancer regions. We found that dREG outperformed groHMM, both in quality and quantity of identified enhancers, by testing each prediction against histone modification signals on DNase-seq open chromatin regions overlapping the putative enhancers.

We developed a two-steps algorithm to predict E-P links. In the first step, we developed two non-parametric validation tests in order to compare between three types of regression methods - OLS, GLM.NB and ZINB, and choose confident gene models. The non-parametric approach allows us to compare between regression methods without relying on the assumptions of the regression method's distribution. In the second step, we applied elastic-net enhancer selection to shrink the gene models in order to prevent models that are over-fitted to the training set. We made sure that gene models that survived the first step had at least one linked enhancer after the second selection step.

Our testing showed that the OLS regression method had better performance over GLM.NB and ZINB. GLM.NB method suffers from zero-inflation, which makes GLM.NB highly inaccurate compared to ZINB method (**Figure 17.a-b**). Further tests should be done to address why the OLS method achieved better results compared to ZINB. Based on these results we proposed the FOCS algorithm (FDR-corrected OLS with Cross-validation and Shrinkage).

We assessed the E-P prediction performance of FOCS versus previous methods: (1) Pearson pairwise correlation between expression patterns of enhancer and promoter (or gene), and (2) gene model construction using OLS regression followed by LASSO enhancer selection. We used ChIA-PET DNA-DNA interactions (mediated by POL2) and GTEx eQTLs as external sources to support functional interactions between enhancers and promoters. FOCS manifested higher percentages of E-P links supported by ChIA-PET and eQTLs compared to the other methods.

The advantage of FOCS over pairwise correlation can be explained by the richer model that takes into account all possible enhancers together when considering each gene. Pairwise correlation tends to capture more correct E-P pairs, but even after multiple testing corrections, the fraction of correct predictions in the pairwise correlation method is lower (**Figure 20**).

The two non-parametric evaluation measures used for selecting gene models in the first step of FOCS, instead of R^2, have an advantage when the relationship between the gene expression pattern and the expression of its closest enhancers is not linear. These measures are still able to select gene models with linear relationship between gene and enhancer expression patterns. Indeed, as shown in **Figure 19.b**, most of selected models (70% of 6,323 models) had $R^2 \geq 0.5$. A biological explanation for this high fraction may be the large fraction of intronic E-P links: 58% of the 22,607 links were between a gene and an enhancer contained in one of its introns. Such intronic enhancers are likely to have expression pattern similar to their target genes, resulting with a linear relationship. When we tested the effect of disallowing intronic enhancers, the number of models and E-P links dropped dramatically: from 6,323 models and 22,607 links to 5032 models and 12,617 links. However, the 30% of our models would have been missed by a criterion of $R^2 \geq 0.5$, as used in the FANTOM5 analysis.

The OLS method significantly outperformed ZINB and GLM.NB, to a large extent due to allowing intronic enhancers linked to genes containing them (**Figure 17.a-b**). Without this option the OLS method had only a modest advantage over ZINB and GLM.NB (**Supplementary Figure S.4.a-b**). Interestingly, while the distribution of the number of enhancers per model shows a linear drop when including intronic E-P links (**Figure 19.d**) it drops down much more rapidly in models without intronic E-P links (**Supplementary Figure S.5.d**). In comparing the two strategies in validations with external data, allowing intronic links fared consistently better: when allowing intronic links, 57% of the links had ChIA-PET data support, compared to 46% disallowing them. The respective numbers for eQTL validations were 33.5% and 28% respectively (**Figure 20**, data for FOCS disallowing intronic links not shown).

As proof of concept, we showed how the predicted E-P links can be used in downstream analysis. Motif finding analysis performed on promoters of up-regulated genes' clusters in experiments taken shortly after treatment (up to 40 minutes) did not show significant motifs, while motif analysis on the genes' linked enhancers found many significant motifs. In contrast, the same analysis on longer time experiments (160 minutes) resulted with less significant motifs compared to enhancers in short time experiments. This may suggest that the response at early time points is mediated by few key TFs while at later time points multiple additional TFs participate in the response making the motif signal in enhancers more diluted and difficult to detect, and if so, such analysis could help in identifying the primary target genes rather than the secondary target genes. In the future, we plan to develop novel methods for utilizing E-P links in downstream analysis. Several other research directions are discussed below.

Motif-finding tools are currently limited to a search area of short promoters (<3kb) or to short conserved regions spanning up to 20kb upstream of the gene's TSS and may miss crucial motifs in enhancers located distal (>40kb) from their target genes. Improving current motif-finding tools to include also the analysis of the genes' linked enhancer regions can shed additional light on unknown regulatory networks that control various biological processes.

FOCS can be extended and improved in several ways. It can be utilized by integrating heterogeneous biological data, including chromatin conformation (e.g., ChIA-PET interactions) and epigenomic data (e.g., data recorded by the ENCODE and Epigenome Road map consortia). Such integration can help to eliminate false positive enhancers (as done in **Section 4.4**) and reduce false E-P links that are not supported by 3D interactions. Integrative analysis that combines these

diverse data resources has the potential to considerably enhance the performance of our methodology for E-P mapping.

Another promising application of FOCS is in noncoding SNP analysis. The vast majority of genetic variants associated with complex traits and diseases map to noncoding genomic regions, and a fraction of them presumably acts by modulating the activity of enhancer elements. Single nucleotide polymorphisms (SNPs) are the most well-studied class of such genetic variants. Using our E-P mapping and data from genome-wide association studies (GWAS), we aim to find SNPs that reside in enhancers and modulate gene activity. An example of such analysis of GWAS data was done by the GTEx project, which associated SNP genotype variation data with gene expression levels across multiple tissues. A possible way to increase the power of current associations of SNPs to genes and to discover new SNP-gene associations is by using E-P links. E-P links intersected with SNPs positions, either in the enhancer or in the gene regions, can indicate the mechanism of action of such risk SNPs.

6. References

1. Consortium Gte, others. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-.). 2015;348:648–60.

2. 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. Proc. Natl. Acad. Sci. 2005;102:15545–50.

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

4. Pathan M, Keerthikumar S, Ang C-S, Gangoda L, Quek CYJ, Williamson NA, et al. FunRich: An open access standalone functional enrichment and interaction network analysis tool. Proteomics. 2015;15:2597–601.

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

6. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat. Methods. 2012;9:796–804.

7. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. Adaptive Computation and Machine Learning. MIT Press, Cambridge; 2000.

8. Yamanishi Y, Vert J-P, Kanehisa M. Protein network inference from multiple genomic data: a supervised approach. Bioinformatics. 2004;20:i363--i370.

9. Huang L, Liao L, Wu CH. Inference of protein-protein interaction networks from multiple heterogeneous data. EURASIP J. Bioinforma. Syst. Biol. 2016;2016:1–9.

10. Hecker M, Lambeck S, Toepfer S, Van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models—a review. Biosystems. 2009;96:86–103.

11. 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.

12. 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.

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

14. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. [Internet]. Nature Publishing Group; 2014 [cited 2014 Jul 9];15:272–86. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24614317

15. 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.

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

17. Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. Cell. 2012;150:12–27.

18. Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell. 2008;134:25–36.

19. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009;461:199–205.

20. Olins AL, Olins DE. Spheroid chromatin units ($ν$ bodies). Science (80-.). 1974;183:330–2.

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

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

23. 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.

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

25. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. Science (80-.). American Association for the Advancement of Science; 1998;281:60–3.

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

27. Banerji J, Olson L, Schaffner W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell. 1983;33:729–40.

28. Hartenstein V, Jan YN. Studying Drosophila embryogenesis with P-lacZ enhancer trap lines. Roux's Arch. Dev. Biol. 1992;201:194–220.

29. 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.

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

31. 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. Proc. Natl. Acad. Sci. 2002;99:757–62.

32. 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. Nature Research; 2015;12:433–8.

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

34. 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.

35. Consortium EP, others. An integrated encyclopedia of DNA elements in the human genome. Nature. Nature Publishing Group; 2012;489:57–74.

36. Perry MW, Boettiger AN, Bothma JP, Levine M. Shadow enhancers foster robustness of

Drosophila gastrulation. Curr. Biol. 2010;20:1562–7.

37. Ron G, Moran D, Kaplan T. Promoter-Enhancer Interactions Identified from Hi-C Data using Probabilistic Models and Hierarchical Topological Domains. bioRxiv. 2017;101220.

38. 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.

39. 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.

40. Moore DS. The developing genome: An introduction to behavioral epigenetics. 2015.

41. Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. Genes Dev. 2009;23:781–3.

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

43. Kovalchuk O, Baulch JE. Epigenetic changes and nontargeted radiation effects—is there a link? Environ. Mol. Mutagen. 2008;49:16–25.

44. Ilnytskyy Y, Kovalchuk O. Non-targeted radiation effects—an epigenetic connection. Mutat. Res. Mol. Mech. Mutagen. 2011;714:113–25.

45. Friedl AA, Mazurek B, Seiler DM. Radiation-induced alterations in histone modification patterns and their potential impact on short-term radiation effects. Front. Oncol. 2012;2:117.

46. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal R, et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 2014;24:1–13.

47. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 1977;74:5463–7.

48. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science (80-.). 2007;316:1497–502.

49. 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.

50. 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.

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

52. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 2006;16:123–31.

53. 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.

54. Gusmao EG, Dieterich C, Zenke M, Costa IG. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. Bioinformatics.

2014;30:3143–51.

55. Boyle AP, Song L, Lee B-K, London D, Keefe D, Birney E, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011;21:456–64.

56. 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.

57. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011;21:447–55.

58. Cuellar-Partida G, Buske FA, McLeay RC, Whitington T, Noble WS, Bailey TL. Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics. 2012;28:56–62.

59. 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.

60. Fullwood MJ, Ruan Y. ChIP-based methods for the identification of long-range chromatin interactions. J. Cell. Biochem. 2009;107:30–9.

61. Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol. 2010;11:R22.

62. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science (80-.). 2008;322:1845–8.

63. Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, Lis JT, et al. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell. 2011;145:622–34.

64. Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature. 2011;474:390–4.

65. 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.

66. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, et al. Divergent transcription from active promoters. Science (80-.). 2008;322:1849–51.

67. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science (80-.). 2007;316:1484–8.

68. 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;16:222.

69. 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.

70. Kruesi WS, Core LJ, Waters CT, Lis JT, Meyer BJ. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. Elife. 2013;2:e00808.

71. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science (80-.). 2013;339:950–3.

72. Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat. Protoc. 2016;11:1455–76.

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

74. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2015;gkv1189.

75. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.

76. Hah N, Murakami S, Nagari A, Danko CG, Kraus WL. Enhancer transcripts mark active estrogen receptor binding sites. Genome Res. Cold Spring Harbor Lab; 2013;

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

78. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. 1970;41:164–71.

79. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inf. Theory. 1967;13:260–9.

80. Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, et al. Identification of active transcriptional regulatory elements from GRO-seq data. 2015;12.

81. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V, others. Support vector regression machines. Adv. Neural Inf. Process. Syst. 1997;9:155–61.

82. Consortium GO, others. Gene ontology consortium: going forward. Nucleic Acids Res. 2015;43:D1049--D1056.

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

84. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B. JSTOR; 1995;289–300.

85. Ulitsky I, Maron-Katz A, Shavit S, Sagir D, Linhart C, Elkon R, et al. Expander: from expression microarrays to networks and functions. Nat. Protoc. Nature Publishing Group; 2010;5:303–22.

86. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, et al. EXPANDER--an integrative program suite for microarray data analysis. BMC Bioinformatics. BioMed Central Ltd; 2005;6:232.

87. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 1998;95:14863–8.

88. Sokal RR. A statistical method for evaluating systematic relationships. Univ Kans Sci Bull. 1958;38:1409–38.

89. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 1987;4:406–25.

90. Sharan R, Shamir R. CLICK: a clustering algorithm with applications to gene expression analysis.

Proc Int Conf Intell Syst Mol Biol. 2000. p. 16.

91. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics. Oxford Univ Press; 2002;18:S136--S144.

92. Shabalin AA, Weigman VJ, Perou CM, Nobel AB. Finding large average submatrices in high dimensional data. Ann. Appl. Stat. JSTOR; 2009;985–1012.

93. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 1998;95:14863–8.

94. Sokal RR. A statistical method for evaluating systematic relationships. Univ Kans Sci Bull. 1958;38:1409–38.

95. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 1987;4:406–25.

96. Hayashi F. Econometrics. 2000.

97. Tibshirani R. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B. JSTOR; 1996;267–88.

98. Tikhonov, A. N., and Arsenin VY. Solutions of ill-posed problems. Vol. 14. Washington, DC: VH Winston and Sons; 1977.

99. Zou H, Hastie T. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Statistical Methodol. 2005;67:301–20.

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

101. Nelder JA, Wedderburn RWM. Generalized Linear Models. J. R. statisitcal Soc. 1972;135:370–84.

102. Lawless JF. Negative binomial and mixed Poisson regression. Can. J. Statisitcs. 1987;15:209–25.

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

104. Cameron AC, Trivedi PK. Regression analysis of count data. second. Cambridge Univ. Press. Cambridge University Press; 2013.

105. Lambert D. Zero-Inflated Poisson Regression , With an Application to Defects in Manufacturing. Taylor Fr. Ltd. behalf Am. Stat. Assoc. Am. Soc. Qual. 1992;34:1–14.

106. Fletcher R. Practical methods of optimization. 2nd ed. New York: John Wiley & Sons; 1987.

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

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

109. Stewart FJ, Ottesen EA, Delong EF. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. ISME J. [Internet]. Nature Publishing Group; 2010;4:896–907. Available from: http://dx.doi.org/10.1038/ismej.2010.18

110. He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, et al. Validation of two ribosomal RNA

removal methods for microbial metatranscriptomics. 2010;7.

111. Tripp HJ, Hewson I, Boyarsky S, Stuart JM, Zehr JP. Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. Nucleic Acids Res. Oxford Univ Press; 2011;gkr576.

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

113. 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 [Internet]. BMC Bioinformatics; 2015;9–11. Available from: http://dx.doi.org/10.1186/s12859-015-0656-3

114. Aboyoun P, Carlson M, Lawrence M, Huber W, Gentleman R, Morgan MT, et al. Software for Computing and Annotating Genomic Ranges. 2013;9:1–10.

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

116. Xie X, Ma W, Songyang Z, Luo Z, Huang J, Dai Z, et al. CCSI: a database providing chromatin--chromatin spatial interaction information. Database. Oxford University Press; 2016;2016:bav124.

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

118. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19:185–93.

119. Mathelier A, Fornes O, Arenillas DJ, Chen C, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016;44:D110--D115.

120. Sharan R, Maron-Katz A, Shamir R. CLICK and EXPANDER: a system for clustering and visualizing gene expression data. Bioinformatics. Oxford Univ Press; 2003;19:1787–99.

121. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.

122. Koritnik DR, Koshy A, Hoversland RC. 17-β-estradiol treatment increases the levels of estrogen receptor and its mRNA in male rat liver. Steroids. 1995;60:519–29.

123. Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science (80-.). 2015;347:1010–4.

7. Supplementary Figures

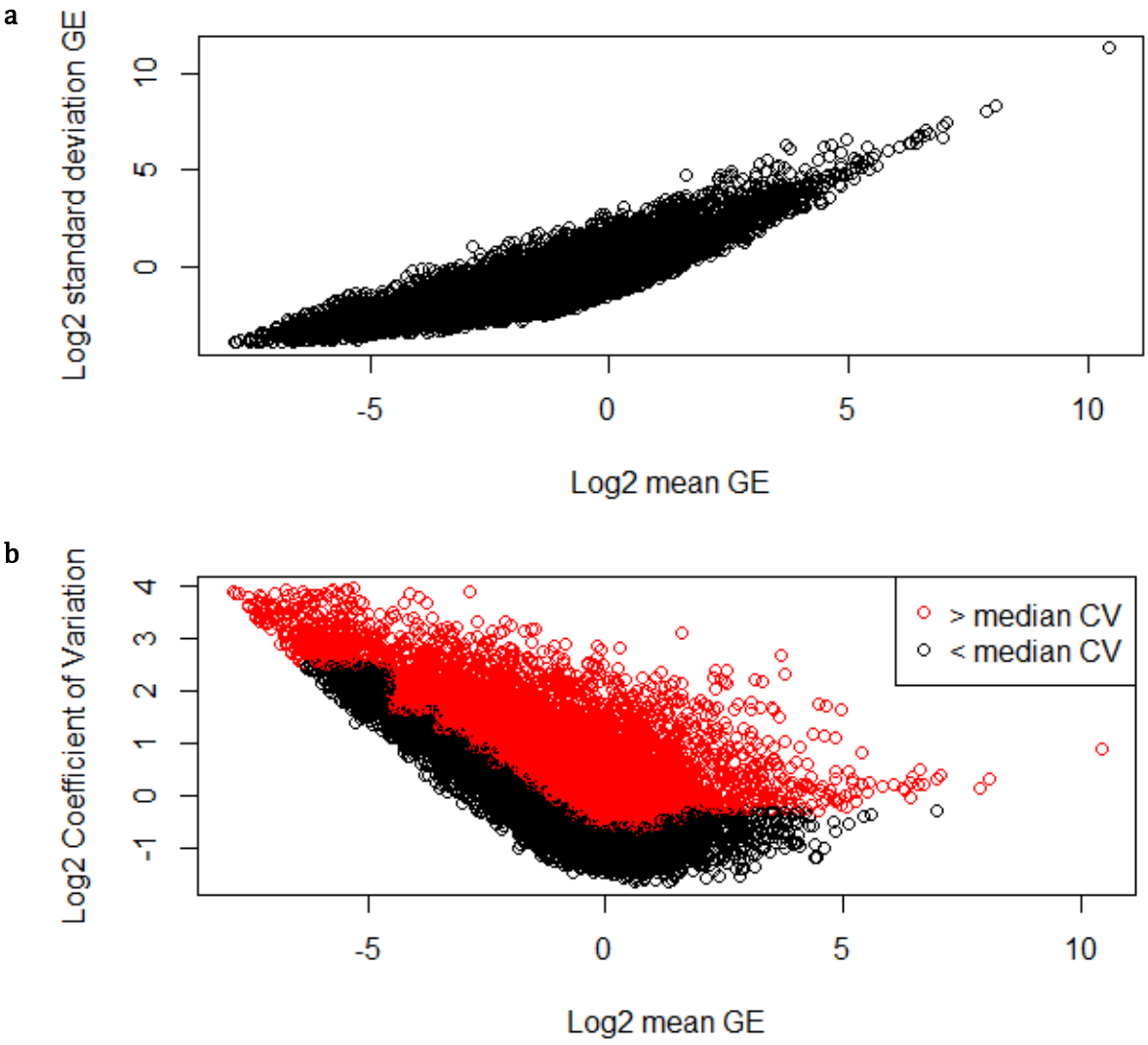


Figure S.1. Filtering analysis of expressed genes (RPKM>1). (a) Genes' Log2 values of Mean RPKM expression (x-axis) and the standard deviation (SD, y-axis) across all samples. There is a positive linear correlation between the mean and the SD values suggesting that filtering genes by variance may give preference to highly or lowly expressed genes. (b) Genes' Log2 values of Mean RPKM expression (x-axis) and the CV (y-axis) across all samples. We partitioned the genes into 20 bins according to their mean RPKM values. We computed the median CV in each bin set of genes. Red/black circles are genes that had CV above/below the median CV in their bin. The plot manifests that by taking only the red genes and filtering the black, most highly expressed variable genes are preserved. In addition, this process filters out lowly expressed genes that are highly variable, possibly due to noise, and genes that do not vary across samples.

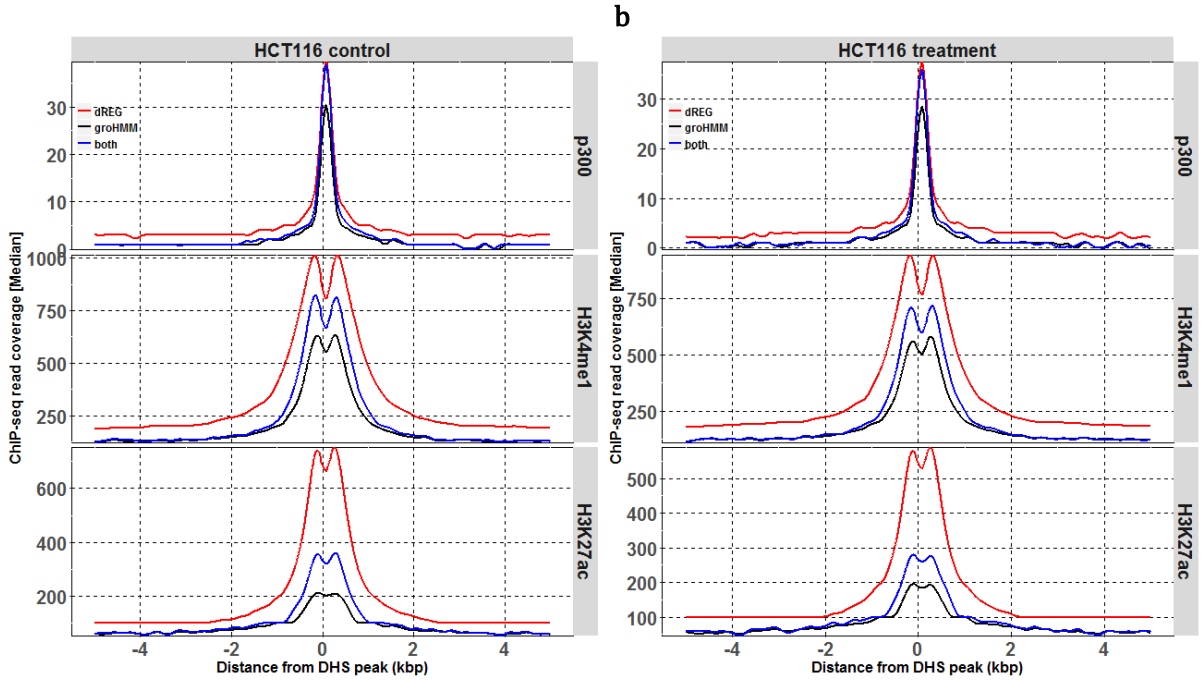


Figure S.2. Epigenetic marks: ChIP-seq median read coverage across DHS peaks. a/b HCT116 control/treatment plots. Red, black, and blue curves show median read coverage for enhancers predicted by dREG, groHMM, and both dREG and groHMM, respectively, that overlap DHS peaks. Rows correspond to the epigenetic markers p300, H3K4me1 and H3K27ac. Negative/Positive distances from DHS center denote upstream/downstream distances respectively. The results show that the regions detected by dREG manifest much higher ChIP-seq signals than the regions detected by groHMM.

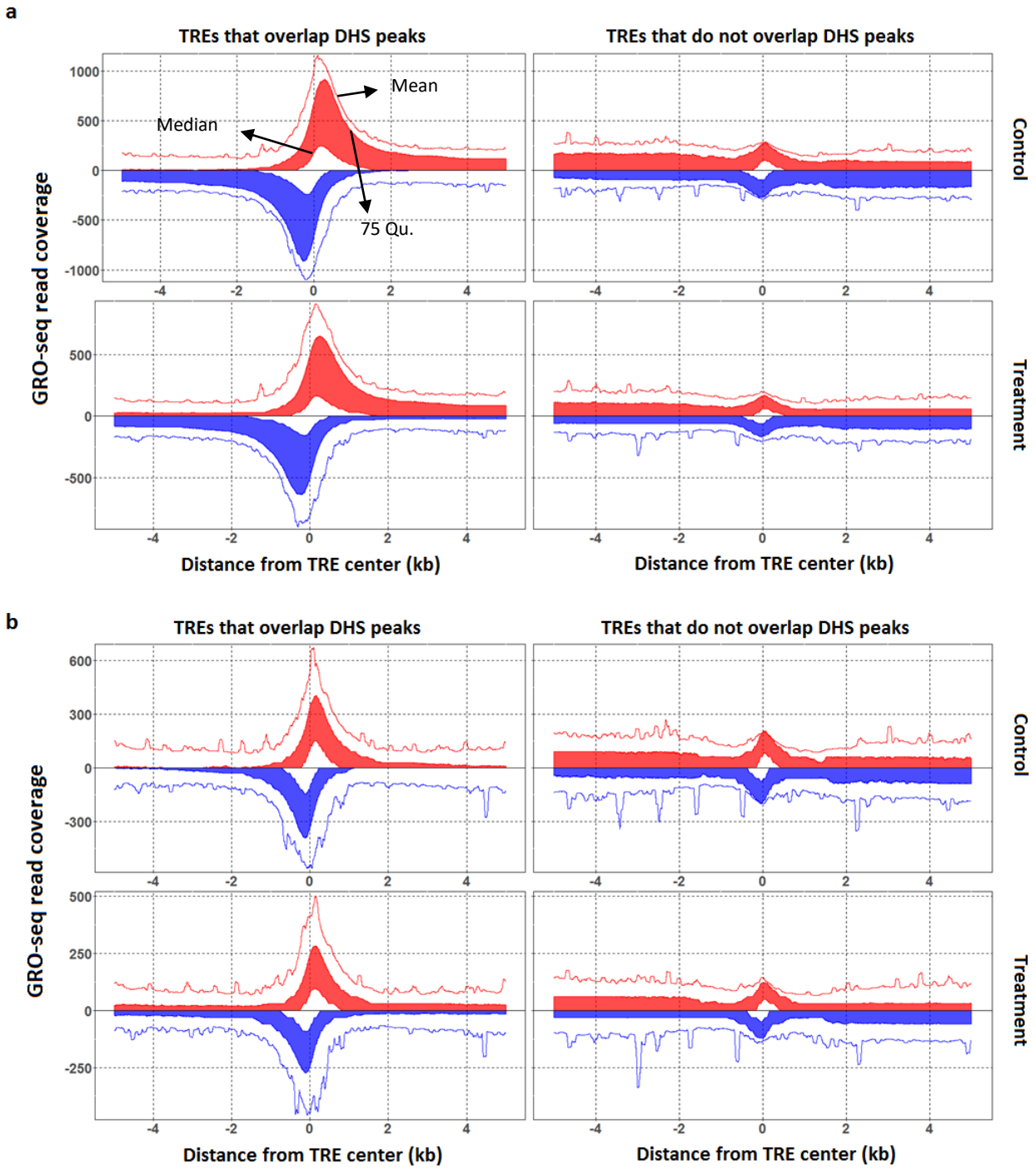


Figure S.3. GRO-seq read coverage of TREs. (a) All TREs and (b) Intergenic/Intronic TREs from HCT116 control and treatment samples were divided into those that overlap (left column) or do not overlap (right column) HCT116 DHS peaks. Red/blue denotes the coverage of the forward/reverse strands respectively. The mean, 75 quantile and median are marked on each strand (see top left figure for details). Positive/Negative distances from TRE center denote downstream/upstream of the TRE center, respectively. TRE regions that overlap DHS peaks manifest different coverage and tail behavior compared to TRE regions that do not overlap DHS peaks.

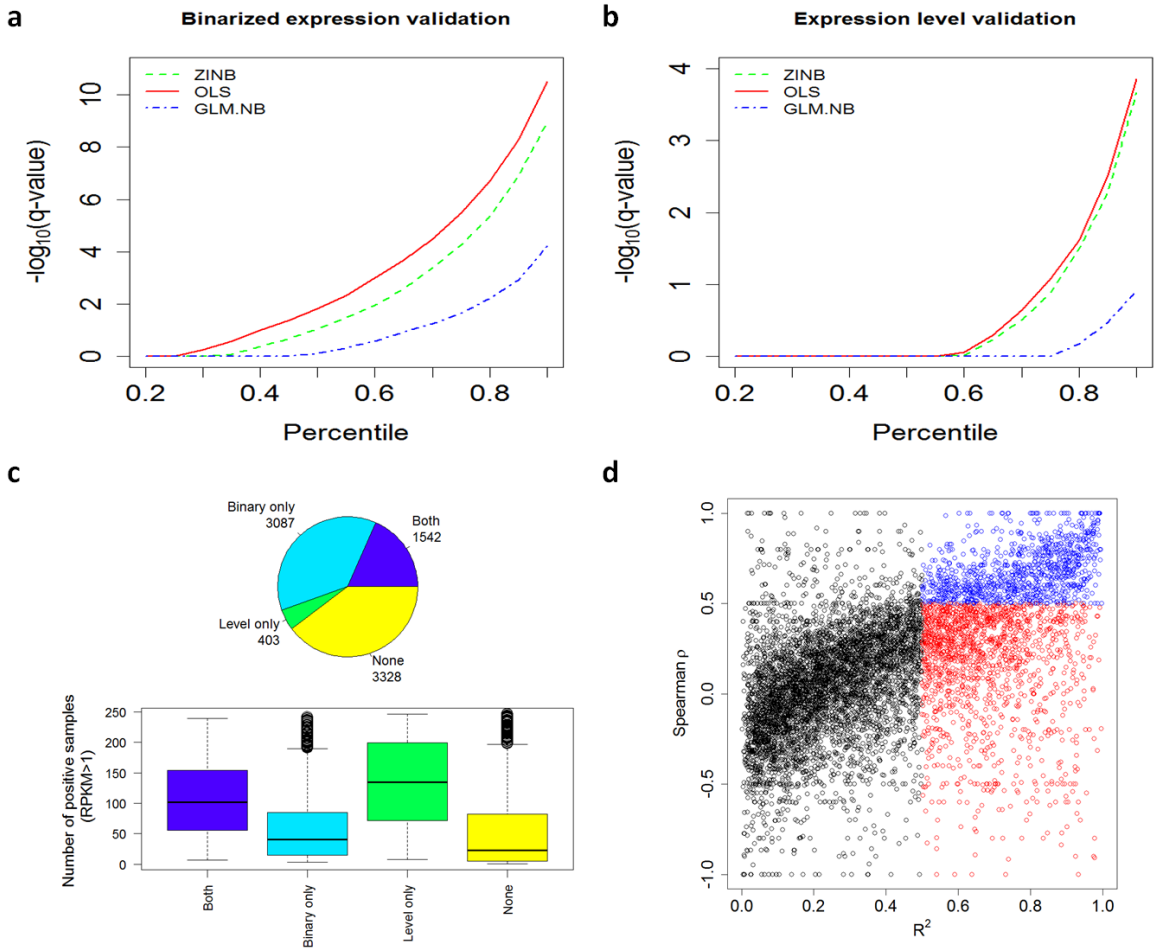


Figure S.4. Performance of methods for constructing enhancer-promoter models without intronic E-P links. (a) Binarized expression validation scores. The x-axis is the percentiles of the $-\log_{10}$[q-values] computed by Wilcoxon rank sum test. (b) Expression level validation scores. The x-axis is the percentiles of the $-\log_{10}$[q-values] computed by the Spearman correlation test. Both plots show advantage of OLS over the other methods. (c) Top: Breakdown of the genes whose OLS models passed each of the validations. Binary/Level only: genes that passed only binary/level validation (q<0.1). The number of genes in each category is shown next to each pie slice. Bottom: The distribution of the number of samples that showed positive expression (RPKM≥ 1) for the genes in each category. The 'Level only' and 'Both' categories capture the majority of the genes that had many samples with positive expression. (d) Comparison between Spearman ρ correlation obtained by the OLS models and gene model R^2 values as computed by FANTOM5 (without cross validation). Blue dots: genes with $R^2 \geq 0.5$ and $\rho \geq 0.5$; red dots: genes with $R^2 \geq 0.5$ and $\rho < 0.5$. Gene model selection based on R^2 might produce many over-fitted models (red dots).

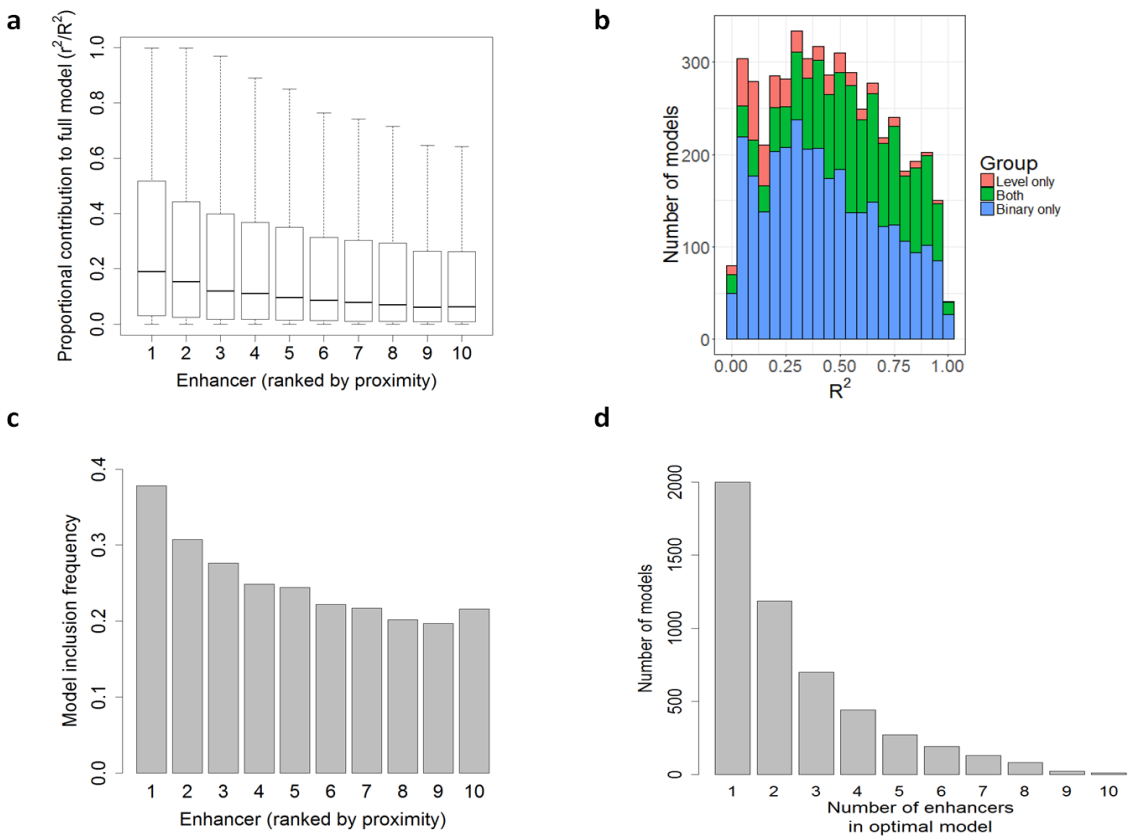


Figure S.5. Enhancer contribution to full and shrunken models constructed without intronic E-P links. In each model the ten most proximal enhancers (within ±500 kb) of the gene's TSS were considered. The full model used all ten, while the shrunken model removed some enhancers based on their scores. (a) The proportional contribution of each enhancer in the full model. Enhancers are ranked by their distance from the TSS (1 is the closest). (b) R^2 value distribution of the full models (n=5,032) by groups (See **Supplementary Figure S.4.c** for details). Approximately 44% of 5,032 models had $R^2 \geq 0.5$. (c) Inclusion rate of enhancers in the models after shrinkage. Enhancers are ranked by their distance from the TSS (1 is the closest). (d) Histogram of the number of enhancers included in the shrunken models.

8. Supplementary Tables

Table S.1. 246 GRO-Seq samples			
Sample accessions	**Cell line**	**Treated with**	**SRAs**
GSM1366021	HCT116	DMSO (10uM) for 0.5hr, wt	SRR1224573
GSM1366022	HCT116	Nutlin3a (10uM) for 0.5hr, wt	SRR1224574
GSM1304424	HCT116	DMSO (10uM) for 1hr, control genotype, rep1	SRR1105736,SRR1105737
GSM1304425	HCT116	Nutlin3a (10uM) for 1hr, control genotype, rep1	SRR1105738,SRR1105739
GSM1304426	HCT116	DMSO (10uM) for 1hr	SRR1105740
GSM1304427	HCT116	Nutlin3a (10uM) for 1hr	SRR1105741
GSM1847255	HCT116	Akti-1/2 treated	SRR2153512
GSM1847256	HCT116	Akti-1/2 treated	SRR2153513
GSM1847257	HCT116	DMSO treated	SRR2153514
GSM1847258	HCT116	DMSO treated	SRR2153515
GSM1727121	HCT116	Scramble shRNA	SRR2084588,SRR2084589, SRR2084590,SRR2084591
GSM1727119	HCT116	PAF1 shRNA #1	SRR2084580,SRR2084581, SRR2084582
GSM1727120	HCT116	Scramble shRNA	SRR2084584,SRR2084585, SRR2084586,SRR2084587
GSM1727118	HCT116	PAF1 shRNA #1	SRR2084576,SRR2084577, SRR2084578,SRR2084579
GSM1727119	HCT116	PAF1 shRNA #1	SRR2084583
GSM1124062	HCT116	DMSO	SRR828695,SRR828696, SRR828729
GSM1422215	MCF7	control, rep1	SRR1501091
GSM1422216	MCF7	control, rep2	SRR1501092
GSM1422217	MCF7	Nutlin3a (8uM) for 16hr, rep1	SRR1501093
GSM1422218	MCF7	Nutlin3a (8uM) for 16hr, rep2	SRR1501094
GSM1115995	MCF7	rep1,100nM E2 1hr treatment	SRR816998
GSM1115996	MCF7	rep2, 100nM E2 1hr treatment	SRR816999
GSM1115997	MCF7	rep1, EtOH 1hr treatment, control	SRR817000
GSM1115998	MCF7	rep2, EtOH 1hr treatment, control	SRR817001

GSM1115999	MCF7	1hr treatment with 100nM E2	SRR817002
GSM1116000	MCF7	1hr treatment with ethanol	SRR817003
GSM1067410, GSM678535	MCF7	untreated,Re-Sequenced GSM678535 to greater depth	SRR653421,SRR497904, SRR497905,SRR497906
GSM1067411, GSM678536	MCF7	untreated,Re-Sequenced GSM678536 to greater depth	SRR653422,SRR497907, SRR497908,SRR497909, SRR497910
GSM1067412, GSM678537	MCF7	treated for 10 minutes with E2,Re-Sequenced GSM678537 to greater depth	SRR653423,SRR497911
GSM1067413, GSM678538	MCF7	treated for 10 minutes with E2,Re-Sequenced GSM678538 to greater depth	SRR653424,SRR497912, SRR497913
GSM1067414, GSM678539	MCF7	treated for 40 minutes with E2,Re-Sequenced GSM678539 to greater depth	SRR653425,SRR497914, SRR497915,SRR497916
GSM1067415, GSM678540	MCF7	treated for 40 minutes with E2,Re-Sequenced GSM678540 to greater depth	SRR653426,SRR497917, SRR497918,SRR497919, SRR497920
GSM678541	MCF7	treated for 160 minutes with 100 nM E2,rep1	SRR497921
GSM678542	MCF7	treated for 160 minutes with 100 nM E2,rep2	SRR497922,SRR497923
GSM1847251	MCF7	Akti-1/2 treated	SRR2153508
GSM1847252	MCF7	Akti-1/2 treated	SRR2153509
GSM1847253	MCF7	DMSO treated	SRR2153510
GSM1847254	MCF7	DMSO treated	SRR2153511
GSM1911184	MCF7	GRO-seq in vehicle treated MCF7 cells with shRNA-mediated knockdown of Luciferase as a control	SRR2724029
GSM1911185	MCF7	GRO-seq in vehicle treated MCF7 cells with shRNA-mediated knockdown of Luciferase as a control	SRR2724030
GSM1911186	MCF7	GRO-seq in vehicle treated MCF7 cells with shRNA-mediated knockdown of PARP-1	SRR2724031

GSM1911187	MCF7	GRO-seq in vehicle treated MCF7 cells with shRNA-mediated knockdown of PARP-1	SRR2724032
GSM2151684	MCF7	GRO-seq in MCF7 cells with shRNA-mediated knockdown of Luciferase as a control with 1 hour 1♀μM PJ34 Treatment	SRR3500459
GSM1523191	MCF7	cells were transfected with siCTL, after 1hr treatment with ethanol	SRR1609863
GSM1523192	MCF7	cells were transfected with siCTL, after 1hr treatment with 100nM E2	SRR1609864
GSM1523193	MCF7	cells were transfected with siNCAPG, after 1hr treatment with ethanol	SRR1609865
GSM1523194	MCF7	cells were transfected with siNCAPG, after 1hr treatment with 100nM E2	SRR1609866
GSM1523195	MCF7	cells were transfected with siCTL, after 1hr treatment with ethanol	SRR1609867
GSM1523196	MCF7	cells were transfected with siCTL, after 1hr treatment with 100nM E2	SRR1609868
GSM1523197	MCF7	cells were transfected with siNCAPD3, after 1hr treatment with ethanol	SRR1609869
GSM1523198	MCF7	cells were transfected with siNCAPD3, after 1hr treatment with 100nM E2	SRR1609870
GSM1438934	MCF7	Vehicle for 40min	SRR1519032
GSM1438935	MCF7	Vehicle for 40min	SRR1519033
GSM1438936	MCF7	Vehicle for 40min	SRR1519034
GSM1438937	MCF7	100 nM 17 β-estradiol (E2) for 40min	SRR1519035
GSM1438938	MCF7	100 nM 17 β -estradiol (E2) for 40min	SRR1519036
GSM1438939	MCF7	100 nM 17 β -estradiol (E2) for 40min	SRR1519037

GSM1438940	MCF7	25 ng/mL TNFa for 40min	SRR1519038
GSM1438941	MCF7	25 ng/mL TNFa for 40min	SRR1519039
GSM1438942	MCF7	25 ng/mL TNFa for 40min	SRR1519040
GSM1438943	MCF7	100 nM 17 β -estradiol + 25 ng/mL TNFa for 40min	SRR1519041
GSM1438944	MCF7	100 nM 17 β -estradiol + 25 ng/mL TNFa for 40min	SRR1519042
GSM1438945	MCF7	100 nM 17 β -estradiol + 25 ng/mL TNFa for 40min	SRR1519043
GSM1470027	MCF7	siCTL, 1hr treatment with vehicle control	SRR1542320
GSM1470028	MCF7	siCTL, 1hr treatment with 100nM E2	SRR1542321
GSM1470029	MCF7	siGATA3, 1hr treatment with vehicle control	SRR1542322
GSM1470030	MCF7	siGATA3, 1hr treatment with 100nM E2	SRR1542323
GSM1470031	MCF7	shCTL, 1hr treatment with vehicle control	SRR1542324
GSM1470032	MCF7	shCTL, 1hr treatment with 1μM RA	SRR1542325
GSM1470033	MCF7	shCTL, 1hr treatment with 100nM E2	SRR1542326
GSM1470034	MCF7	shRARs, 1hr treatment with vehicle control	SRR1542327
GSM1470035	MCF7	shRARs, 1hr treatment with 1μM RA	SRR1542328
GSM1470036	MCF7	shRARs, 1hr treatment with 100nM E2	SRR1542329
GSM1470037	MCF7	shCTL, 1hr treatment with vehicle control	SRR1542330
GSM1470038	MCF7	shCTL, 1hr treatment with 100nM E2	SRR1542331
GSM1470039	MCF7	shAP2g, 1hr treatment with vehicle control	SRR1542332
GSM1470040	MCF7	shAP2g, 1hr treatment with 100nM E2	SRR1542333
GSM1014637	MCF7	none	SRR579299,SRR579300, SRR579301,SRR579302, SRR579303,SRR579304, SRR579305,SRR579306, SRR579307,SRR579308,

			SRR579309
GSM1014638	MCF7	none	SRR579310,SRR579311, SRR579312,SRR579313, SRR579314,SRR579315, SRR579316,SRR579317, SRR579318
GSM1014639	MCF7	E2 100 nM, 0min	SRR579319,SRR579320, SRR579321,SRR579322, SRR579323
GSM1014640	MCF7	E2 100 nM, 10min	SRR579324,SRR579325, SRR579326,SRR579327, SRR579328,SRR579329, SRR579330,SRR579331, SRR579332,SRR579333, SRR579334
GSM1014641	MCF7	E2 100 nM, 10min	SRR579335,SRR579336, SRR579337,SRR579338, SRR579339,SRR579340, SRR579341
GSM1014642	MCF7	E2 100 nM, 10min	SRR579342,SRR579343, SRR579344,SRR579345, SRR579346
GSM1014643	MCF7	E2 100 nM, 25min	SRR579347,SRR579348, SRR579349,SRR579350, SRR579351,SRR579352
GSM1014644	MCF7	E2 100 nM, 25min	SRR579353,SRR579354, SRR579355,SRR579356, SRR579357,SRR579358, SRR579359
GSM1014645	MCF7	E2 100 nM, 40min	SRR579360,SRR579361, SRR579362,SRR579363
GSM1014646	MCF7	E2 100 nM, 40min	SRR579364,SRR579365, SRR579366,SRR579367, SRR579368,SRR579369
GSM1014647	MCF7	E2 100 nM, 40min	SRR579370,SRR579371, SRR579372,SRR579373, SRR579374,SRR579375, SRR579376,SRR579377, SRR579378
GSM1382433	A375	DMSO	SRR1275489
GSM1382434	A375	DMSO	SRR1275490
GSM1382435	A375	25 μM Leflunomide	SRR1275491
GSM1382436	A375	25 μM Leflunomide	SRR1275492
GSM1382437	A375	25 μM A771726	SRR1275493
GSM1382438	A375	25 μM A771726	SRR1275494
GSM1634453	U2OS	control	SRR1916552
GSM1634454	U2OS	Myc activation (5 hr)	SRR1916553

GSM1634455	U2OS	control	SRR1916554
GSM1634456	U2OS	Myc activation (5 hr)	SRR1916555
GSM1622612	K562	control (37 degree C)	SRR1823901
GSM1622613	K562	control (37 degree C)	SRR1823902
GSM1622614	K562	heat shock (30min at 43 degree C)	SRR1823903
GSM1622615	K562	heat shock (30min at 43 degree C)	SRR1823904
GSM1480325	K562	none	SRR1552484
GSM1579367	hESC	untreated	SRR1745515
GSM1579368	hESC	untreated	SRR1745516
GSM1579369	hESC	Wnt3a 6h	SRR1745517
GSM1579370	hESC	Wnt3a 6h	SRR1745518
GSM1579371	hESC	Wnt3a+Activin A (WA) 6h	SRR1745519
GSM1579372	hESC	Wnt3a+Activin A (WA) 6h	SRR1745520
GSM1579373	hESC	Activin A 6h	SRR1745521
GSM1006728	hESC	none	SRR574824,SRR574825, SRR574826
GSM1006729	hESC	RPMI, 0hr	SRR574827,SRR574828
GSM1006730	hESC	Activin 50ng/ml, 1hr	SRR574829,SRR574830
GSM1006731	hESC	Activin 50ng/ml, 48hr	SRR574831
GSM1648604	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible LUC construct, 24hr	SRR1950491,SRR1950492
GSM1648605	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible LUC construct, 24hr	SRR1950493,SRR1950494
GSM1648606	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible TEL-AML1 construct, 0hr	SRR1950495,SRR1950496
GSM1648607	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible TEL-AML1 construct, 0hr	SRR1950497,SRR1950498
GSM1648608	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible TEL-AML1 construct, 4hr	SRR1950499,SRR1950500
GSM1648609	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible TEL-AML1 construct, 4hr	SRR1950501,SRR1950502
GSM1648610	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible TEL-AML1 construct, 12hr	SRR1950503,SRR1950504
GSM1648611	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible TEL-	SRR1950505,SRR1950506

		AML1 construct, 12hr	
GSM1648612	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible TEL-AML1 construct, 24hr	SRR1950507,SRR1950508
GSM1648613	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible TEL-AML1 construct, 24hr	SRR1950509,SRR1950510
GSM1648614	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible TEL-AML1-mut (dna-binding deficient) construct, 24hr	SRR1950511,SRR1950512
GSM1648615	Nalm6_pre_B_ALL	Nalm6 pre-B-ALL cell line with inducible TEL-AML1-mut (dna-binding deficient) construct, 24hr	SRR1950513,SRR1950514
GSM1278354	A549	control siRNAs	SRR1041870
GSM1278355	A549	control siRNAs	SRR1041871
GSM1278356	A549	siSetx	SRR1041872
GSM1278357	A549	siSetx	SRR1041873
GSM874647	A549	none (uninfected)	SRR408117
GSM874648	A549	Flag-NS1 virus, 12hr wt	SRR408118
GSM874649	A549	Flag-delta PAF virus, 12hr	SRR408119
GSM1524923	Ramos_cell_line	none	SRR1611840
GSM1480326	GM12878	none	SRR1552485
GSM980644	GM12004	none	SRR531824
GSM980645	GM12750	none	SRR531825
GSM1543777	LNCaP	GRO-seq, siControl, 1h treatment with vehicle	SRR1648886
GSM1543778	LNCaP	GRO-seq, siControl, 1h treatment with 100 nM DHT	SRR1648887
GSM1543779	LNCaP	GRO-seq, siTOP1, 1h treatment with vehicle	SRR1648888
GSM1543780	LNCaP	GRO-seq, siTOP1, 1h treatment with 100 nM DHT	SRR1648889
GSM1543781	LNCaP	GRO-seq, siControl, 1h treatment with vehicle	SRR1648890
GSM1543782	LNCaP	GRO-seq, siControl, 1h treatment with 100 nM DHT	SRR1648891
GSM1543783	LNCaP	GRO-seq, siMRE11, 1h treatment with vehicle	SRR1648892

GSM1543784	LNCaP	GRO-seq, siMRE11, 1h treatment with 100 nM DHT	SRR1648893
GSM1543785	LNCaP	GRO-seq, siNKX3.1, 1h treatment with vehicle	SRR1648894
GSM1543786	LNCaP	GRO-seq, siNKX3.1, 1h treatment with 100 nM DHT	SRR1648895
GSM1543787	LNCaP	GRO-seq, siControl, 1h treatment with vehicle	SRR1648896,SRR1648897
GSM1543788	LNCaP	GRO-seq, siControl, 1h treatment with 100 nM DHT	SRR1648898,SRR1648899
GSM1543789	LNCaP	GRO-seq, siMRE11, 1h treatment with vehicle	SRR1648900
GSM1543790	LNCaP	GRO-seq, siMRE11, 1h treatment with 100 nM DHT	SRR1648901
GSM1543796	LNCaP	GRO-seq, siControl, 1h treatment with vehicle	SRR1648909
GSM1543797	LNCaP	GRO-seq, siControl, 1h treatment with 100 nM DHT	SRR1648910
GSM1543798	LNCaP	GRO-seq, siTOP1, 1h treatment with vehicle	SRR1648911
GSM1543799	LNCaP	GRO-seq, siTOP1, 1h treatment with 100 nM DHT	SRR1648912
GSM1348226	LNCaP	DHT	SRR1192053
GSM1348227	LNCaP	vehicle only	SRR1192054
GSM1348228	LNCaP	SD70 + DHT	SRR1192055
GSM1348229	LNCaP	SD70	SRR1192056
GSM1159899	LNCaP	siCTL	SRR892025
GSM1159900	LNCaP	siPCGEM1	SRR892026
GSM1159901	LNCaP	siPRNCR1	SRR892027
GSM1159902	LNCaP	siCTL,100 nM DHT	SRR892028
GSM1159903	LNCaP	siPCGEM1,100 nM DHT	SRR892029
GSM1159904	LNCaP	siPRNCR1,100 nM DHT	SRR892030
GSM1159895	LNCaP	siCTL	SRR892016
GSM1159896	LNCaP	siCTL,100 nM DHT	SRR892017
GSM1159897	LNCaP	siPYGO2	SRR892018
GSM1159898	LNCaP	siPYGO2,100 nM DHT	SRR892019
GSM686948	LNCaP	siCTRL (1027280),vehicle	SRR122339
GSM686949	LNCaP	siCTRL (1027280), DHT	SRR122340

GSM686950	LNCaP	siFoxA1 (M-010319), DHT	SRR122341
GSM1553207	HEK293	doxycycline induced, a-amanitin (2.5 mg/ml) 42 hr	SRR1661564
GSM1553208	HEK293	DRB (100 mM) 3hr	SRR1661565
GSM1553209	HEK293	DRB washout t=10min.	SRR1661566
GSM1553210	HEK293	DRB washout t=20min.	SRR1661567
GSM1553211	HEK293	doxycycline induced, a-amanitin (2.5 mg/ml) 42 hr	SRR1661568
GSM1553212	HEK293	DRB (100 mM) 3hr	SRR1661569
GSM1553213	HEK293	DRB washout t=10min.	SRR1661570
GSM1553214	HEK293	DRB washout t=20min.	SRR1661571
GSM1553215	HEK293	doxycycline induced, a-amanitin (2.5 mg/ml) 42 hr	SRR1661572
GSM1553216	HEK293	DRB (100 mM) 3hr	SRR1661573
GSM1553217	HEK293	DRB washout t=10min.	SRR1661574
GSM1553218	HEK293	DRB washout t=20min.	SRR1661575
GSM1553219	HEK293	doxycycline induced, a-amanitin (2.5 mg/ml) 42 hr	SRR1661576
GSM1553220	HEK293	DRB (100 mM) 3hr	SRR1661577
GSM1553221	HEK293	DRB washout t=10min.	SRR1661578
GSM1553222	HEK293	DRB washout t=20min.	SRR1661579
GSM1249869	HEK293T	siCTL	SRR3317155
GSM1249870	HEK293T	siJMJD6-1	SRR3317156
GSM1249871	HEK293T	siJMJD6-2	SRR3317157
GSM1249872	HEK293T	siBrd4-1	SRR3317158
GSM1249873	HEK293T	siBrd4-2	SRR3317159
GSM1249874	HEK293T	siCTL	SRR3317160
GSM1249875	HEK293T	siJMJD6-1	SRR3317161
GSM1249876	HEK293T	siJMJD6-2	SRR3317162
GSM1249877	HEK293T	siBrd4-1	SRR3317163
GSM1249878	HEK293T	siBrd4-2	SRR3317164
GSM1273483	HUVEC	Notx-2h-rep1	SRR1035898
GSM1273484	HUVEC	VEGFA-2h-rep1	SRR1035899
GSM1412749	HUVEC	Notx-2h-rep2	SRR1406747
GSM1412750	HUVEC	VEGFA-2h-rep2	SRR1406748
GSM1273485	HAEC	Notx-2h	SRR1035900
GSM1273486	HAEC	VEGFA-2h	SRR1035901

GSM1405106	HeLa	CTRL,Nascent RNA was profiled after EGF stimulation in a doxocycline inducible knock-down cell line for INTS11, rep1	SRR1342250
GSM1405107	HeLa	DOX,Nascent RNA was profiled after EGF stimulation in a doxocycline inducible knock-down cell line for INTS11, rep1	SRR1342251
GSM1405108	HeLa	CTRL,Nascent RNA was profiled after EGF stimulation in a doxocycline inducible knock-down cell line for INTS11, rep2	SRR1342252
GSM1405109	HeLa	DOX,Nascent RNA was profiled after EGF stimulation in a doxocycline inducible knock-down cell line for INTS11, rep2	SRR1342253
GSM1518913	HeLa	none	SRR1596500
GSM1518914	HeLa	none	SRR1596501
GSM1240738	AC16	DMSO, TNFa, 0m	SRR1015583
GSM1240739	AC16	DMSO, TNFa, 0m	SRR1015584
GSM1240740	AC16	TNFa 25 ng/ml, 10m	SRR1015585
GSM1240741	AC16	TNFa 25 ng/ml, 10m	SRR1015586
GSM1240742	AC16	TNFa 25 ng/ml, 30m	SRR1015587
GSM1240743	AC16	TNFa 25 ng/ml, 30m	SRR1015588
GSM1240744	AC16	TNFa 25 ng/ml, 120m	SRR1015589
GSM1240745	AC16	TNFa 25 ng/ml, 120m	SRR1015590
GSM1240746	AC16	none	SRR1015591
GSM1240747	AC16	none	SRR1015592
GSM1240748	AC16	a-amanitin 1 µg/ml	SRR1015593
GSM1240749	AC16	a-amanitin 1 µg/ml	SRR1015594
GSM1014631	AC16	none	SRR579293
GSM1014632	AC16	none	SRR579294
GSM1014633	AC16	TNFa 25 ng/ml, 10m	SRR579295
GSM1014634	AC16	TNFa 25 ng/ml, 10m	SRR579296
GSM1014635	AC16	TNFa 25 ng/ml, 30m	SRR579297
GSM1014636	AC16	TNFa 25 ng/ml, 30m	SRR579298

GSM1055806	IMR90	none	SRR639050
GSM1055807	IMR90	TNF-a (10ng/mL) 1hr	SRR639051
GSM1171524	Immortalized human breast cancer cells	Vehicle rep1	SRR915731
GSM1171525	Immortalized human breast cancer cells	Vehicle rep2	SRR915732
GSM1171526	Immortalized human breast cancer cells	10nM E2 rep1	SRR915733
GSM1171527	Immortalized human breast cancer cells	10nM E2 rep2	SRR915734
GSM1045177	NTera2_D1	none	SRR620530
GSM1045178	NTera2_D1	1 μM atRA, 2 days	SRR620531
GSM2235679	VCaP	Ethanol,2h	SRR3923617
GSM2235680	VCaP	Ethanol,2h	SRR3923618
GSM2262426	VCaP	Ethanol,4h	SRR4001595
GSM2262427	VCaP	Ethanol,4h	SRR4001596
GSM2262428	VCaP	100nM DHT,4h	SRR4001597
GSM2262429	VCaP	100nM DHT,4h	SRR4001598
GSM2235681	VCaP	10nM R1881,30min	SRR3923619
GSM2235682	VCaP	10nM R1881,30min	SRR3923620
GSM2235683	VCaP	10nM R1881,2h	SRR3923621
GSM2235684	VCaP	10nM R1881,2h	SRR3923622

Table S.2. Available data in ENCODE project					
Cell line	**DNase-seq**	**H3K4me1**	**H3K4ac27**	**P300**	**POL2 ChIA-PET**
MCF7	+	+	+	+	+
HCT116	+	+	+	+*	+
IMR90	+	+	+	-	-
K562	+	+	+	+	+
LNCaP	+	-	-	-	-
* P300 ChIP-seq data was taken from GSE51176 GEO series (sample id GSM1240110)					

Table S.3. Comparison of dREG and groHMM: results of the HCT116 cell line			
Sample type	**Group**	**Set size**	**Percent covered enhancers***
Control	DHS-TRE	13,602	73.0
Control	DHS-DNT	4,584	30.0
Control	DHS-TRE-DNT	2,385	74.4
Treatment	DHS-TRE	15,853	66.0
Treatment	DHS-DNT	3,983	32.9
Treatment	DHS-TRE-DNT	2,290	68.1
* The percent of reported enhancers that had any overlap with some DHS peak. For DHS-TRE-DNT we computed the percentage as follows: (1) we found the set E of dREG enhancers that overlap groHMM enhancers, (2) we found the set $\boldsymbol{E'} \subseteq \boldsymbol{E}$ of enhancers that overlap some DHS peak in DHS-TRE-DNT group, and (3) we computed the percentage by dividing $\lvert\boldsymbol{E'}\rvert$ with $\lvert\boldsymbol{E}\rvert$.			

Table S.4. Number of gene models in each regression method under FDR 0.1				
Method	**Binary only**	**Expression level only**	**Both**	**None**
OLS	2,580	236	3,507	2,037
GLM.NB	2,659	377	606	4,718
ZINB	2,844	657	1,334	3,525
Each gene model contained 10 enhancers as features. The number of E-P links is $\boldsymbol{y} \cdot \mathbf{10}$ links where $\boldsymbol{y}$ is the number of gene models in each category				

Table S.5. Number of gene models in each regression method under FDR 0.2				
Method	**Binary only**	**Expression level only**	**Both**	**None**
OLS	2,509	249	3,745	1,857
GLM.NB	2,830	453	798	4,279
ZINB	2,907	681	1,566	3,206
Each gene model contained 10 enhancers as features. The number of E-P links is $\boldsymbol{y} \cdot \mathbf{10}$ links where $\boldsymbol{y}$ is the number of gene models in each category				

תקציר

זיהוי נרחב של אזורי מעצמים (enhancers, מקדמי שיעתוק גנים) בגנום ומיפוי אזורים אלו לגן המטרה שלהם הם מטרות מפתח בגנומיקה פונקציונלית. כיום, זיהוי אזורי המעצמים בעיקר מבוסס על סמנים אפיגנטיים, כגון H3K4me1, H3K27ac ו-p300, הנמצאים בקורלציה עם מעצם פעיל. יחד עם זאת, ניסויים בהיקף גדול שנעשו לאחרונה קובעים שהרבה מאזורי המעצמים האלה אינם פונקציונאליים.

Enhancer RNAs (eRNAs) הם קבוצה חדשה של תעתיקים המשועתקים מאזורי מעצמים אקטיביים ובדרך כלל בשיעתוק דו-כיווני. לאחרונה, הוצע להשתמש בביטוי של eRNAs כסמן למעצמים אקטיביים ופונקציונאליים. eRNAs לרוב אינם יציבים ומתפרקים מהר בהשוואה ל-RNA שליח המשועתק מגנים עקב מחסור בפולי-אדנילציה. לכן, לא ניתן למדוד את ביטוי ה-eRNAs בשיטות RNA-seq הסטנדרטיות. שיטת ה- GRO-seq מאפשרת לאמוד את ביטוי ה-eRNAs דרך מדידת קצב השיעתוק של RNAs שלא עברו שיחבור חלופי בסקאלה גנומית.

בעבודה זו פיתחנו שיטה חישובית חדשה הנקראת FOCS (FDR-corrected OLS with Cross-validation and Shrinkage) למיפוי מעצמים לגן המטרה שלהם. FOCS עושה שימוש בשיטת רגרסיה לינארית למידול תבנית ביטוי של גן באמצעות תבניות הביטוי של המעצמים הקרובים לגן תוך שימוש ב-elastic-net לצמצום מספר המעצמים בכל מודל של גן וביצוע אימות הצלבה (Cross-validation) לאורך כל סוגי התאים כדי למנוע התאמה יתירה של המודל לנתונים. הפעלנו את FOCS על מידע זמין של GRO-seq מ-40 ניסויים המקיפים 246 דגימות, שמקורן מ-23 סוגי תאים ונבחנו תחת תנאים רגילים ולחץ. השתמשנו במידע חיצוני, ChIA-PET ו-GTEx eQTLs, כדי לאמת את החיזויים של מעצם-גן. אנו מראים ש-FOCS ממפה בצורה טובה יותר, גם באיכות וגם בכמות האינטראקציות, לעומת שיטות קודמות למיפוי כגון קורלציה בין כל זוג מעצם-גן או מידול ביטוי הגן באמצעות רגרסיה מבוססת LASSO.

אוניברסיטת תל אביב

הפקולטה למדעים מדוייקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע"ש בלבטניק

שימוש בנתוני עתק גנומיים לחיזוי אינטראקציות מעצם-פרומוטר

חיבור זה הוגש כעבודת גמר לקראת התואר "מוסמך אוניברסיטה" בבית הספר למדעי המחשב

על ידי

תום אהרן עיט

העבודה נעשתה בבית הספר למדעי המחשב, אוניברסיטת תל אביב

בהנחיית

פרופ' רון שמיר

דר' רן אלקון

אייר התשע"ז