# Deciphering the Quantitative Effects of Cooperativity and Mutations on Transcription Factor Binding

by

Vincentius Martin

Department of Computer Science
Duke University

Date: _____
Approved:

_____
Raluca Gordân, Supervisor

_____
Alberto Bartesaghi

_____
Gregory E. Crawford

_____
Alexander Hartemink

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Computer Science
in the Graduate School of
Duke University

2022

# ABSTRACT

## Deciphering the Quantitative Effects of Cooperativity and Mutations on Transcription Factor Binding

by

Vincentius Martin

Department of Computer Science
Duke University

Date: _____
Approved:

_____
Raluca Gordân, Supervisor

_____
Alberto Bartesaghi

_____
Gregory E. Crawford

_____
Alexander Hartemink

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Computer Science
in the Graduate School of
Duke University

2022

Copyright © 2022 by Vincentius Martin
All rights reserved

# Abstract

Transcription factor (TF) proteins bind to DNA in a sequence specific manner to regulate gene expression. The binding affinity of TFs for individual sites is well characterized and can be represented using DNA motif models such as position weight matrices. However, there are many factors influencing TF-DNA recognition in the cell, leading to complexities than cannot be captured by motif models alone. Here, we present our studies on two factors: cooperative TF binding and alterations in TF binding due to DNA mutations. Both factors require quantitative and rigorous approaches to distinguish real effects from random noise.

First, we present a new method for characterizing cooperative binding of TFs to DNA. This method addresses the issue that TF binding sites located in close proximity, which occurs frequently across the human genome, are not necessarily bound cooperatively. To distinguish between cooperative and independent binding, we developed a high-throughput on-chip binding assay designed specifically to measure TF binding to neighboring sites. Using the experimental data from our assay, we trained machine learning models to differentiate between cooperative and independent binding of TFs. This method enabled us to reveal molecular mechanisms used by TFs to bind DNA cooperatively.

Second, we introduce QBiC-Pred (Quantitative Predictions of TF Binding Changes Due to Sequence Variants), an ordinary least squares based method to predict the magnitude of the effect of DNA mutations on TF-DNA recognition. We implemented QBiC-Pred as a web service: qbic.genome.duke.edu, which allows users to run our models through a user-friendly web interface. We used this method to identify non-recurring putative regulatory driver mutations in cancer. Our approach is novel because we prioritize mutations based on their effects on transcription factor (TF)

binding, instead of relying on the recurrence of the mutations among tumor samples—which is often difficult to perform as individual non-coding mutations are rarely seen in more than one donor. Focusing on the functional effects of non-coding mutations across regulatory regions, we identified dozens of genes whose regulation in tumor cells is likely to be significantly perturbed by non-coding mutations.

# Contents

**Abstract** **iv**

**List of Tables** **x**

**List of Figures** **xi**

**Acknowledgements** **xiv**

**1 Introduction** **1**

1.1 DNA recognition by transcription factors . . . . . 2

1.2 Factors modulating TF-DNA interactions . . . . . 5

1.2.1 Indirect TF-DNA interactions . . . . . 5

1.2.2 Competitive binding between paralogous TFs . . . . . 6

1.2.3 TF binding in the presence of nucleosomes . . . . . 6

1.2.4 Other factors modulating TF-DNA interactions . . . . . 8

1.3 Cooperative binding of TFs to DNA . . . . . 8

1.3.1 Clusters of TF binding sites . . . . . 10

1.3.2 Physical mechanisms of cooperative binding . . . . . 12

1.3.3 Technologies for identifying cooperative binding . . . . . 13

1.3.4 Our contribution: Cooperative gcPBM . . . . . 17

1.4 Quantifying the effect of DNA mutations on TF binding . . . . . 18

1.4.1 Role of non-coding mutations in diseases . . . . . 18

1.4.2 Computational methods for quantifying the effects of DNA mutations on TF binding . . . . . 20

1.4.3 Our Contribution: QBiC-Pred . . . . . 22

**2 High-throughput data and modeling reveal insights into the mechanisms of cooperative DNA-binding by transcription factors 24**

2.1 Introduction . . . . . 24

2.2 Materials and methods . . . . . 26

2.2.1 Identifying neighboring binding sites across the genome . . . . 26

2.2.2 Design of DNA library for 'cooperative gcPBM' assay . . . . . 27

2.2.3 Cooperative gcPBM assay . . . . . 30

2.2.4 Identifying ETS1-RUNX1 cooperative versus independent binding events . . . . . 32

2.2.5 Identifying ETS1-ETS1 cooperative versus independent binding events . . . . . 37

2.2.6 Designing genomic features to distinguish cooperative versus independent TF binding to neighboring sites . . . . . 39

2.3 Results . . . . . 41

2.3.1 Cooperative gcPBM data reveals ETS1 cooperative binding to neighboring sites . . . . . 41

2.3.2 Features encoded in the DNA are highly predictive of ETS1 cooperative binding . . . . . 43

2.3.3 Binding site strength is highly predictive of ETS1-RUNX1 cooperative DNA binding . . . . . 45

2.3.4 The role of binding site distance and orientation in ETS1-RUNX1 cooperative DNA binding . . . . . 49

2.3.5 Binding site distance and orientation are predictive of ETS1-ETS1 cooperative binding . . . . . 51

2.3.6 Using DNA shape and sequence features to distinguish cooperative versus independent binding . . . . . 57

2.4 Discussion . . . . . 58

**3 QBiC-Pred: Quantitative Predictions of Transcription Factor Binding Changes Due to Sequence Variants 62**

3.1 Introduction . . . . . 62

3.2 Materials and methods . . . . . 65

3.2.1 OLS models of TF-DNA binding specificity . . . . . 65

3.2.2 *In vitro* measurements of TF binding changes due to single nucleotide variants . . . . . 67

3.2.3 *In vivo* allele-specific binding data . . . . . 68

3.2.4 QBiC-Pred web server implementation . . . . . 69

3.3 Results . . . . . 71

3.3.1 OLS performs better than Elastic Net in predicting *in vitro* TF binding changes . . . . . 71

3.3.2 OLS models of TF binding specificity outperform PWMs and DeepBind models in predicting *in vitro* TF binding changes . 73

3.3.3 The cross-validation accuracy of OLS models correlates with their accuracy in predicting *in vitro* TF binding changes . . . 75

3.3.4 OLS models of TF binding specificity outperform PWMs and DeepBind models in predicting *in vivo* allele-specific binding variants . . . . . 77

3.4 Discussion . . . . . 79

**4 Transcription Factor-Centric Approach to Identify Non-Recurring Putative Regulatory Drivers in Cancer 82**

4.1 Introduction . . . . . 82

4.2 Materials and methods . . . . . 85

4.2.1 ICGC simple somatic mutations and gene expression data . . 85

4.2.2 Promoter and enhancer data . . . . . 85

4.2.3 Defining the effects of mutations on TF binding, and the significance of these effects . . . . . 87

4.2.4 Analytical and simulation-based approaches to compute the significance of mutation effects on TF binding . . . . . 91

4.2.5 Integrating results across all regulatory regions of a gene . . . 94

4.3 Results . . . . . 97

4.3.1 Integrated analysis across regulatory regions identifies 54 genes with significant TF binding changes due to mutations in regulatory DNA . . . . . 97

4.3.2 Genes with significant mutations in their regulatory regions show large expression differences in mutated versus non-mutated samples . . . . . 102

**5 Conclusions** **108**

**Bibliography** **113**

**Biography** **125**

# List of Tables

4.1 Genes with significant TF binding changes due to mutations in regulatory DNA . . . . . 99

# List of Figures


1.1 TF-DNA interactions represent a key mechanism in the regulation of gene expression . . . 2

1.2 Factors modulating TF-DNA interactions . . . 3

1.3 Clusters of TF binding sites . . . 9

1.4 Cooperative binding mechanisms . . . 13

1.5 QBiC-Pred workflow . . . 23

2.1 Experimental approach to detect cooperative TF binding . . . 28

2.2 ChIP-seq signal for peaks with multiple putative Ets1 binding sites . 29

2.3 Distribution of distances between ChIP-seq peak summits for Ets1 and Runx1 . . . 30

2.4 Designing DNA probes with mutated Ets1 binding sites for the Ets1-Ets1 DNA library . . . 31

2.5 Ets1 binding signal distribution in sequences containing neighboring Ets1+Runx1 binding sites . . . 33

2.6 Possible p-value for a one-sided Mann-Whitney $U$ test . . . 34

2.7 Normalization of the Ets1 gcPBM data between the Ets1-only and the Ets1+Runx1 experiments . . . 35

2.8 Runx1-Ets1 cooperative gcPBM experiment . . . 36

2.9 Runx1 binding signal distribution in sequences containing neighboring Runx1-Ets1 binding sites . . . 36

2.10 Normalization of the Runx1 gcPBM data between the Runx1-only and the Runx1+Ets1 experiments . . . 37

2.11 Ets1 binding signal distribution . . . 38

2.12 Features used in cooperative binding classification models . . . . . 40

2.13 Cooperative vs. independent binding revealed by cooperative gcPBM data . . . . . 42

2.14 Cooperative vs. independent binding of Runx1, revealed by cooperative gcPBM data . . . . . 44

2.15 Distributions of Runx1 and Ets1 binding scores as identified from ChIP-seq peaks . . . . . 44

2.16 Random Forest classification models performance for ETS1-RUNX1 and ETS1-ETS1 systems . . . . . 45

2.17 Features used to train cooperative binding models on the ETS1-RUNX1 data . . . . . 46

2.18 Performance of Random Forest models trained separately on the binding site strength of either the main TF or the cooperator TF . . . . . 47

2.19 Random Forest classification models can accurately distinguish between cooperative and independent binding for Runx1-Ets1 . . . . . 47

2.20 Analyses of features used to train cooperative binding models on the Runx1-Ets1 data . . . . . 48

2.21 Binding strength of ETS1 and RUNX1 in various configurations . . . 52

2.22 Binding site strength of RUNX1 in ETS1-RUNX1 experiments . . . . 53

2.23 Features used to train cooperative binding models on the ETS1-ETS1 data . . . . . 54

2.24 Performance of Random Forest models trained separately on the binding site strength of ETS1-ETS1 . . . . . 56

2.25 DNA sequence and shape features are predictive of cooperative vs. independent binding . . . . . 58

3.1 TF binding change calculation . . . . . 66

3.2 QBiC-pred web server architecture . . . . . 69

3.3 Snapshot of QBiC-Pred web server results page . . . . . 70

3.4 The performance of OLS versus Elastic Net models . . . . . 73

3.5 Performance of OLS models . . . . . 75

3.6 Measured and predicted effects of SNPs in an ELK binding site and its flanking regions . . . . . 76

3.7 Relationship between OLS model quality and prediction accuracy . . 77

3.8 OLS vs other models in distinguishing between ASB and non-ASB variants . . . . . 79

4.1 Number of mutations in regulatory regions . . . . . 86

4.2 TF-centric approach to prioritize genes based on mutations in their regulatory regions . . . . . 88

4.3 Number of mutations in regulatory elements . . . . . 90

4.4 Entropy versus enhancer length in enhancer . . . . . 97

4.5 Genes prioritized based on mutations in their regulatory regions . . . 98

4.6 Number of prioritized genes in regulatory elements . . . . . 101

4.7 Analysis of differences in gene expression for genes with regulatory mutations . . . . . 103

4.8 Mutations in the ATXN3 enhancers that alter the binding specificity of TF RUNX1 . . . . . 105

4.9 Comparison between the genes' significance in the promoter-only versus the combined analysis . . . . . 106

4.10 Significance of gene expression differences in promoters and enhancers 106

# Acknowledgements

PhD journey has been a humbling experience. The more I learned, the more I realized I don't know. The experience propels me to keep learning and seek new ways to develop myself; not only as a scientist, but also as a person. And reaching the finish line would not have been possible without the presence of many people who challenged, supported, or (potentially) stuck with me, over the course of completing this journey.

First and foremost, I would like to thank my PhD advisor, Dr. Raluca Gordân for her constant guidance and support. From her, I learned how to be critical in dealing with my research problems and how to clearly communicate science in writings and presentations. Furthermore, I am grateful for her kindness in letting me exploring science in my own pace and more importantly, making mistakes and learned from them. I remember joining her group and told her that I know nothing about biology. And a simple utterance she said: "whenever there is a will, there is a way", allowed me to circumvent my ignorance and learned so much in the process. Raluca is a great role model and I am fortunate to finish my PhD as her student.

Next, I would like to thank my PhD committee members for their inputs and suggestions. Dr. Alexander Hartemink, whom I was fortunate to be his teaching assistant for two semesters in the Introduction to Computational Genomics class. Alex is a great teacher; the fluidity and clarity of his lectures inspired how I present science (and possibly, present anything in life). A notebook containing the notes I took from watching his lectures will probably accompany me wherever I go. Dr. Greg Crawford, for a prolific discussion during my dissertation defense, which invited me to think on the possible future applications of the computational models I developed. Dr. Alberto Bartesaghi for insight in how cryo-EM can possibly be used to shed light

in studying the cooperative binding mechanisms of transcription fators.

I would like to thank the former and current members of gordanlab for the great experience in exploring science and for the fellowships. Dr. Jingkang Zhao and Yuning Zhang for great collaborations in my PhD projects. Dr. Vishwa Nellore and Dr. Dinesh Manandhar for organizing the lab hang outs and making the first summer I joined the lab felt exciting. Farica Zhuang for being a close friend and a great research partner; also for the good conversations during rides to/from school and the myriad hang outs. Harshit Sahay for the inputs on the dissertation document. And eveyone whom I had a great pleasure meeting during my time in gordanlab: Dr. Ariel Afek, Zack Mielko, Yuze Hou, Wei Zhu, Olivier Boivin, Shengyu Li, Kyle Pinheiro, and Hana Wasserman.

I am also grateful for the friendship I made along the way. Sneha Mitra, Xioanan Hu, Yesenia Velasco, Lesia Semenova, Shuzhi Yu, and Hannah Kim have been friends whom I shared my PhD journey with, since I started the CS PhD program in 2016. Sara Harper and Stephen Harper for the fellowships, for introducing me to many US cultures, and for being like parents. Margie and Ed Secord for helping in my transition to US in 2016 and for introducing me to other Indonesians in NC.

I had a great time with many Indonesians I met during my PhD journey. Christina Sofian, definitely the best housemate and a great friend. I am thankful for all her support during the time I was stressed out with my research, all the life advice and life support she gave; and more importantly: great foods and desserts she made. Kevin Erning, Emilio Bachtiar, and Anastasia Sharon for the friendship and for making NC feels like home.

Aside from being a scientist, I am fortunate to be an active musicians where I studied with two inspiring music faculties in classical guitar and composition. First, I want to thank Randy Reed for polishing my skill as a classical guitarist. Classical

guitar has always been a part of me and being able to play many interesting repertoires kept me sane during the difficult times when I stuck with my research. Second, I thank Dr. Anthony Kelley, for the guidance in my composition study, which enabled me to express myself through writing music. I am fortunate to have some of my works performed by some great musicians during my study with Dr. Kelley.

Starting a PhD journey would not be possible if it was not because of the guidance from my former mentor, Dr. Haryadi Gunawi from The University of Chicago, Computer Science department. For two years starting from 2014, he taught me how to do research and he guided me through all the steps in applying for the PhD program in the US. I am grateful for the collaboration I had with him and I am thankful for the opportunity he gave to join his research.

Finally, I would like to thank my family for their constant support. Most importantly, I want to thank my wife Ventina Claudia for her love and presence, for keeping me sane during difficult times, and for her patience during busy times. I am eternally grateful and looking forward to continue my post-PhD journey with her.

# Chapter 1

# Introduction

Nearly every cell in a multicellular eukaryotic organism contains the same genetic information which is encoded in its DNA. Despite having the same DNA, each cell may differentiate into different cell types and perform different set of biological tasks. Cell differentiation happens through careful regulation of gene expression by binding of transcription factor (TF) proteins to specific locations on the DNA, enabling each cell to express a set of genes and synthesize proteins specific to its function (Figure 1.1). Therefore, understanding how TFs recognize their target DNA sequences is a critical step in understanding the regulation of gene expression.

DNA recognition by TFs is complex (Slattery et al., 2014; Siggers and Gordân, 2014) and there are many factors modulating these interactions (Lemon and Tjian, 2000; Inukai et al., 2017) (shown in Figure 1.2 and explained in Chapter 1.2). Furthermore, alterations in the DNA, such as DNA mutations, can change the binding of TFs and cause dysregulation of gene expression. These alterations have been linked to various diseases (Tseng et al., 2021) including cancer (Huang et al., 2013; Horn et al., 2013). In this study, we focus on two factors modulating TF-DNA interactions: cooperative binding between TFs (Figure 1.2A) and the effect of DNA mutations on TF binding (Figure 1.2B). Both problems require quantitative and rigorous approaches to distinguish real effects from random noise.

This chapter provides the necessary background to understand the mechanisms of TF-DNA interactions and briefly, on the factors modulating these interactions. Then we will focus on the two factors we studied. For both, we will review existing work

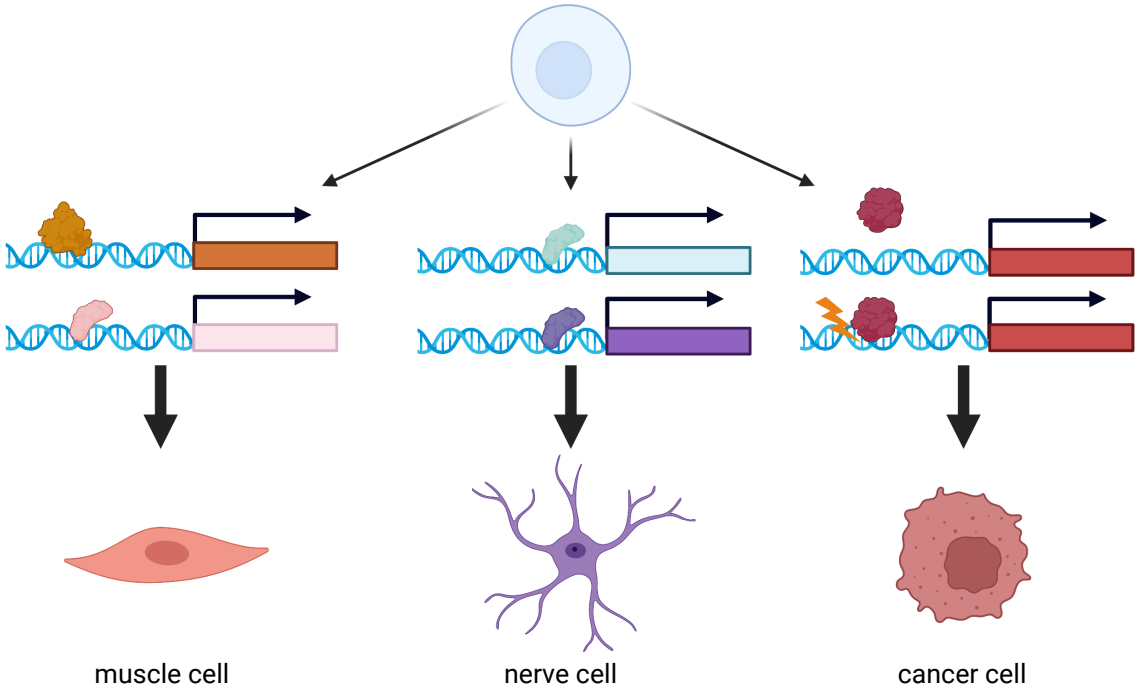



**Figure 1.1**: **TF-DNA interactions represent a key mechanism in the regulation of gene expression**. Combinations of TF binding events enable undifferentiated cells to express different genes and differentiate into specific cell types. Alterations in the DNA such as caused by mutations (as shown on the right) can alter TF binding and result in the dysregulation of gene expression. Created with BioRender.com.

and discuss how our studies fill the gap in current scientific knowledge.

## 1.1 DNA recognition by transcription factors

TFs regulate gene expression by binding to specific short DNA sequences, known as TF binding sites. This recognition of short sequences is governed by TF binding specificity. The differences in their DNA binding specificities enable different TFs to identify and bind to selected DNA sequences with different affinities. These binding events can result in the activation or repression of the genes regulated by the TFs. Therefore, deciphering the binding specificity of TFs is important to predict where

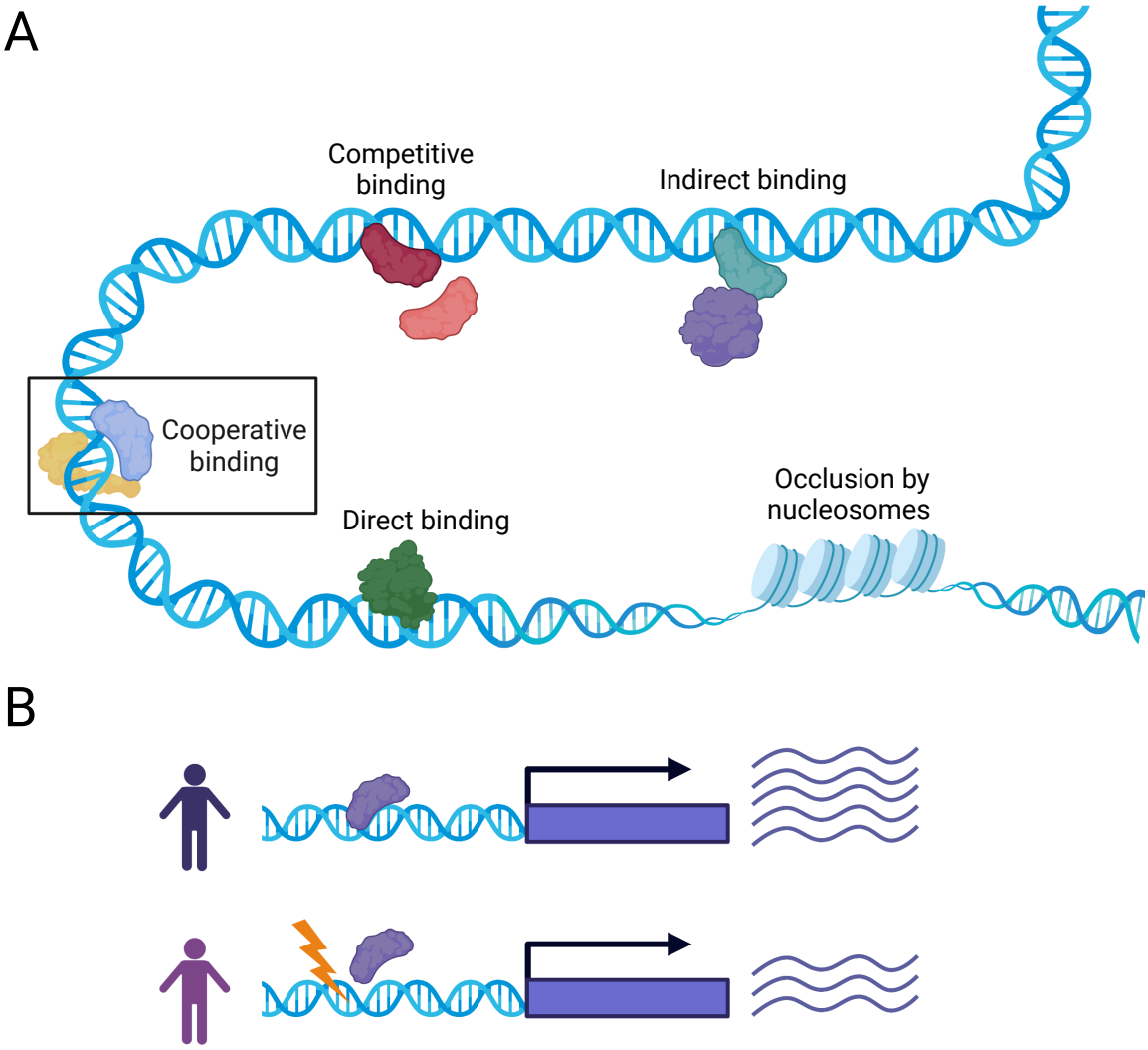



**Figure 1.2**: **Factors modulating TF-DNA interactions** (A) Different mechanistic categories by which TFs recognize their cognate sites *in vivo*. In this study, we focus on the cooperative binding between TFs. (B) Genetic mutations can alter TF binding to DNA and affect gene regulation. Created with BioRender.com.

they bind in the genome and which genes they regulate. This can further be used to understand the gene regulatory networks and infer how TFs contribute to the determination of cell fate.

The TF-DNA recognition involves two distinct modes: the recognition of specific base-pairs (base readout), as well as recognition of specific DNA structural characteristics (shape readout) (Rohs et al., 2010; Siggers and Gordân, 2014). Given

that structural analyses are time consuming and data intensive, the DNA shape features used in models of TF-DNA binding are typically derived from the sequence of the DNA (Zhou et al., 2013), and combined with direct sequence information (i.e. the identity of the DNA bases) to generate "motif" models that can predict a TF's specificity to any sequence.

The most commonly used model to represent TF-DNA binding specificity is the position weight matrix (PWM), which specifies the probability of seeing each nucleotide at each position within a TF's binding site. PWMs are simple, easy to train, and also easy to visualize as motif logos. However, they assume that each position in a TF binding site is independent of the other positions, an assumption that is not always true.

Subsequently, sequence-based models that use more complex features than PWM have been developed (Slattery et al., 2014). These models usually take into account dependencies between positions within the binding sites, for example by using k-mer features (i.e. nucleotide substring of length $k$; e.g. Siddharthan, 2010; Annala et al., 2011), and generally perform better than PWM alone in predicting the DNA binding specificity of TFs (Kähärä and Lähdesmäki, 2013). Furthermore, models using DNA shape features such as groove geometries and helical shapes (e.g. Yang et al., 2014) can help to uncover DNA binding specificity that are not apparent through nucleotide sequence alone. Features outside the TF binding sites also contribute to the TFs' binding specificity. Shen et al. (2018) developed k-mer based models using the DNA sequence of TF binding sites and their flanking regions to study the differences in DNA-binding specificity between TF family members, i.e. paralogous TFs. This study shows the importance of incorporating DNA flanking regions (also shown by Gordan et al. (2013)), especially to detect binding in medium to low affinity sites, which are usually indistinguishable by other DNA-binding data and models.

## 1.2 Factors modulating TF-DNA interactions

While TF-DNA binding specificity models have been able to represent *in vitro* (i.e. direct) TFs' DNA recognition, there are many factors modulating how TFs recognize their binding sites *in vivo* (Figure 1.2) (Inukai et al., 2017). Indirect binding, cooperative binding, competitive binding, and chromatin environments (Figure 1.2A) can contribute in determining the exact genomic binding locations of TFs. Furthermore, mutations in non-coding regions of the DNA can have functional effects by disrupting interactions between TFs and their genomic target sites (Figure 1.2B). These factors lead to complexities that cannot be captured by motif models alone (Chapter 1.1), presenting a need for models designed specifically to capture the influence of each factor.

This chapter provides a brief overview of each factor that modulates TF binding and then focuses on how our study contributes to a deeper understanding of cooperative binding and the effect of DNA mutations on TF binding.

### 1.2.1 Indirect TF-DNA interactions

Individual TFs can bind DNA indirectly (i.e. 'piggy-backing') through interactions with protein partners (Figure 1.2A) (Gordân et al., 2009). For example, the yeast TF Swi6 is known to bind TF Mbp1, while Mbp1 binds directly to its binding site (Taylor et al., 2000). However, the indirect TF binding is hard to distinguish from direct TF binding and can be a confounding factor in *in-vivo* experiments such as chromatin immunoprecipitation (ChIP) assay. Occasionally, peak regions found in ChiP experiments for a TF have no sequences matching its binding motifs (Gordân et al., 2009; Bailey and Machanick, 2012). However, these peak regions are often enriched with motifs for different TFs that have distinct DNA binding preferences, which sug-

gest the possibility of indirect binding. Another possible explanation is that TFs bind through cooperative binding mechanisms by utilizing low-affinity binding sites, which is often not captured by DNA motif models (Chapter 1.1). The details on cooperative binding will be explained in Chapter 1.3. Nevertheless, understanding how TFs interact with their protein partners can uncover an additional layer of complexity that TFs use to select their genomic binding sites.

### 1.2.2 Competitive binding between paralogous TFs

Paralogous TFs are TFs from the same protein family. Despite performing distinct regulatory functions, paralogous TFs have similar DNA-binding domains (DBDs) and recognize similar DNA sequence motifs. One solution to this paradox is through the differences in their intrinsic DNA-binding specificities, which results in different DNA binding preferences in medium to low-affinity sites (Shen et al., 2018).

However, because of their similar motifs, paralogous TFs can still compete for binding sites. Zhang et al. (2021) showed that competitive binding between yeast TF Cbf1 and Pho4 determine their *in vitro* binding preferences and the TFs' binding patterns are also reflected *in vivo*. They further confirmed that competitive binding between Cbf1 and Pho4 determines their genomic occupancy in the cell and contributes to the differential regulation of Pho4 target genes genome-wide. This shows the importance of taking competition into account when analyzing *in vivo* binding data for paralogous TFs, such as obtained from ChIP-based assays.

### 1.2.3 TF binding in the presence of nucleosomes

Eukaryotic genomes are packaged into chromatin, which consists of nucleosome subunits where DNA is wrapped around the core histone proteins. The DNA packaging

enables precise control of gene regulation by the closing and opening of the chromatin. A large fraction of the eukaryotic genome is closed since most DNA sequences are wrapped around the nucleosomes, occluding the binding of most TFs to their target DNA sites.

The fraction of TFs that are able to bind nucleosomal DNA in closed chromatin environment are called 'pioneer TFs'. The binding of pioneer TFs often leads to the opening of the chromatin by displacement of linker histones and recruitment of chromatin remodelers (Guertin and Lis, 2013; Iwafuchi-Doi, 2019). This enables subsequent binding of other TFs to the DNA in the already-open chromatin regions. For example, the FoxA family of TFs bind to their target sites on the compacted nucleosomal DNA, which opens the DNA for liver-specific TFs to bind and stimulate gene activation (Iwafuchi-Doi et al., 2016). In the reprogramming of fibroblast into pluripotency; TF Oct4, Sox2, and Klf4 act as pioneer factors for c-Myc, which leads to the enhanced binding of Oct4, Sox2, Klf4, and c-Myc to the DNA (Soufi et al., 2012). Therefore, the presence of pioneer TFs enable precise gene regulation by opening the DNA for cell-type-specific TFs to access their target DNA sites.

TFs can also bind to nucleosomal DNA through competition with nucleosomes, i.e. "collaborative competition" (Wasson and Hartemink, 2009). This interaction occurs when TFs in high concentration compete with nucleosomes for a binding site, which leads to the opening of neighboring sites. This process is also known as "nucleosome-mediated cooperativity" (Figure 1.4C), since TFs cooperate indirectly (i.e. without any protein-protein interactions) by performing synergistic binding to the nucleosomal DNA (Mirny, 2010). For example, simultaneous binding by Gal4 TF derivatives has been shown to destabilize and displace the nucleosomes (Adams and Workman, 1995). Therefore, using collaborative competition, it is possible for TFs to bind and gain occupancy of adjacent sites in the closed chromatin region.

### 1.2.4 Other factors modulating TF-DNA interactions

The aforementioned factors show the complexity of TF-DNA interactions beyond a TF's intrinsic preferences for binding site motifs. There are other factors that can influence TF-DNA interactions. For example, TFs often utilize low-affinity binding sites in the cells, which usually are not well represented by any particular binding models (Siggers and Gordân, 2014). Low-affinity binding sites present binding flexibility for functionally relevant sites beyond consensus binding motifs, as shown for paralogous TFs (Shen et al., 2018). Other factors such as DNA regions flanking the binding sites can contribute to additional specificity, which are often exerted through DNA shape (Gordan et al., 2013). Epigenetic changes such as histone modifications and DNA methylation can also modulate TF binding with the effects being TF-specific (Xin and Rohs, 2018; Héberlé and Bardet, 2019).

In the genome, TFs also bind cooperatively with each other. Furthermore, mutations in the DNA can affect TF binding to their target sites. These two factors are the main focus of this study and the remaining portion of this chapter will focus on exploring the existing technologies and studies, followed by our contribution in studying these factors.

## 1.3 Cooperative binding of TFs to DNA

In high-level organisms with large genomes, such as multicellular eukaryotes, TFs recognize sequences that are too short to define unique genomic positions. In contrast, TF binding sites in prokaryotes are usually long enough that individual sites are sufficient to define unique locations for different genes (Mirny, 2010). For example, the average length of human TF binding sites are usually between 5-10bp while in prokaryotes such as in *E. coli*, the average length of TF binding sites is ∼24.5bp. At

first, this might seem paradoxical. Considering that eukaryotes have relatively larger genome size than prokaryotes, why are their TF binding sites relatively short? The answer to this paradox lies in the mechanistic advantages provided by short binding sites; which enable cooperative DNA recognition by binding of multiple TFs to binding sites in close proximity to each other, i.e. binding site clusters (Figure 1.3A). Thus, defining unique genomic positions by using combinations of multiple binding sites (Morgunova and Taipale, 2017).

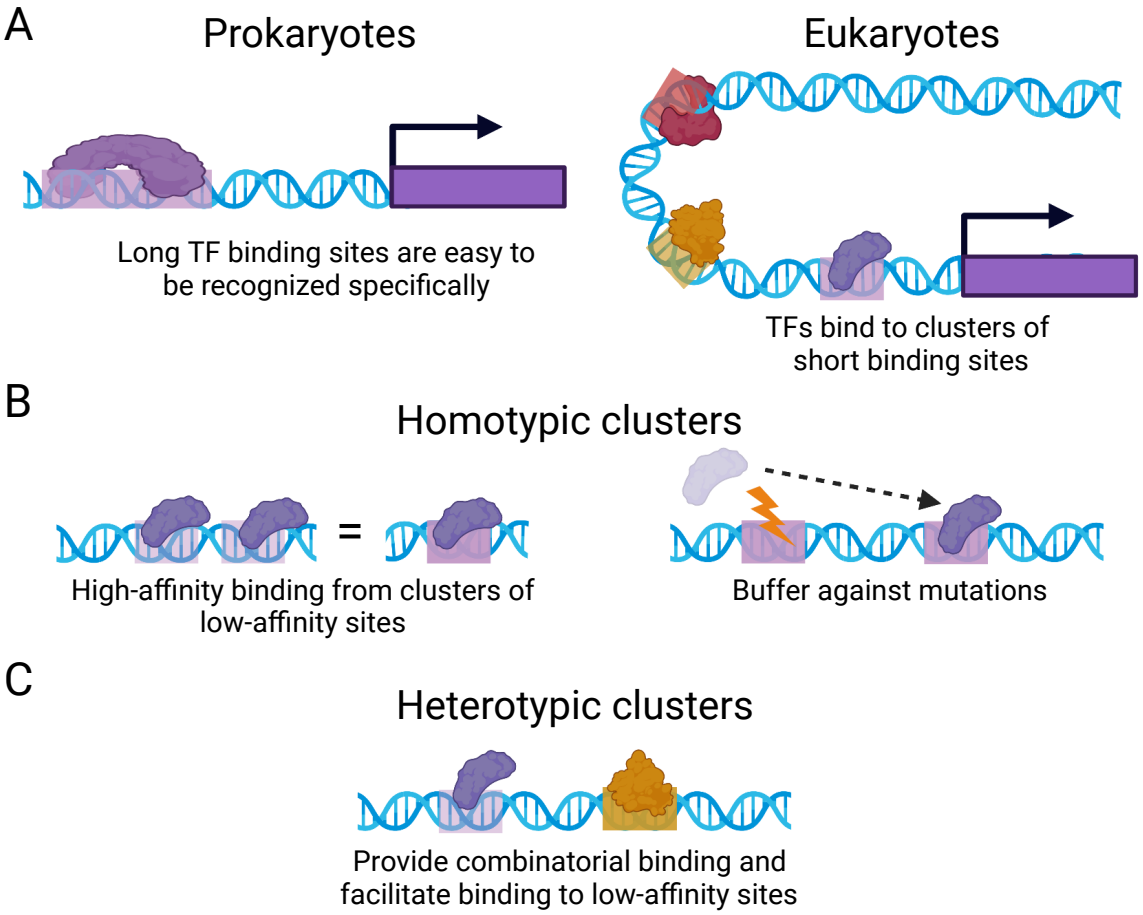



**Figure 1.3**: **Clusters of TF binding sites** (A) Left: long binding sites, which are common in prokaryotic genomes, are easy to be recognized by the corresponding TFs. Right: in eukaryotes, binding sites are usually short, and thus combinatorial binding of TFs to DNA facilitates the selection of unique binding targets in the genome. (B,C) The advantage of homotypic and heterotypic clusters, respectively. The opacity of the binding sites represents their affinity. Created with BioRender.com.

### 1.3.1 Clusters of TF binding sites

A cluster of TF binding sites is a region of DNA which consists of multiple binding sites located in close proximity to each other. There are two categories of binding site clusters. When the binding sites are recognized by the same TF, they are said to form a "homotypic cluster". When recognized by different TFs, they are said to form a "heterotypic cluster".

#### Homotypic clusters

Homotypic clusters are a pervasive feature in various organisms, epsecially in eukaryotic genomes; and they play important roles in gene regulation (Lifanov et al., 2003; Gotea et al., 2010; Ezer et al., 2014). They occur in more than 50% of human promoters and are especially abundant in developmental enhancers (Gotea et al., 2010). It has been shown that in *Drosophila*, the presence of homotypic clusters correlates with functional developmental enhancers (Lifanov et al., 2003). For example, Bicoid is one of the TFs that form the main anterior-posterior (AP) axis in the early embryo. The positions of clusters containing the Bicoid binding motif correlate with the positions of known functional enhancers in the *even-skipped* locus (Lifanov et al., 2003), which agrees with the role of Bicoid in AP patterning (Driever and Nüsslein-Volhard, 1988).

There are some mechanistic advantages provided by homotypic clustering. First, compared to heterotypic clusters, homotypic clusters have the simplicity of not having to rely on the combinatorial problem of binding from different types of TFs. Instead, they utilize combinations of DNA features such as binding site distance, orientation, and strength to modulate binding (Ezer et al., 2014). Furthermore, homotypic clusters provide a mechanism for achieving high-affinity binding regions by facilitating

binding to multiple low affinity sites. These low affinity sites are easier to create through naturally-occurring mutations than a strong affinity binding site; which is beneficial from an evolutionary perspective. Lastly, they buffer against single detrimental mutations that would abolish binding to an individual site (Lu and Rogan, 2018, also illustrated in Figure 1.3B).

**Heterotypic clusters**

The large size of eukaryotic genomes poses the need of combinatorial TF binding. Thus, heterotypic clusters provide the mean of achieving target specificity by using combinations of short TF binding motifs (Morgunova and Taipale, 2017, illustrated in Figure 1.3C). Unlike homotypic clusters, there are a myriad of possible combinations of adjacent binding sites for different TFs in the genome, making it difficult to study how TF combinations in heterotypic clusters would influence gene expression. Smith et al. (2013) uses random permutations of TF binding sites to show that different permutations can produce significantly different levels of transcription *in vivo*. On the other hand, the heterotypic TF binding patterns in the genome might be more definitive than expected. As shown by Smith et al. (2013), some TF binding site combinations are bound more favorably than others. Therefore, a study using TF pairs with known regulatory effect might reveal more relevant mechanistic insight on TF binding in heterotypic clusters compared to testing random TF combinations. It can further reveal how the difference in the DNA features used in heterotypic binding influence the TF binding. Validations using structural data, such as available from Jolma et al. (2015), can then be used to investigate the features used in heterotypic cluster binding.

### 1.3.2 Physical mechanisms of cooperative binding

The presence of adjacent binding sites in TF clusters does not imply that TFs bind cooperatively. In other words, it does not imply that the TFs influence each other's binding. To understand whether a collection of TF binding events in a cluster is cooperative, it is important to first understand the mechanisms used by TFs to bind cooperatively.

The simplest mechanism by which TFs can cooperate is through DNA facilitation (Figure 1.4A). In this case, binding to DNA enhances the interaction between two TFs by bringing the TFs closer or by altering the conformation of one TF so that the interaction with the other TF is more favorable (Morgunova and Taipale, 2017). It has been shown that in DNA-facilitated interactions, the distance and orientation of the neighboring TF binding sites is critical (Jolma et al., 2015; Morgunova and Taipale, 2017). Two TFs that can bind to DNA in multiple orientations can have different preferred distances for each orientation. Furthermore, when two TFs bind closely, this interaction can change the DNA binding motifs possibly in the positions that are recognized by weak bonds such as water mediated hydrogen bonds or DNA backbone contacts.

TF cooperativity can also be mediated entirely through DNA, without a physical interaction between TFs (Figure 1.4B). The binding of one TF can cause favorable allosteric changes to the shape of the DNA, leading to increased TF binding affinity at a neighboring site (Morgunova and Taipale, 2017). In this case, the TFs do not interact directly with each other; instead, the increase in affinity is caused by the allosteric changes in the DNA. For DNA-mediated cooperativity, the linker sequence (i.e. the DNA sequence between the two TF binding sites) has been shown to be the key player that determines the nature of this interaction (Kim et al., 2013).

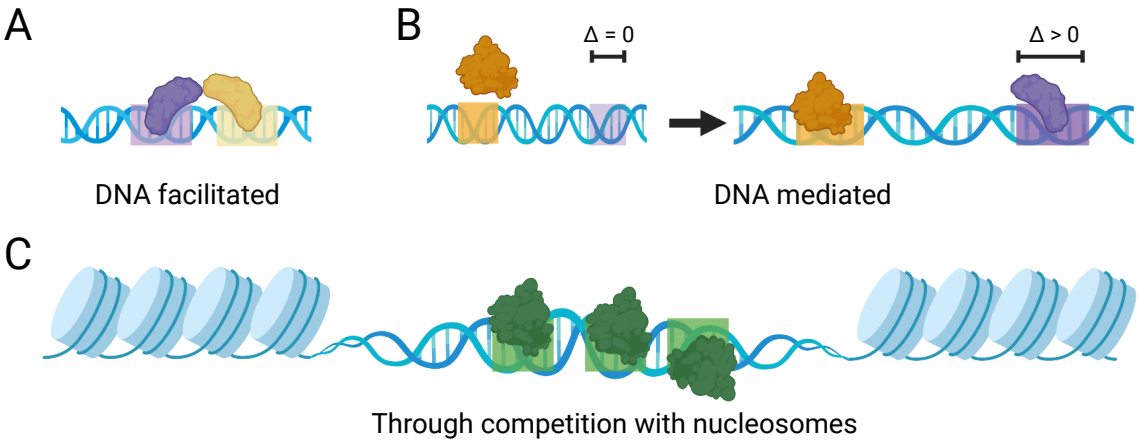



**Figure 1.4**: **Cooperative binding mechanisms** (A) The presence of DNA can facilitate some TF combinations to form highly specific complexes. (B) In the DNA-mediated cooperative binding, a TF binding can cause allosteric changes in the DNA, for example by widening the DNA which could energetically favor other TF binding at positions that are already widened. (C) Indirect cooperative binding through competition with nucleosomes, which can happen at a relatively long range. Created with BioRender.com.

Therefore, features such as linker GC content or ratios between different combinations of mono/di/tri-nucleotides might be relevant, especially on the positions that are close to the binding sites of the TFs.

Another possible mechanism of cooperative binding is through competition with nucleosomes (i.e. collaborative competition, Figure 1.4C) as explained in Chapter 1.2.3. However, this mechanism will not be discussed further as this study focuses on the intrinsic cooperativity of TFs to neighboring binding sites, i.e. the cooperativity that is due solely to the TFs and the DNA sequence, without influences from other cellular factors such as nucleosomes.

### 1.3.3 Technologies for identifying cooperative binding

The first step in identifying cooperative TF binding is to obtain adjacent TF binding sites. However, the presence of TF binding sites in close proximity does not

imply cooperative binding as each binding site can be bound independently. Therefore, subsequent analyses are needed to differentiate between cooperatively versus independently bound regions.

In this section, we review current approaches in identifying cooperative TF binding. We explain how each approach identifies TF binding site clusters and differentiates between cooperative from independent binding. Then, we elaborate on the merits and shortcomings of each approach.

### ChIP-seq based approach

Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) is a technology to detect protein-DNA interactions *in vivo*. In the ChIP experiment, antibodies specific to the protein of interest are used to immunoprecipitate DNA fragments that are bound by the protein (Park, 2009). These DNA fragments are then sequenced and mapped to the genome. The location of the binding sites are denoted as peaks obtained by overlapping multiple reads.

Binding site locations from ChIP-seq can be utilized to obtain genomic regions containing putative TF binding sites in close proximity, for example through motif discovery procedures (Levitsky et al., 2019). To differentiate between cooperative versus independent binding between a pair of different TFs, Datta et al. (2018) performed a knock-out based analyses by utilizing three different sets of ChIP-seq experiments. First, two ChiP-seq experiments to determine the binding locations of each TF from the pair of interest, which consists of the target and the partner TFs. Then the third ChIP-seq containing the binding data of the target TF after the partner TF is knocked-out. A binding location is determined to be cooperatively bound if the knock-out results in the loss or lower target TF's peak intensity.

This approach has the benefit of capturing cooperative binding using only ChIP-

seq peak intensities without relying on the DNA sequences. However, knock-out TF data is not easy to obtain and knock-out can lead to a wide range of effects. Datta et al. (2018) is aware of the difficulty in obtaining knock-out TF data and they also present an Expectation-Maximization based method to compute the probability that a location is cooperatively bound by both TFs using only peak intensity of the two TFs as input (i.e. without the need of the knock-out ChIP-seq data). Nevertheless, another problem with this approach is that, ChIP-seq data consists of the average between million cells and thus the presence of adjacent peaks does not imply that the binding of the two TFs happen in the same cell. Furthermore, this approach does not detect any particular mechanisms of cooperative binding and could not differentiate indirect binding from cooperative binding. Therefore, additional analyses to detect indirect TF binding might be necessary (Gordân et al., 2009; Bailey and Machanick, 2012). Furthermore, this approach only works for heterotypic clusters and will not apply to homotypic clusters.

**SELEX-based approach**

Systematic evolution of ligands by exponential enrichment, i.e. SELEX (Tuerk and Gold, 1990) is an *in vitro* technique to detect either single-stranded DNA or RNA interactions with ligands. A modified SELEX approach has been developed to specifically analyze the DNA binding specificity of a large number of TFs in parallel, known as high-throughput SELEX, i.e. HT-SELEX (Jolma et al., 2010). Briefly, a library of double-stranded DNA (dsDNA) molecules is incubated with immobilized TF. After washing and elution, the bound dsDNA molecules are amplified with PCR and used for the subsequent selection cycles. After the last round of selection, the bound dsDNA population is subjected to high-throughput single molecule sequencing to infer specificity. Furthermore, HT-SELEX can be designed to have multiple binding

sites for the same TF and thus, can be used to study cooperative binding in homotypic clusters by comparing TF binding in the sequences with two sites to individual binding of the TF to each site.

CAP-SELEX (Systematic Evolution of Ligands by Exponential Enrichment coupled to Consecutive Affinity Purification) is a SELEX-based method to identify the interactions between two different TFs to specific DNA sequences (Jolma et al., 2015). In the CAP-SELEX experiments, the TF pair of interest are incubated with a library of random DNA sequences. The bound DNA sequences are subject to similar selection cycles with HT-SELEX, which is followed by sequencing to generate DNA binding motifs obtained from the binding of the pair, i.e. composite motifs. However, composite motifs obtained from CAP-SELEX only show the co-binding of the two TFs but do not state whether the binding is cooperative or independent.

Ibarra et al. (2020) developed a method that combines CAP-SELEX and HT-SELEX experiments to identify cooperative binding between two different TFs. They defined cooperative binding events based on the relative affinity of the TF pair (derived from CAP-SELEX) versus the mean relative affinity of the two individual TFs (derived from HT-SELEX data for each TF). They developed a regression model using DNA features as input and revealed the importance of DNA shape features for cooperativity between Forkhead-Ets TF pairs. However, both CAP-SELEX and HT-SELEX use exponential enrichment and thus cannot accurately capture the affinities in the medium to low affinity binding sites which have been shown to be important for the cooperative TF binding (Crocker et al., 2015; Kribelbauer et al., 2019).

### 1.3.4 Our contribution: Cooperative gcPBM

To model cooperative TF binding to both homotypic and heterotypic clusters we need accurate, quantitative binding data for sequences with multiple binding sites for the TF of interest. Such data is not widely available, as most assays are designed to measure single TF binding events. Even assays that are designed to detect cooperative binding are usually not sensitive enough to capture low-affinity binding sites or are done *in vivo*, where they are subject to other factors influencing TF binding in the cells (as explained in Chapter 1.3.3).

To measure cooperative binding, we used a modification of the genomic-context protein-binding microarray (gcPBM) assay developed in our lab (Gordan et al., 2013; Shen et al., 2018), which we termed as *cooperative gcPBM*. Since the experiments are done in controlled *in vitro* environments, our assays are free of other *in vivo* factors influencing TF-DNA interactions (Chapter 1.2). Furthermore, we selected sequences containing neighboring sites with various distances and orientations that are bound *in vivo*, thus reflecting the actual sites that the TFs encounter in the cell. Instead of using enrichment, this unbiased selection of binding site clusters enables us to capture medium- to low-affinity binding sites that can be important for cooperative binding.

To differentiate between cooperative versus independently bound sequences, we compared the binding intensity of the sequences when they are bound by both TFs versus when they are bound by each TF individually, for both homotypic and heterotypic clusters. Then, we used this data to train machine learning models that predict whether the binding to a sequence is cooperative or independent, using features encoded in the DNA. Using these models, we are able to decipher distinct rules that govern the cooperative binding of the TF ETS1 in homotypic and heterotypic clusters. Details about the cooperative gcPBM and how we used it to decipher ETS1

cooperative binding are explained in Chapter 2.

## 1.4 Quantifying the effect of DNA mutations on TF binding

Genetic mutations in protein-coding regions can have important functional impacts and thus have received a lot of attention. However, numerous research studies have also shown the importance of mutations in non-coding portions of the genome in various complex diseases (Khurana et al., 2016; French and Edwards, 2020). It has been shown that most human genetic mutations (∼93%) occur in non-coding sequences (Maurano and others., 2012), where they can play a functional role by disrupting interactions between transcription factors (TFs) and their binding sites. However, this functional impact is difficult to assess and quantify.

Our study aims to quantify the impact of non-coding DNA mutations on TF binding and assess the impact of these mutations in complex diseases such as cancer. It is important to note that non-coding mutations can have other functional impacts such as alteration of RNA splicing, UTRs, or non-coding RNA (French and Edwards, 2020); which are outside the scope of this study.

### 1.4.1 Role of non-coding mutations in diseases

*Cis*-regulatory elements such as promoters and enhancers are scattered around the non-coding regions of the genome and they have diverse roles in the regulation of protein-coding genes. Mutations in these regions can cause dysregulation of gene expression and lead to various diseases.

Promoter sequences are typically located around the transcription start sites and

recruit TFs and RNA polymerase II to initiate transcription (Haberle and Stark, 2018). Mutations in promoter sequences have been linked to various diseases. For example, mutations in the OVOL2 promoter sequence have been shown to create binding sites for various TFs and a major cause of dominant corneal endothelialdystrophies in over 100 affected individuals (Davidson et al., 2016). Mutations in the Cx32 promoter can result in the loss of activity of the TF SOX10 and has been associated to X-linked form of Charcot–Marie–Tooth (CMTX) disease (Bondurand et al., 2001).

In cancer, mutations in the TERT promoter regions are highly recurrent and act as driver mutations in various cancer types (Bell et al., 2016). The first discovery of TERT promoter mutations was in human melanoma (Huang et al., 2013; Horn et al., 2013). These mutations can create binding sites for the ETS family of transcription factors and drive the expression of the TERT (telomerase reverse transcriptase) gene, which encodes the catalytic reverse transcriptase subunit of telomerase. Telomerase lengthens telomeres, allowing cells to escape apoptosis and become cancerous. TERT expression is generally repressed in normal somatic cells, but they can be overexpressed in cancer.

Mutations in distal regulatory elements such as enhancers can lead to dysregulation of gene expression since TF binding to these elements can either enhance or silence the transcription of the genes they regulate. In the IRF6 enhancer, mutations can abrogate the binding of TFs p63 and E47; and have been linked to Van der Woude syndrome (VWS). Furthermore, recurrent non-coding mutations have been identified within the enhancer of TAL1 gene in T-cell acute lymphoblastic leukemia (T-ALL), creating binding sites for the TF MYB and forming a super-enhancer upstream of TAL1, which results in its overexpression (Mansour et al., 2014).

With the various effects that non-coding mutations can have in diseases, it is

important to identify functional non-coding mutations versus mutations that occur by chance. Public dataset are now available, for example through genome-wide association study (GWAS) which associate specific genetic variants with particular diseases. Such studies are conducted by comparing genomic DNA from two groups of participants: those with and without the disease being studied. The result gives correlations but not causation) between the associated genetic variations and the disease. For cancer associated variants, the Pan-Cancer Analysis of Whole Genomes (PCAWG) provides mutation dataset from over 2,600 cancer whole genomes, enabling us to systematically survey non-coding regions for potential driver events. As more non-coding mutation datasets become available, we can expect to see new discovery of disease-causing mutations in the non-coding regulatory regions.

### 1.4.2 Computational methods for quantifying the effects of DNA mutations on TF binding

PWMs provide a score reflecting the likelihood of TF binding which can be used to score TF binding changes from mutations, such as used by Wagih et al. (2018). Given a mutation at position $q$, they first took all PWM scores using sliding window of length $k$ from position $q - k + 1$ to $q + k - 1$ (i.e. all windows should overlap $q$), so we have $k$ windows of PWM scores for both wild type ($w$) and mutant ($t$) sequences. The binding change score is calculated using the difference between the maximum PWM scores of $t$ and $w$. More formally, given the set of indices $S = 1, .., k$ for each sliding window position, the binding change score for TF $T$ from a mutation in position $q$ can be calculated as:

$$\Delta_{T,q} = \max_{\forall i \in S} t_i - \max_{\forall i \in S} w_i \tag{1.1}$$

Vanilla PWM-based models usually require a threshold to call a position as a binding site; and thus, might not be able to capture low-affinity binding sites. Improvement over PWM models such as the PWM-based TRAP model (Thomas-Chollier et al., 2011) can circumvent this limitation by using the total of the TF binding affinity values from each position, represented as p-values. To study the effect of mutations on TF binding, a variation of TRAP called sTRAP can by used. It compares the affinity values between a wild-type and a mutated sequence, resulting in the change in TF binding affinity calculated as the difference in $log(p)$ between mutant and wild-type sequences.

Models using more complex techniques than PWM can more accurately captures the sequence specificities. DeepBind (Alipanahi et al., 2015) is a deep learning models to predict TFs' sequence specificities and can be trained using sequence data from a wide range of assays such as PBM, SELEX, and ChIP-seq. Similar to PWM, DeepBind outputs a likelihood score representing the TF binding. Thus, to calculate the TF binding change, the same method shown in equation 1.1 can be used.

Both PWM and DeepBind models use the likelihood scores to train and predict the TF binding change, so they might not directly represent the actual binding affinity of the TF. The sTRAP model circumvents this drawback using its affinity-based PWM method. However, PWM-based model makes assumption that each position in the binding site is independent, an assumption that is not always true (Badis et al., 2009; Bulyk et al., 2002; Tomovic and Oakeley, 2007). Complex models such as DeepBind can better capture the dependencies between positions, but deep-learning models can be expensive to compute. Another drawback from both PWM and DeepBind models is, one TF can have multiple models available. Different models can result in different predictions and there is no definitive way to choose the best model to use, as quality metrics are not always reported for these models. Ideally, a method for

characterizing the effect of mutations on TF binding should be able to capture both large and small changes in binding, as long as the changes are 'significant' given the quality/precision of the model.

### 1.4.3 Our Contribution: QBiC-Pred

We introduce QBiC-Pred (Quantitative Predictions of TF Binding Changes Due to Sequence Variants) (Martin et al., 2019), or QBiC for short, a new method to assess the impact of non-coding mutations on TF-DNA binding (Figure 1.5). QBiC is trained using an ordinary least squares (OLS) method from our previous study (Zhao et al., 2017). The OLS method uses *in vitro* TF binding data from universal protein-binding microarray (uPBM) experiments (Berger and Bulyk, 2009) to train regression models of TF-DNA binding specificity using OLS estimation.

QBiC models are trained directly from actual TF binding measurements obtained from uPBM, which contains both the wild type and the mutant sequences. Thus, QBiC results directly represent the binding change obtained from experiments, in contrast to the methods that use the likelihood of TF binding (Chapter 1.4.2). Furthermore, in our OLS-based approach, we are able to rank the models based on cross-validation accuracy on the uPBM training dataset. Thus, when a TF has multiple OLS models, we can use the model with the highest cross-validation accuracy. In addition, using OLS, we have a way to statistically test the significance of each variant, taking the quality of the predictive models into account. Thus, instead of just reporting the magnitude of the effect, QBiC also outputs the significance of the predictions (i.e. p-value). Moreover, OLS is simple and much faster for training and prediction, in comparison to the deep-learning based models.

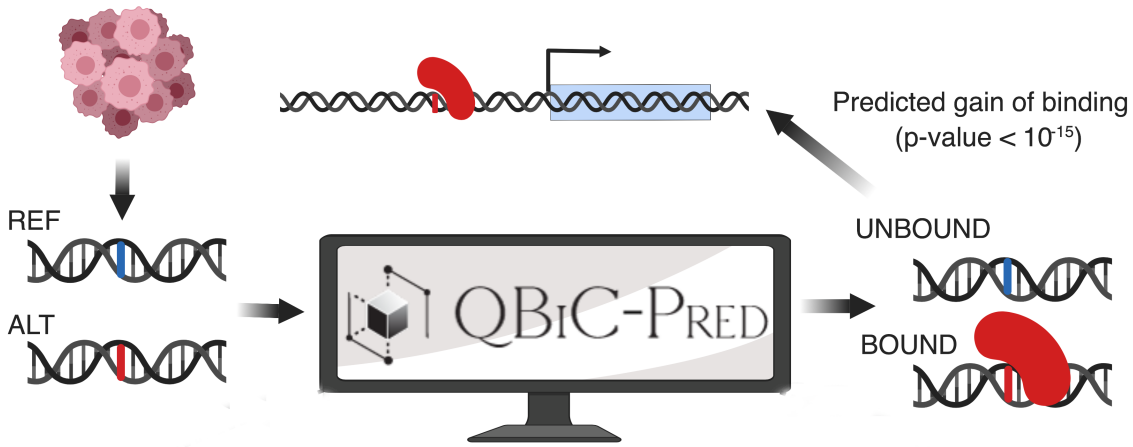



**Figure 1.5**: **QBiC-Pred workflow.** For each input mutation, QBiC uses the reference sequence (i.e. wild type) and the alternate sequence (i.e. mutant) as input to the OLS model which computes the change in TF binding as a linear combination of the coefficient estimates for all 6-mers overlapping the variant. Finally, QBiC outputs the predicted binding change along with its confidence score.

We implemented QBiC as a web service, which allows users to run our OLS models through a user-friendly web interface: qbic.genome.duke.edu. Details on the web server and the accuracy of QBiC in comparison to other methods are available in Chapter 3. We also used QBiC to identify non-recurring putative regulatory drivers in cancer, which is presented in Chapter 4.

# Chapter 2

# High-throughput data and modeling reveal insights into the mechanisms of cooperative DNA-binding by transcription factors

## 2.1 Introduction

Transcription factors bind to DNA in a sequence-specific manner, a key process in the regulation of gene expression. In human cells, TF binding sites are oftentimes present in clusters, i.e. multiple sites located in close proximity to each other (Ezer et al., 2014; Gotea et al., 2010). Clusters of binding sites recognized by the same TF (i.e. homotypic clusters) occur in more than 50% of human promoters and are especially abundant in developmental enhancers (Lifanov et al., 2003; Gotea et al., 2010; Ezer et al., 2014). During evolution, clustering of TF binding sites provides a mechanism for achieving high-affinity binding regions while buffering against single detrimental mutations that would completely abolish binding (Lu and Rogan, 2018). Clusters of sites bound by different TFs (i.e. heterotypic clusters) are critical for combinatorial regulation, which is especially important in complex multicellular organisms (Morgunova and Taipale, 2017). In addition, given that most eukaryotic TFs have short motifs that are insufficient to specify unique target sites across the genome, heterotypic clustering of binding sites helps increase target specificity by

using combinations of short motifs (Morgunova and Taipale, 2017).

Mechanistically, binding sites located in close proximity to each other allow TFs to bind cooperatively, thus stabilizing the binding and enhancing each TF's contribution to gene regulation (Siggers and Gordân, 2014). Sites that are likely bound cooperatively *in vivo* are oftentimes bound weakly when only individual TFs are present in the cell (Datta et al., 2018). Thus, cooperative interactions serve to enhance the overall DNA-binding affinity of the TF-TF complex and enable the utilization of low-affinity binding sites in gene regulation. Moreover, cooperativity has been used to explain the rapid rate of evolution of TF binding sites in multicellular organisms (Tugrul et al., 2015), by facilitating the emergence of shorter sites with low to medium affinity. This shows that cooperativity is an indispensable aspect in the regulation of gene expression in higher organisms.

Characterizing the cooperative DNA-binding patterns of TFs is not trivial. Even when two binding sites are located in close proximity in the genome, cooperative binding is not necessarily implied, as TFs can also bind independently to each of their target sites. Recently, high-throughput assays based on Systematic Evolution of Ligands by Exponential Enrichment (SELEX) have been used, on a large scale, to identify composite motifs that reflect orientation and spacing preferences for cooperative TF binding at neighboring sites (Jolma et al., 2013, 2015). In these studies, at most three composite motifs were identified per TF (Jolma et al., 2013) or pair of TFs (Jolma et al., 2015), leaving open the question: at spacings and orientations for which composite motifs were not identified by SELEX-based methods, can cooperative binding events still occur? In addition, even for the preferred spacing/orientation between neighboring sites, are all binding events cooperative, or can TFs also bind independently? Finally, what features other than spacing and orientation are important for cooperative TF binding?

To address these questions, we developed an experimental approach to distinguish between cooperative and independent binding of TFs to neighboring DNA sites in high throughput. We used on-chip assays based on the genomic-context protein-binding microarray (gcPBM) technology (Gordan et al., 2013; Shen et al., 2018), but with DNA libraries designed specifically to assess cooperative TF binding at binding sites located in close proximity in the genome. To develop and evaluate our approach, we used as a case study human TF ETS1, which is known to bind cooperatively with itself as well as with other TFs (Lamber et al., 2008; Fitzsimmons et al., 1996; Sieweke et al., 1997, 1998; Bradford et al., 1997; Gégonne et al., 1993; Wasylyk et al., 1990; Garvie et al., 2002; Choy et al., 2014; Baillat et al., 2002; Babayeva et al., 2010), including RUNX1 (also known as CBF$\alpha$2, AML1, or PEBP2) (Wotton et al., 1994; Kim et al., 1999; Goetz et al., 2000; Gu et al., 2000; Shrivastava et al., 2014; Shiina et al., 2015). Studying both ETS1-ETS1 and ETS1-RUNX1 binding to DNA enabled us to compare mechanisms of ETS1 cooperative binding with different partners. In addition, existing in-depth knowledge on ETS1 cooperativity, from low-throughput structural, biochemical and molecular biology studies, allowed us to validate the findings from our high-throughput approach.

## 2.2 Materials and methods

### 2.2.1 Identifying neighboring binding sites across the genome

We used chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) data from ENCODE (Moore et al., 2020) and Cistrome (Liu et al., 2011) to identify genomic regions containing putative ETS1 binding sites in close proximity (4-24 bp) to other ETS1 sites or to RUNX1 binding sites, where the TFs might

bind cooperatively (Figure 2.1A). For each available ETS1 ChIP-seq data set, we scanned the peaks using position weight matrix (PWM) models and *in vitro* binding data, as detailed below, to identify potential binding sites. As expected, we found that peaks containing multiple binding sites typically have higher ChIP-seq signal than peaks with single binding sites (Figure 2.2). When comparing the ChIP-seq peaks of ETS1 and RUNX1, we found that RUNX1 peaks were oftentimes in close proximity to ETS1 ChIP-seq peaks (Figure 2.3) consistent with cooperative binding between the two TFs (Shrivastava et al., 2014; Gu et al., 2000). We used the ETS1 peaks that overlap RUNX1 peaks to identify potential sites of ETS1-RUNX1 cooperativity (Figure 2.1A). Importantly, in this study we focused on genomic rather than artificial DNA sequences because genomic regions represent sequences that the TFs encounter in the cell. Thus, we expect that the relevant determinants for cooperative vs. independent TF binding will be present in these genomic regions, and thus in our dataset.

### 2.2.2 Design of DNA library for 'cooperative gcPBM' assay

We adapted the gcPBM assay (Gordan et al., 2013) to measure cooperative and independent ETS1-ETS1 and ETS1-RUNX1 binding to thousands of genomic regions. In our original gcPBM assay, only individual TF binding sites were tested, using DNA libraries containing 36-bp genomic sequences centered on putative binding sites for the TF of interest (Gordan et al., 2013; Shen et al., 2018). Here, the 36-bp probes in our DNA library are designed to contain genomic sequences with two binding sites, selected from ChIP-seq peaks (Figure 2.1A). We focused on sites located at a distance of 4-24 bp, calculated between the centers of the binding sites. Smaller distances were not considered, as the cores of the binding sites (i.e. the regions where DNA bases make direct contacts with the proteins) would be overlapping. Distances larger than

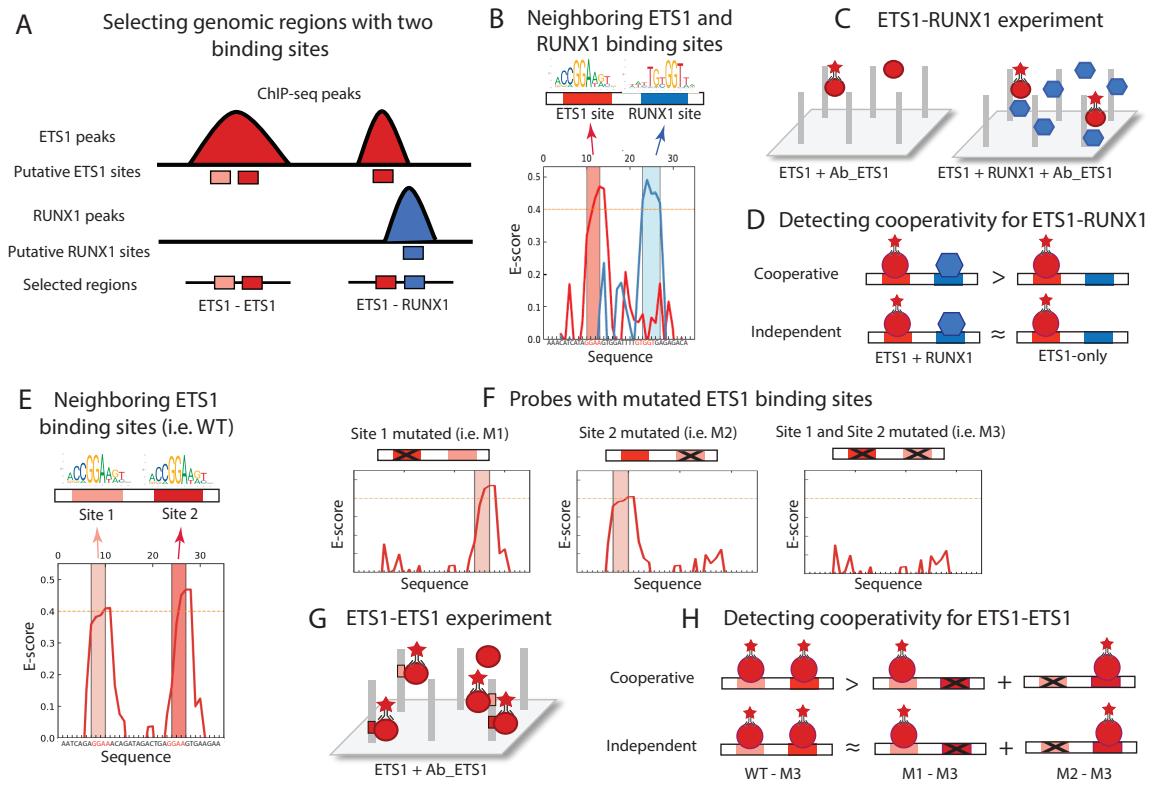



**Figure 2.1**: **Experimental approach to detect cooperative TF binding.** (A) Neighboring ETS1-ETS1 and ETS1-RUNX1 sites were identified within ChIP-seq peaks. (B) Example of a 36-bp genomic sequence with neighboring ETS1 (red) and RUNX1 (blue) sites. The sites were identified using universal PBM enrichment scores (E-scores) (Berger and Bulyk, 2009) and PWM models. PWM logos for ETS1 and RUNX1 are shown. The core base pairs in each binding site are highlighted in color. The dotted orange line denotes the E-score cutoff, 0.4, for calling specific TF binding sites. (C) Cooperative gcPBM experiments designed to detect cooperative binding events by comparing binding of ETS1 alone vs. in the presence of a high concentration of cooperator TF, RUNX1. Ab_ETS1: anti-ETS1 antibody. (D) Binding of ETS1-RUNX1 is considered cooperative when the intensity is significantly higher when both TFs are present. (E) Example of a 36-bp genomic sequence with neighboring ETS1 sites. (F) DNA probes that represent mutants of the wild type (WT) sequence with neighboring ETS1 sites from panel E. (G) Cooperative gcPBM experiments designed to detect cooperative binding events of ETS1-ETS1, based on ETS1 binding measurements for the WT, M1, M2, and M3 probes. (H) Binding of ETS1-ETS1 is considered cooperative when the intensity of the WT sequences is higher than the sum of the binding intensities to the individual sites, adjusted for non-specific binding (M3).

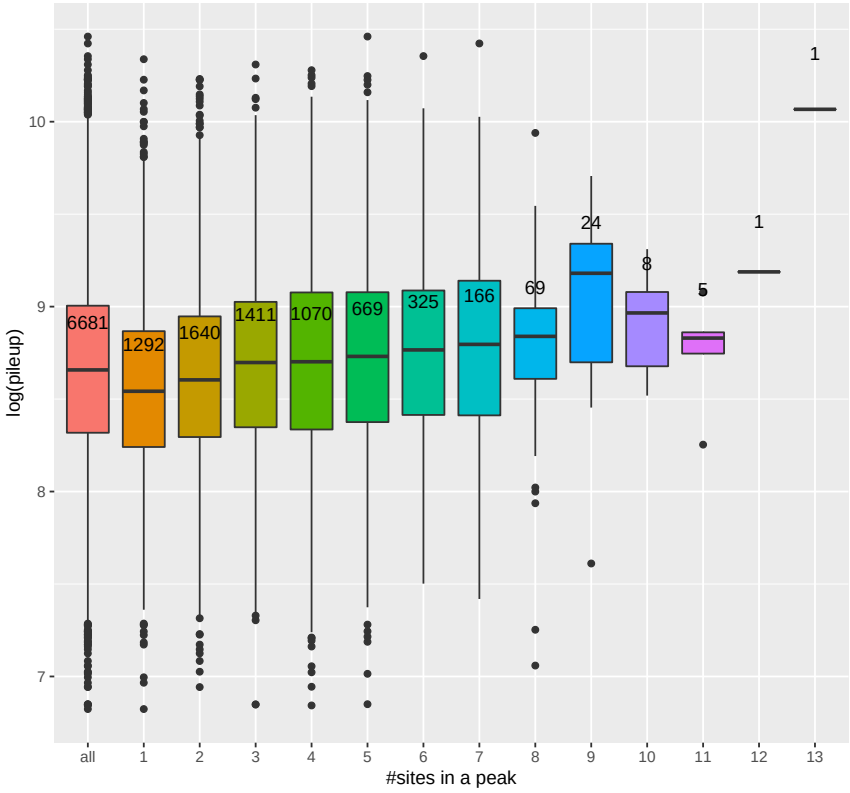



**Figure 2.2**: **ChIP-seq peaks with multiple putative Ets1 binding sites have higher ChIP-seq signal**, calculated as the natural logarithm of the peak pileup score reported by the MACS software tool (https://github.com/macs3-project/MACS). A representative plot is shown, for the Ets1 ChIP-seq data in K562 cells (ENCODE ID: ENCSR000BKQ). Boxplots show median signals, with boxes extending to the 25th and 75th percentiles. Whiskers extend to the largest/smaller values no further than 1.5 times the inter-quartile range, and the points show the most extreme data points.

24 bp could not be tested due to the length limitation for the probes in our DNA library.

Putative TF binding sites were identified using PWM models and enrichment scores (E-scores) from universal PBM experiments (Berger and Bulyk, 2009). As in previous work, we searched for two consecutive 8-mers with E-score> 0.4 (which generally indicates specific TF binding) (Gordan et al., 2013; Shen et al., 2018), and we aligned the putative binding sites using PWMs (Figure 2.1B,E). Using this

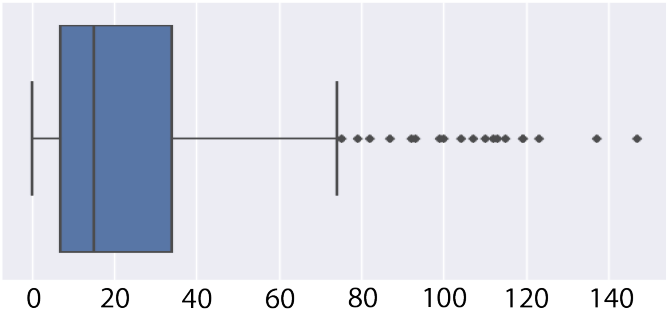



**Figure 2.3**: **Distribution of distances between ChIP-seq peak summits for Ets1 and Runx1.** For each Runx1 ChIP-seq peak summit in K562 cells (ENCODE ID: ENCSR588AKU) we calculated the distance to the nearest Ets1 peak summit (ENCODE ID: ENCSR000BKQ). Boxplot shows the median distance (17 bp), with the box extending to the 25th and 75th percentiles. Whiskers extend to the largest/smaller values no further than 1.5 times the inter-quartile range, and the points show the most extreme data points.

strategy, we identified 5093 neighboring ETS1-ETS1 sites and 3,197 ETS1-RUNX1 sites. For each region, we designed a 36-bp DNA probe centered on the middle position(s) between the two sites. As in previous work (Gordan et al., 2013; Shen et al., 2018; Zhang et al., 2021), our DNA library also included negative control probes that lacked TF binding sites (400 for ETS1-RUNX1, 257 for ETS1-ETS1).

To allow the identification of ETS1-ETS1 cooperative binding events, our DNA library also contained, for each genomic region with two neighboring ETS1 sites, probes where one or both of the sites were mutated (Figure 2.1F, Figure 2.4). Similar probes were added to the ETS1-RUNX1 library, with either the ETS1 or the RUNX1 site mutated; in this case, the mutated probes were used to normalize the binding data between different experiments, as described below.

### 2.2.3 Cooperative gcPBM assay

The designed DNA sequences were synthesized as single-stranded molecules on microarrays (Agilent) in 4x180k format, with three or four replicate spots per sequence

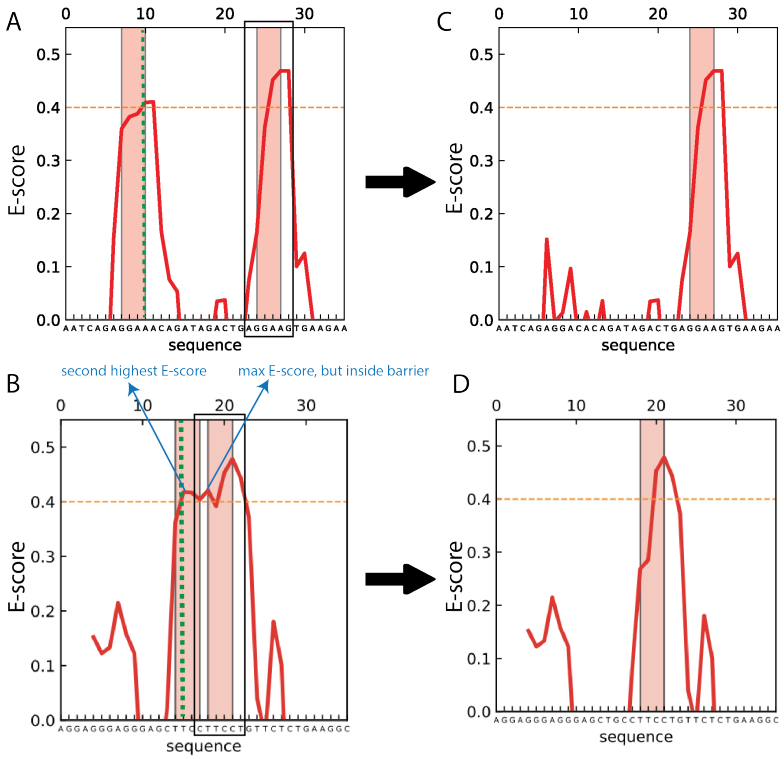



**Figure 2.4**: **Designing DNA probes with mutated Ets1 binding sites for the Ets1-Ets1 DNA library**. In order to detect cooperative binding of Ets1 to pairs of neighboring sites extracted from the human genome, for each genomic regions of interest we generated mutated sequences where one or both of the binding sites were mutated. (A,B) Example genomic regions containing two Ets1 binding sites. The cores of the binding sites (GGAA or TTCC) are highlighted using light red rectangles. The y-axes show the universal PBM E-scores for all 8-mers (red line) covering the genomic sequences shown on the x-axes, focusing on the [0,0.5] range. E-scores above 0.4 typically correspond to specific TF-DNA binding (Berger et al., 2006; Berger and Bulyk, 2009). (C,D) For the genomic regions shown in panels A and B, we illustrate the corresponding mutated probes where the first of the two binding sites was mutated so that none of the 8-mers overlapping the binding core had E-scores above 0.4. To mutate a binding site, we targeted the 8-mer with the largest E-score and we chose the single base-pair mutation that resulted in the largest decrease in E-score. The mutated positions are marked with vertical green dotted lines in panels A and B. To make sure that the neighboring binding site is not affected by the mutation, we defined a "barrier" region (i.e. positions that could not be mutated) around the binding sites, marked with a black box in panels A and B. If the position with the largest E-score fell within the barrier region (as in the example in panel B), then we chose the next largest E-score position in order to find an appropriate mutation.

for the ETS1-ETS1 and ETS1-RUNX1 neighboring sites, respectively. The arrays was double-stranded and the protein-binding steps were performed following standard PBM protocols (Berger and Bulyk, 2009). Briefly, the double-stranded DNA libraries were incubated with the proteins of interest (ETS1, RUNX1, or ETS1+RUNX1), and the level of bound protein at each spot was measured using primary antibodies specific for either ETS1 (Cell Signaling Technology 14069) or RUNX1 (Abcam ab23980) followed by Alexa488-conjugated secondary antibody (Invitrogen A-21206). Full-length, recombinant human ETS1 and RUNX1 proteins were expressed and purified as described in our previous work (Shen et al., 2018). The arrays were scanned using a GenePix 4400A microarray scanner (Molecular Devices), and the fluorescence intensity at each DNA spot was used to represent the TF binding level for the DNA sequence at that spot. As shown in previous work (Shen et al., 2018; Penvose et al., 2019; Afek et al., 2020), the PBM fluorescence intensities correlate remarkably well with independently measured equilibrium dissociation constants, and can be interpreted as the levels of bound TF at each of the sequences in our DNA library.

### 2.2.4 Identifying ETS1-RUNX1 cooperative versus independent binding events

To identify cooperative binding events at the genomic regions containing neighboring ETS1-RUNX1 sites, we first measured ETS1 binding in the presence vs. absence of saturating levels of RUNX1 (Figure 2.1C), similarly to previous work (Goetz et al., 2000). For ETS1 we used a concentration of 1nM, selected because it resulted in a broad range of binding signals on our DNA arrays (Figure 2.5). For RUNX1 we chose a high concentration (25nM) in order to ensure that, even at spots with low-affinity RUNX1 sites, a large fraction of the DNA molecules in each spot were bound.

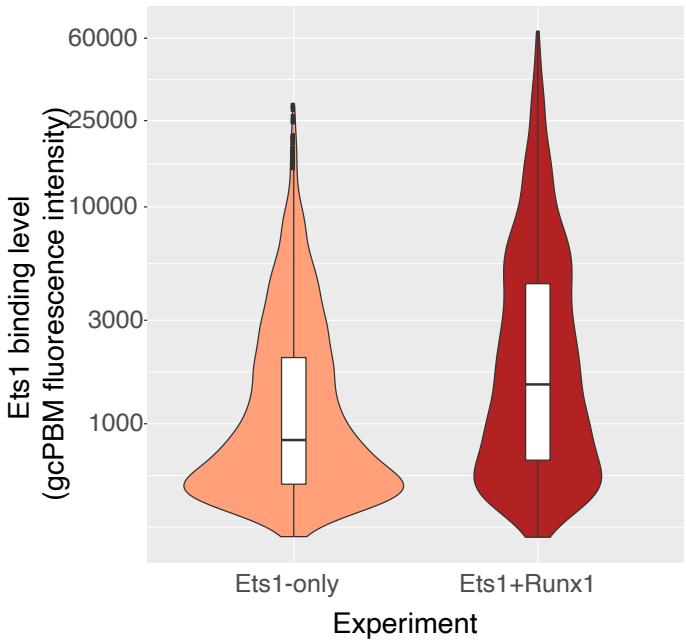



**Figure 2.5**: Violin plots showing Ets1 binding signal (gcPBM fluorescence intensity) for DNA sequences containing neighboring Ets1+Runx1 binding sites, in the Ets1-only (left) and the Ets1+Runx1 (right) experiments.

Next, we compared the ETS1 binding signals in the ETS1-only vs. the ETS1+RUNX1 experiments, for each genomic region of interest (Figure 2.1D). If the later signal was significantly higher (p-value<0.02, one-sided Mann-Whitney $U$ test applied to the four replicates spots), then we labelled the binding as 'cooperative'. If the signals from the two experiments were similar (p-value≥0.1), then we labelled the binding 'independent'. Sequences with intermediate p-values were considered ambiguous and were not included in further analyses. The p-value cutoffs were determined based on the possible values for the Mann-Whitney $U$ test between two groups of four measurements (Figure 2.6A). Before performing the statistical tests, we applied a minor normalization to the data from the ETS1+RUNX1 experiment so that, for the DNA sequences without RUNX1 binding sites, the best fit line between the ETS1-only and the ETS1+RUNX1 data sets was the diagonal (Figure 2.7).

We also removed from further consideration all genomic sequences with ETS1

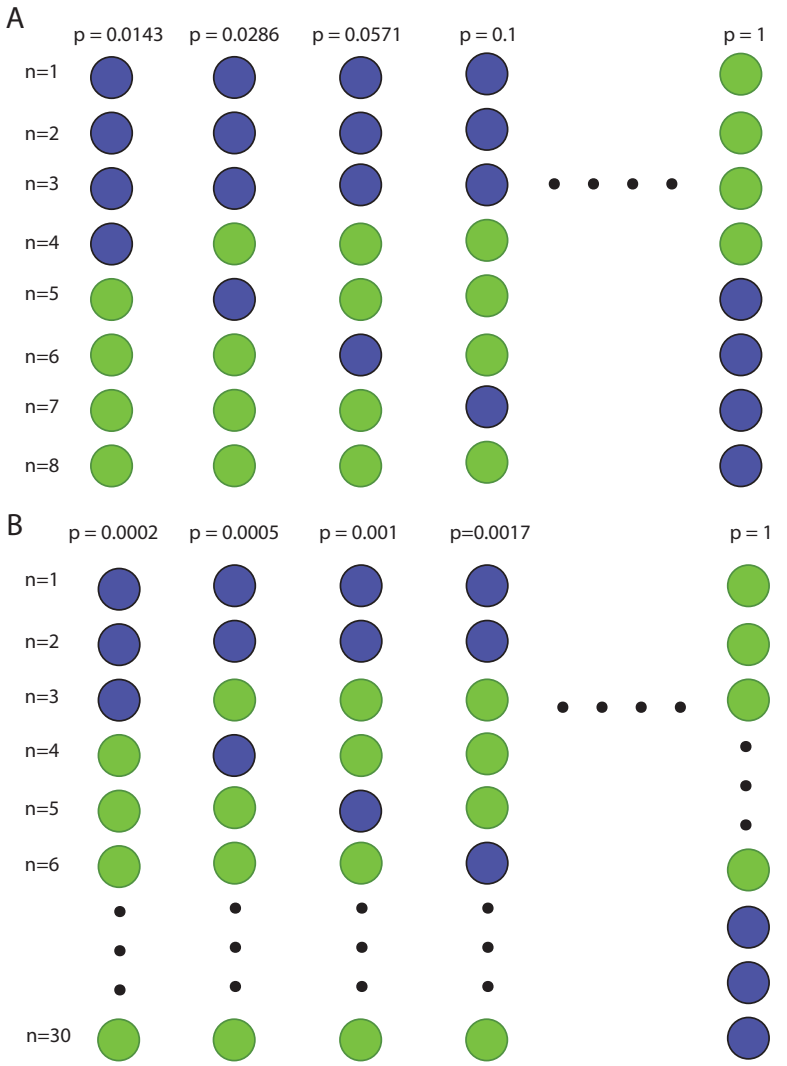



**Figure 2.6**: **Possible p-value for a one-sided Mann-Whitney $U$ test**, between two groups of 4 measurements each (A), or a group of 3 measurements compared to a group of 27 measurements (B). The two colors illustrate the two groups. The alternative hypothesis is that the blue group contains higher values than the green group. The data points (measurements) are sorted in decreasing order, with the rank numbers shown on the left. The smallest p-values (left-most) correspond to the situation when the two groups have no overlap, and all values for the blue data point are higher than all values for the green data points.

binding signals lower than the $75^{th}$ percentile of negative control probes, as we expect these sequences to be bound mostly non-specifically. Importantly, we chose the 75% rather than 100% cutoff in order to include putative ETS1 sites with very low affinity

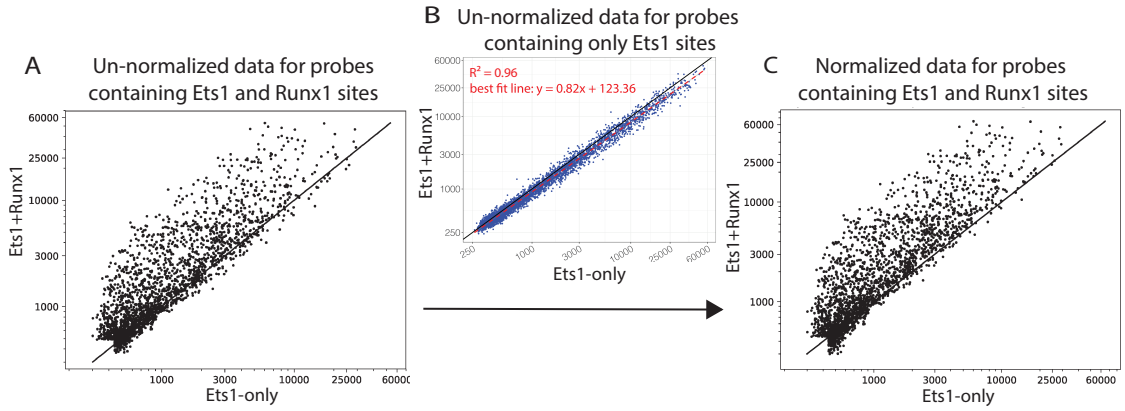



**Figure 2.7**: **Normalization of the Ets1 gcPBM data between the Ets1-only (x-axis) and the Ets1+Runx1 (y-axis) experiments**. (A) Comparison between the un-normalized Ets1 binding data from the Ets1-only versus the Ets1+Runx1 experiments. Binding levels are represented as fluorescence intensities from the two gcPBM experiments. Each data point corresponds to a genomic sequence contain one Ets1 site and one Runx1 site in close proximity. (B) Similar to panel A, but each point represents a DNA sequence containing one Ets1 site and no Runx1 site. Black line shows the diagonal. Red dotted line shows the best fit line. (C) Similar to panel A, but the y-axis shows data after normalization, which was performed by applying the function $f(y) = (y - 123.36)/0.82$.

that might benefit significantly from cooperative binding with RUNX1. After these steps, we obtained 2300 sequences from the ETS1-RUNX1 experiment, with 976 (42%) of sequences showing cooperative binding.

The strategy described above tested the effect of RUNX1, as the 'cooperator TF', on ETS1-DNA binding. We also performed the reverse test, which we refer to as the 'RUNX1-ETS1' experiment (Figure 2.8, 2.9, 2.10), by considering ETS1 as the cooperator (at 25nM concentration) and testing its effect on DNA-binding by RUNX1, at 2nM concentration.

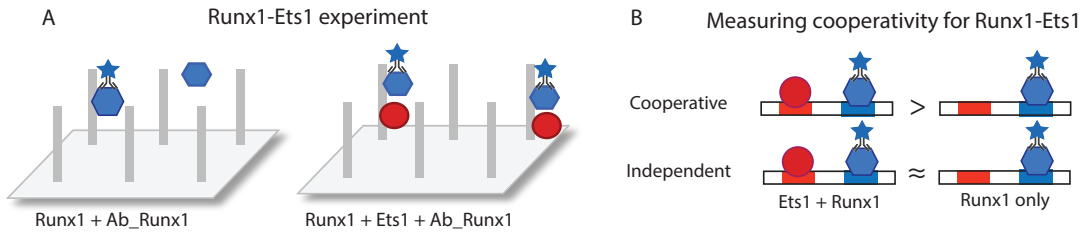



**Figure 2.8**: **Runx1-Ets1 cooperative gcPBM experiment** (A) Cooperative gcPBM experiment designed to detect cooperative binding events by comparing binding of Runx1 (blue) alone and in the presence of a high concentration of cooperator TF (Ets1, red). Ab_Runx1: anti-Runx1 antibody. (B) Binding of Runx1-Ets1 is considered cooperative when the Runx1 binding intensity is significantly higher in the presence compared to the absence of Ets1.

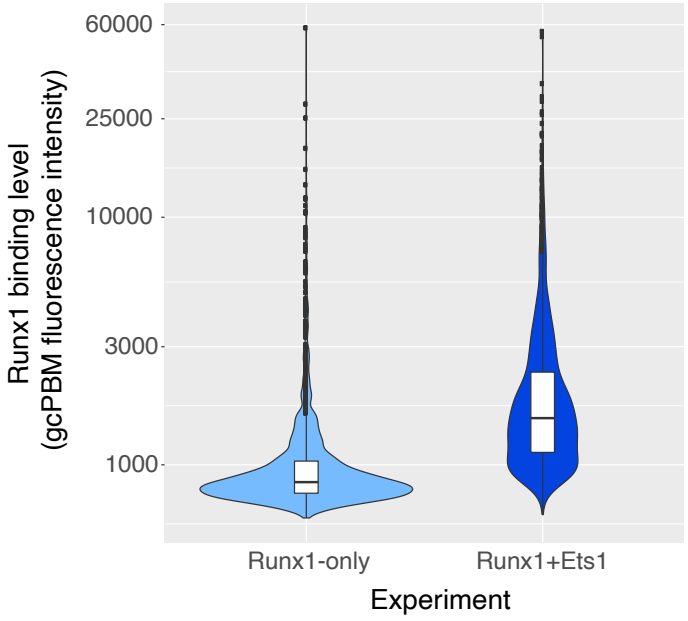



**Figure 2.9**: Violin plots showing Runx1 binding signal (gcPBM fluorescence intensity) for DNA sequences containing neighboring Ets1+Runx1 binding sites, in the Runx1-only (left) and the Runx+Ets1 (right) experiments.

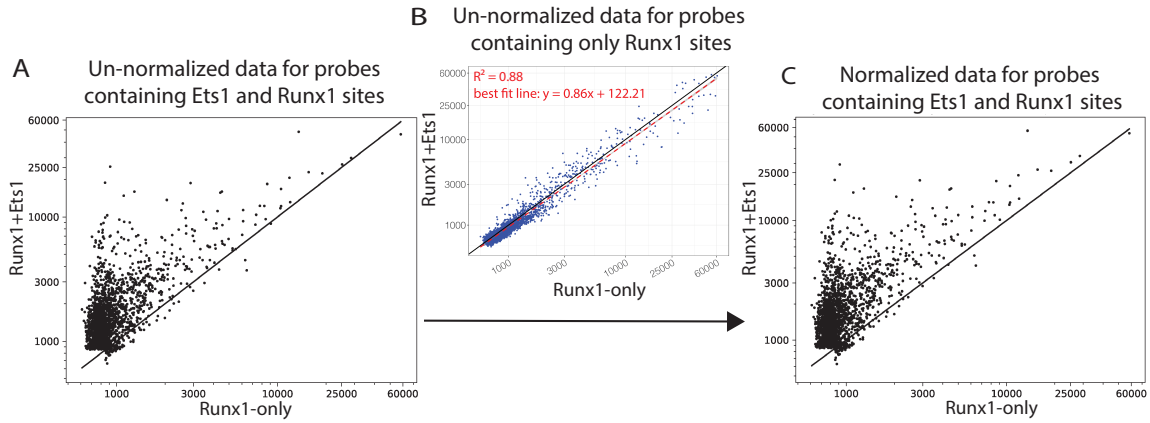



**Figure 2.10**: **Normalization of the Runx1 gcPBM data between the Runx1-only (x-axis) and the Runx1+Ets1 (y-axis) experiments**. (A) Comparison between the un-normalized Runx1 binding data from the Runx1-only versus the Runx1+Ets1 experiments. Binding levels are represented as fluorescence intensities from the two gcPBM experiments. Each data point corresponds to a genomic sequence contain one Ets1 site and one Runx1 site in close proximity. (B) Similar to panel A, but each point represents a DNA sequence containing one Runx1 site and no Ets1 site. Black line shows the diagonal. Red dotted line shows the best fit line. (C) Similar to panel A, but the y-axis shows data after normalization, which was performed by applying the function $f(y) = (y - 122.21)/0.86$.

### 2.2.5 Identifying ETS1-ETS1 cooperative versus independent binding events

To identify cooperative DNA-binding by ETS1, for each genomic region of interest that contained two ETS1 sites (e.g. Figure 2.1E), we also included in our DNA library three probes generated by mutating one or both of the binding sites, using mutations that are likely to render the binding non-specific (Figure 2.4). We refer to these probes as M1, M2, and M3 (Figure 2.1F). We performed ETS1 binding measurements (Figure 2.11) for the mutated sequences, as well the wild-type (WT) sequence, each represented in three replicate spots, at a protein concentration of 10nM.

Next, for each WT probe (i.e. genomic sequence containing two neighboring

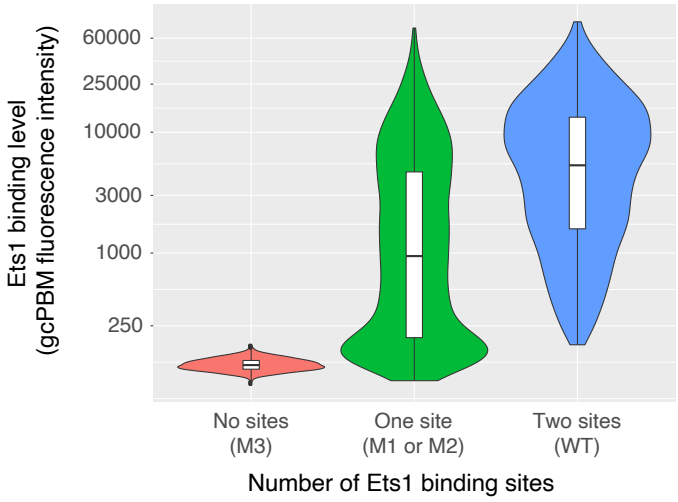



**Figure 2.11**: Violin plots showing Ets1 binding signal (gcPBM fluorescence intensity) for DNA sequences containing two neighboring Ets1 binding sites (right), a single Ets1 binding site (middle) or no Ets1 binding sites (left). See Materials and Methods for details on the design of the DNA probes (WT, M1, M2, M3).

ETS1 sites), we asked whether its protein binding level was larger than the sum of the binding levels at probes containing the individual sites (i.e. M1 and M2), adjusting for the level of non-specific binding at these probes (i.e. M3). Thus, to test for cooperative ETS1 binding at the sites in the WT probe we formed the null hypothesis $H_0 : \mu_{WT} - \mu_{M3} = \mu_{M1} - \mu_{M3} + \mu_{M2} - \mu_{M3}$ and the alternative hypothesis $H_1 : \mu_{WT} - \mu_{M3} > \mu_{M1} - \mu_{M3} + \mu_{M2} - \mu_{M3}$, where $\mu$ denotes the mean over replicate measurements (Figure 2.1H). Finally, we simplified our hypotheses by adding $\mu_{M3}$ to both sides of the equations above, to obtain $H_0 : \mu_{WT} = \mu_{M1} + \mu_{M2} - \mu_{M3}$ and $H_1 : \mu_{WT} > \mu_{M1} + \mu_{M2} - \mu_{M3}$. Wild-type probes for which we rejected the null hypothesis were considered 'cooperative'; we used the most stringent p-value cutoff (0.0003) for the one-sided Mann-Whitney $U$ test applied to the three replicate measurements for WT vs. the 27 measurements for M1+M2-M3 (Figure 2.6B). Wild-type probes with p>0.1 were considered 'independent', also known as 'additive' (Liu et al., 2020) since the binding level at the two-site probes was equal the sum of the

binding levels at the individual sites. In this study, we use the term 'independent' binding, for consistency with the ETS1-RUNX1 analyses.

As in previous work (Shen et al., 2018), and as described above for the ETS1-RUNX1 analyses, we filtered out WT sequences in the negative control range. In addition, to ensure that the introduced mutations indeed abolished the two binding sites, we filtered out all probes for which M3 was outside of the negative control range.

### 2.2.6 Designing genomic features to distinguish cooperative versus independent TF binding to neighboring sites

To investigate the rules that govern cooperative versus independent binding of TFs to neighboring sites, we trained classification models to distinguish between cooperative and independent binding events, using as training data the genomic sequences labeled as described above, based on cooperative gcPBM data. Feature design is critical for training accurate and interpretable classification models. Here, we designed features based on the genomic sequences themselves, taking into account characteristics important for TF-DNA recognition.

Figure 2.12 shows the main categories of features used in our analyses: 1) the distance between the two binding sites; 2) the orientations of the binding sites, which can either be positive (if the site matches the PWM in Figure 2.1B,E) or negative (if the site matches the reverse complement of the PWM); and 3) the strengths of the two binding sites, predicted using the PWM models. In addition, we designed features to represent the sequence composition and the DNA shape at and around the binding sites. We computed sequence and shape features taking into account the core binding sites (4bp for ETS1 and 5bp for RUNX1, as in Shen et al. (2018)), and up to

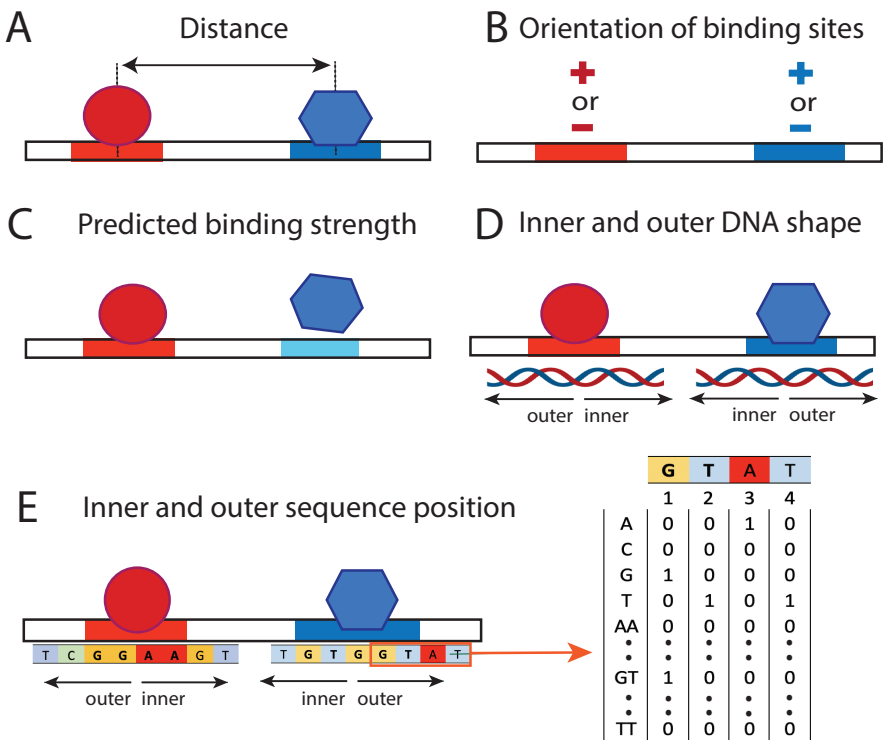



**Figure 2.12**: Features used in classification models: (A) distance between the two sites, (B) binding site orientation (defined based on PWM models), (C) predicted binding site strengths (using PWM scores), (D) positional DNA shape features, (E) positional k-mer features, shown for k=1 and 2.

4 positions in the inner and outer flanks (Figure 2.12D,E). The exact parameters were selected by nested 10-fold cross-validation. Shape features (minor groove width, roll, propeller twist, and helical twist) were computed using DNAshape (Zhou et al., 2013), and the identities of mono-, di- and trinucleotides were used to compute sequence features, as in Shen et al. (2018).

We trained Random Forest classification models, using the sklearn.ensemble module in Python. We tuned the parameters (n_estimators, max_depth, min-samples_leaf, and min_samples_split) by nested 10-fold cross-validation. We chose to focus on Random Forest as it provides interpretability and it allows us to reason about the features that drive the high classification accuracy. Furthermore, considering that the types of

features used are very diverse, Random Forest models allowed us to take into account nonlinear relationships between the features and the class.

## 2.3 Results

### 2.3.1 Cooperative gcPBM data reveals ETS1 cooperative binding to neighboring sites

Our data showed that ETS1 binds DNA cooperatively with itself, as well as with RUNX1. The patterns of cooperativity, however, and the features that distinguish cooperative from independent binding events are different for ETS1-ETS1 compared to ETS1-RUNX1, as described below.

Using the strategy described in Materials and Methods, we identified 976 cooperative and 1,185 independent binding events for the ETS1-RUNX1 experiment (Figure 2.13A), and 447 cooperative and 874 independent binding events for the ETS1-ETS1 experiment (Figure 2.13B). These results indicate that ETS1-ETS1 cooperativity might be more restrictive compared to ETS1-RUNX1, i.e. might depend more on certain features of the neighboring binding sites, such as distance and orientation.

In our RUNX1-ETS1 experiment, where we measured the influence of ETS1, present at a saturating level, on the RUNX1-DNA binding, we found 1,449 cooperative and 324 independent binding events (Figure 2.14), i.e. ∼82% of all genomic RUNX1 sites located in close proximity to ETS1 sites benefited from cooperative binding between the two proteins. In comparison, ∼45% of ETS1 sites benefited from cooperative binding with RUNX1, i.e. ETS1 binding at these sites increased significantly in the presence of a saturating level of RUNX1. Examining our data in

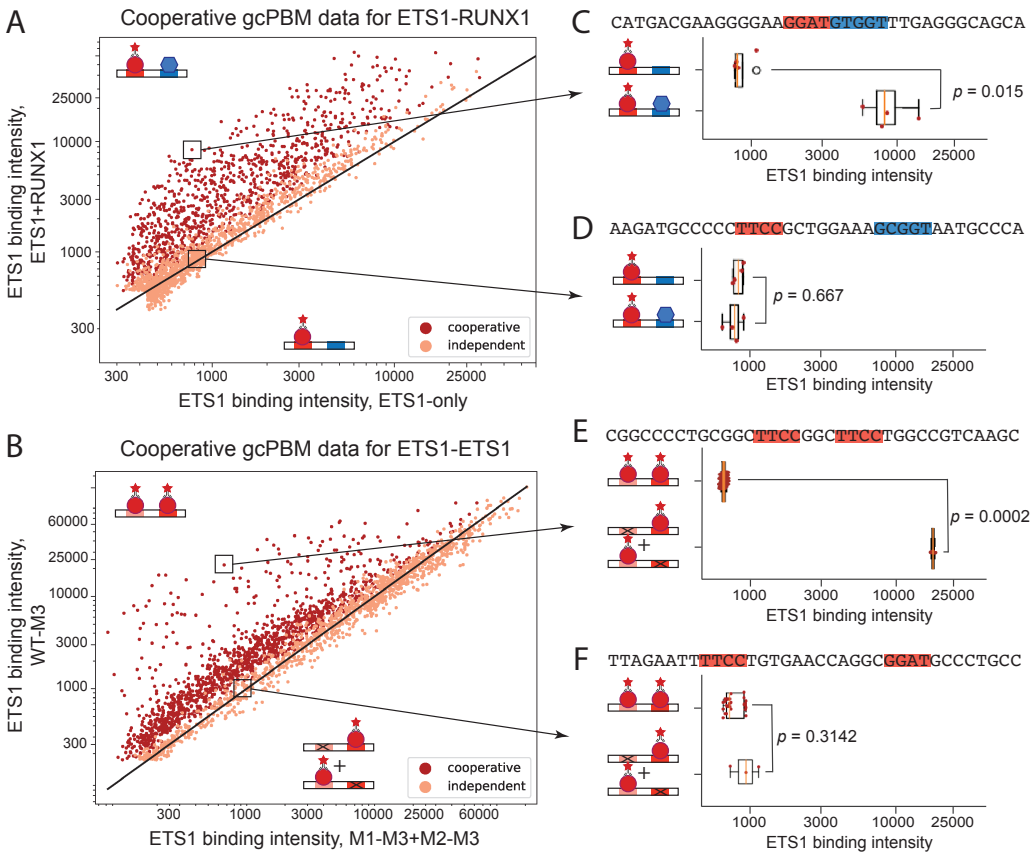



**Figure 2.13**: **Cooperative vs. independent binding revealed by cooperative gcPBM data.** (A) Comparison between the ETS1 binding intensity in the absence of RUNX1 (x-axis) versus the presence of RUNX1 (y-axis). Each point corresponds to a genomic DNA sequence with neighboring ETS1 and RUNX1 sites. Values shown are medians over four replicate spots. (B) Comparison between the ETS1 binding intensity at genomic loci containing two neighboring ETS1 sites (y-axis) versus the sum of binding intensities at the individual sites (x-axis). Each point corresponds to a genomic DNA sequence with two neighboring ETS1 sites. Values shown are medians over three replicate spots. (C,D) Examples of probes bound cooperatively (C) and independently (D) from the ETS1-RUNX1 experiment. (E,F) Similar to (C,D), but for the ETS1-ETS1 experiment. Boxplots show median signals, with boxes extending to the 25th and 75th percentiles. Whiskers extend to the largest/smaller values no further than 1.5 times the inter-quartile range. Points show the individual binding intensity measurements.

more depth, we found that the RUNX1 binding sites selected for our DNA library (i.e. the RUNX1 sites that are proximal to ETS1 sites) were generally of much lower affinity than RUNX1 sites found generally within ChIP-seq peaks (Figure 2.15A).

This trend was not as pronounced in the case of ETS1, i.e. we found many medium and high affinity genomic ETS1 binding sites in close proximity to RUNX1 sites (Figure 2.15B). This observation also explains why the total numbers of binding events that we could confidently call either cooperative or independent were different between the ETS1-RUNX1 and the RUNX1-ETS1 experiments: many of the RUNX1 sites tested were of very low affinity in the RUNX1-only experiment, and practically indistinguishable from negative controls. Overall, these results show that cooperativity is more frequent at lower affinity binding sites, and in the human genome the TF binding sites that are potentially cooperatively bound are not among those with the highest affinity, consistent with previous studies highlighting the importance of cooperativity at weak and medium-affinity TF binding sites (Tugrul et al., 2015; Ezer et al., 2014).

The genomic sequences included in our DNA libraries were selected in an unbiased manner, based simply on high-throughput *in vivo* binding data for ETS1 and RUNX1, and without biasing the library toward known sites of cooperative interaction between these proteins. Nevertheless, as described below, our data and results are in great agreement with previous, low-throughput studies on ETS1-ETS1 and ETS1-RUNX1 interactions, and they help strengthen and refine previous observations on the cooperative binding of the two proteins to their neighboring sites across the human genome.

### 2.3.2 Features encoded in the DNA are highly predictive of ETS1 cooperative binding

To investigate the determinants of cooperative versus independent DNA-binding by ETS1 and RUNX1, we trained classification models on our cooperative gcPBM data

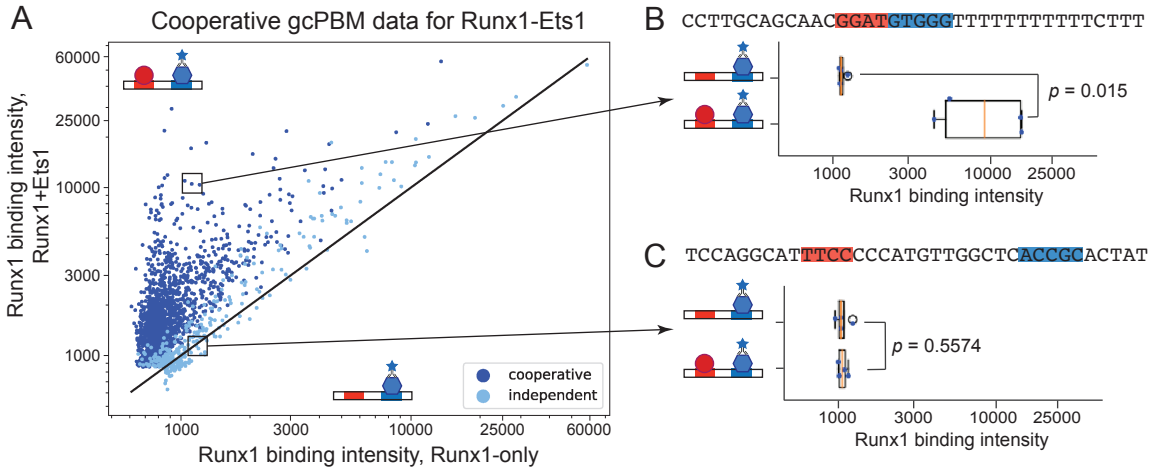



**Figure 2.14**: **Cooperative vs. independent binding of Runx1, revealed by cooperative gcPBM data.** (A) Comparison between the Runx1 binding intensity in the absence of Ets1 (x-axis) versus the presence of Ets1 (y-axis). Each point corresponds to a genomic DNA sequence with neighboring Ets1 and Runx1 sites. Values shown are medians over four replicate spots. (B,C) Examples of probes bound cooperatively (B) and independently (C) from the Runx1-Ets1 experiment. Boxplots show median signals, with boxes extending to the 25th and 75th percentiles. Whiskers extend to the largest/smaller values no further than 1.5 times the inter-quartile range. Points show the individual binding intensity measurements. P-values were calculated using a one-sided Mann-Whitney $U$ test.

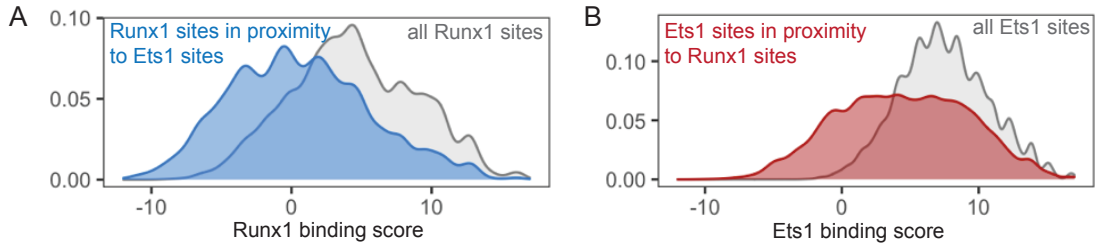



**Figure 2.15**: **Distributions of TF binding scores** (calculated based on PWM models) for (A) Runx1 binding sites and (B) Ets1 binding sites identified in ChIP-seq peaks.

using features derived from the genomic sequences containing the binding site pairs (Figure 2.12). Using the area under receiver-operating characteristic curve (AUC), we found that genomic features are highly predictive of cooperative binding (Figure 2.16), although different features are important in the ETS1-RUNX1 compared

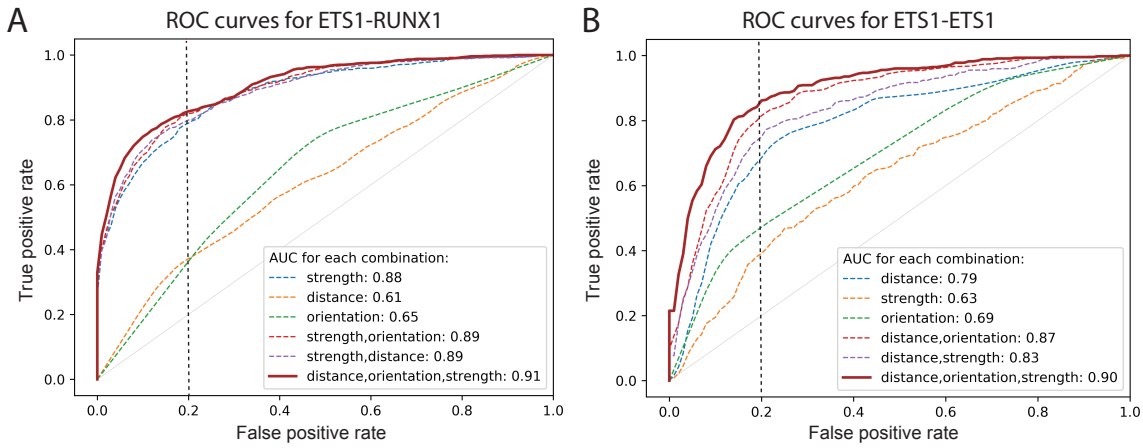



**Figure 2.16**: **Random Forest classification models can accurately distinguish between cooperative and independent binding events** for (A) the ETS1-RUNX1 system, and (B) the ETS1-ETS1 system. For each combination of features, the false positive and true positive rates were computed by averaging over the 10 folds of a cross-validation test. Dotted vertical lines show the 0.2 false positive rate.

to the ETS1-ETS1 systems.

### 2.3.3 Binding site strength is highly predictive of ETS1-RUNX1 cooperative DNA binding

For the ETS1-RUNX1 system, the most predictive feature type was the binding site strength (AUC: 0.88, Figure 2.16A), and in particular the binding strength of the cooperator factor (Figure 2.18A)), which is significantly different between cooperatively and independently bound regions ($p < 10^{-173}$, Figure 2.17A). In other words, our results indicate that at genomic regions containing one ETS1 and one RUNX1 site in close proximity, the higher the affinity of the RUNX1 site (and thus the stronger the RUNX1 binding), the more likely we are to observe that ETS1 benefits from cooperative binding with RUNX1. A similar trend was observed in the RUNX1-ETS1 experiment, where the strength of the cooperator's binding site—in this case, ETS1—was the most predictive feature for distinguishing cooperative from indepen-

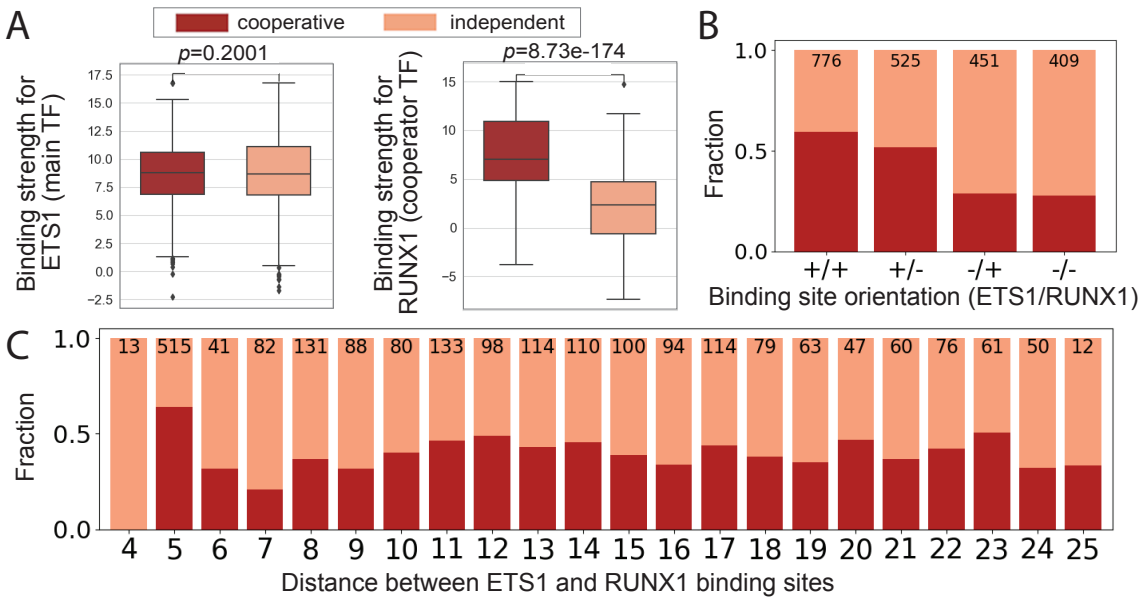



**Figure 2.17**: **Features used to train cooperative binding models on the ETS1-RUNX1 data.** (A) Binding site score distributions for sequences with cooperative vs. independent binding. Y-axes show PWM scores. (B,C) Fractions of probes bound cooperatively vs. independently, for different orientations and distances between the ETS1 and RUNX1 binding sites. Numbers above the barplots show the total number of sequences in each category.

dent binding events (Figures S13B, S14, S15). The orientations of the binding sites and the distance between them had only minor contributions to model accuracy (Figure 2.16A, Figure 2.19), consistent with the fact that we observed cooperative binding at all orientations and almost all distances tested (Figure 2.17B,C; Figure 2.20B,C).

This result suggests that the cooperative ETS1-RUNX1 binding to DNA is likely driven by protein-protein interactions through domains that are flexible enough to act at a wide range of distances and relative orientations of the binding sites. Our data argues against cooperativity through DNA, e.g. by allosteric DNA changes, as a dominating mechanism of ETS1-RUNX1 cooperativity, given the large variety of binding site sequences and configurations that can be bound cooperatively by the two proteins.

The hypotheses above are fully consistent with literature. Studies of select ETS1-

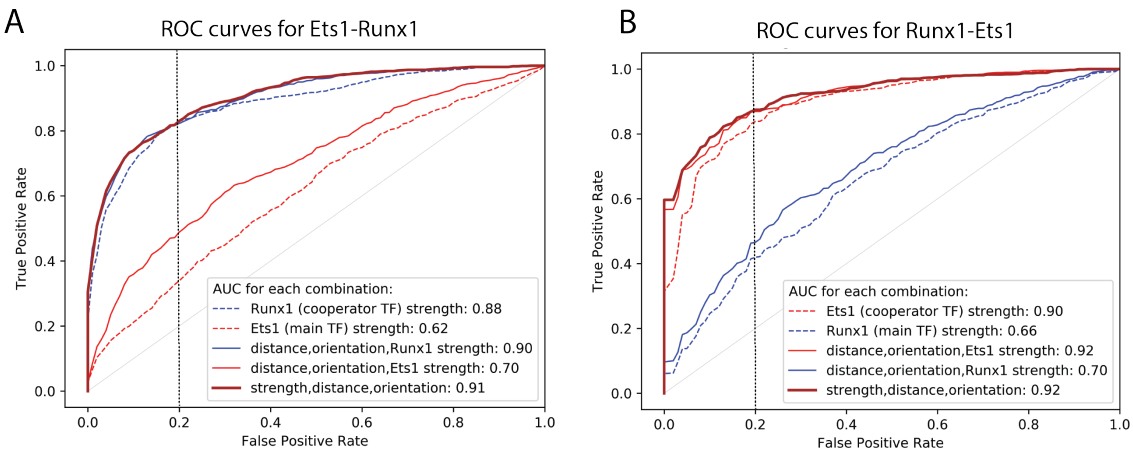



**Figure 2.18**: **Performance of Random Forest models trained separately on the binding site strength of either the main TF or the cooperator TF.** (A) Ets1-Runx1 system, where Ets1 is the main TF and Runx1 is the cooperator. (B) Runx1-Ets1 system, where Runx1 is the main TF and Ets1 is the cooperator. Dotted vertical lines show the 0.2 false positive rate.

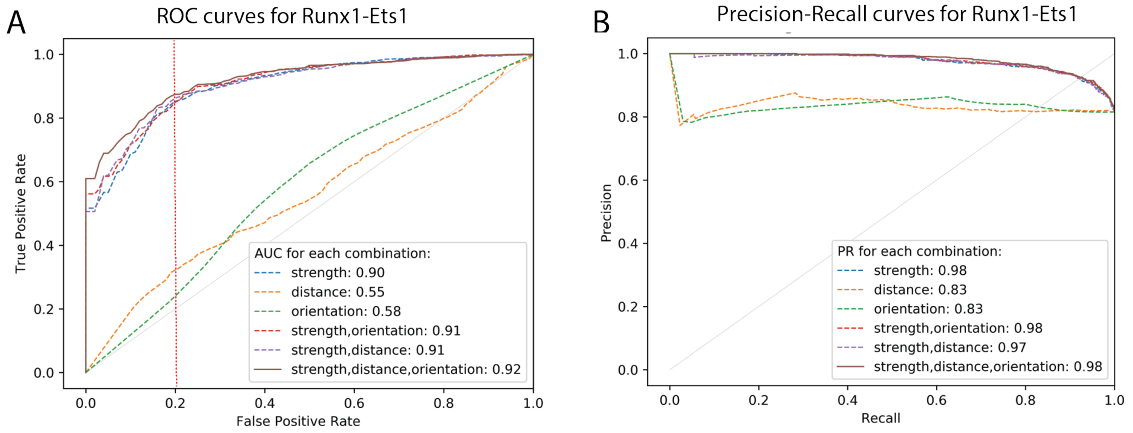



**Figure 2.19**: **Random Forest classification models can accurately distinguish between cooperative and independent binding for Runx1-Ets1.** (A) ROC curve. (B) Precision recall curve is also shown, to account for the small number of cooperative binding events in Runx1-Ets1 data set. For each set of features, the false positive and true positive rates were computed by averaging over the 10 folds of a cross-validation test. Dotted vertical lines show the 0.2 false positive rate.

RUNX1 binding site pairs showed that the cooperativity between the two TFs was retained even when nicks were introduced between their binding sites, thus arguing against a potential role for effects through the DNA (Goetz et al., 2000). Instead,

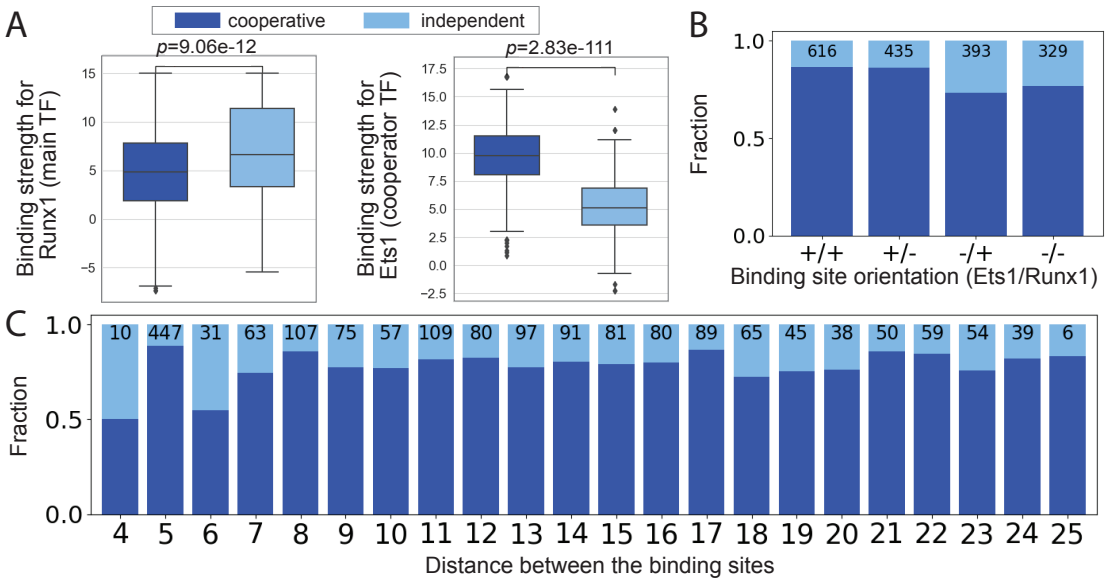



**Figure 2.20**: **Analyses of features used to train cooperative binding models on the Runx1-Ets1 data.** (A) Binding site score distributions for sequences with cooperative vs. independent binding. Y-axes show PWM scores. (B,C) Fractions of probes bound cooperatively vs. independently, for different orientations and distances between the Ets1 and Runx1 sites. Numbers above the barplots show the total number of sequences in each category.

ETS1 and RUNX1 were shown to enhance each other's binding to DNA primarily through direct interactions between their auto-inhibitory domains (Kim et al., 1999; Goetz et al., 2000; Gu et al., 2000; Shrivastava et al., 2014). The DNA-binding activity of ETS1 is negatively regulated by inhibitory regions that pack against the DNA-binding domain, with the N-terminal region of the inhibitory module unfolding upon ETS1-DNA binding (Petersen et al., 1995; Jonsen et al., 1996; Skalicky et al., 1996; Garvie et al., 2002). RUNX1 enhances ETS1-DNA binding by interacting directly with ETS1, disrupting the packing of the auto-inhibitory module and thus countering the auto-inhibition; this can increase the DNA-binding affinity of ETS1 by up to ∼30-fold (Goetz et al., 2000). Furthermore, the crystal structure of the ETS1-RUNX1-DNA ternary complex (Shrivastava et al., 2014) supports the hypothesis of a flexible linker region in the RUNX1 protein that connects its DNA-binding and its

ETS1-interaction domains, consistent with our data showing cooperative binding at a wide range of binding site distances and orientations (Figure 2.17).

### 2.3.4 The role of binding site distance and orientation in ETS1-RUNX1 cooperative DNA binding

The distance between neighboring binding sites can be a strong determinant of cooperative TF-TF interactions (Morgunova and Taipale, 2017; Kim et al., 2013). For the ETS1-RUNX1 system, however, we found that the distance feature does not contribute significantly to the accuracy of our predictive model (Figure 2.16A), suggesting only a small effect on cooperativity (at least in the range of distances tested). This finding is in agreement with previous, small-scale studies, and generalizes the results of those studies by significantly expanding the number of sequences tested.

Wotton et al. (1994), Sun et al. (1995), and Goetz et al. (2000) reported cooperative binding between ETS1 and RUNX1 when their sites were at distances d=5,7-9,12-14,16,19,27 and 38; for most of these distances only a single pair of sites, and thus a single orientation, was tested. Our data revealed cooperative binding between ETS1 and RUNX1 at all distances d=5-25, and all orientations, thus strengthening the point that the cooperativity between these proteins is largely insensitive to the distance between their DNA-binding sites. Nevertheless, the fraction of cooperatively-bound regions was different at different distances, varying from 63.9% at d=5 to 20.7% at d=7 (Figure 2.17).

The distance d=5 is clearly the most preferred for ETS1-RUNX1 cooperativity: it showed the largest fraction of cooperative binding events (63.9%, or 329/515) and it contains the most probes in our library (515 out of 2,161; Figure 2.17), which means that it is the most common distance between ETS1 and RUNX1 sites in the genomic

regions co-bound *in vivo*. Cooperativity between ETS1 and RUNX1 at d=5, and specifically in the +/+ orientation (where 81.8%, or 266/325 regions were labelled as cooperative), has been widely investigated in molecular, structural and bioinformatic studies (Wotton et al., 1994; Sun et al., 1995; Kim et al., 1999; Goetz et al., 2000; Shrivastava et al., 2014; Hollenhorst et al., 2009). This is the configuration (i.e. the spacing and orientation) between the known cooperative ETS1-RUNX1 sites in the TCR$\alpha$ and TCR$\beta$ enhancers, and it showed the largest cooperativity ($\sim$33-fold) in a synthetic sequence tested by Goetz et al. (2000). Furthermore, Hollenhorst et al. (2009) reported that a composite motif corresponding to this configuration was enriched in ETS1-bound genomic regions near genes important for T-cell activation, regions also bound by RUNX1 *in vivo*. No other ETS1-RUNX1 configurations were enriched in the ChIP-seq data, so the authors speculated that this spacing and orientation may also be important for transcriptional activation. Our data, however, suggests that this particular configuration is highly preferred even for the DNA-binding itself.

Interestingly, when comparing the ETS1-RUNX1 sites in this configuration compared to all other configurations present in our DNA library, we found that cooperative regions with +/+ orientation and d=5 (Figure 2.22A) have lower RUNX1 binding affinity than cooperative binding regions with other configurations (Figure 2.22B); no significant difference was observed for the independent regions (Figure 2.22C), nor for the ETS1 binding strength (Figure 2.21A). Considering all configurations tested in our study, we found a significant negative correlation between the RUNX1 binding strength of cooperatively bound sites and the fraction of cooperative sites observed in that configuration (Figure 2.22D); the trend was even stronger when focusing on the +/+ orientation (Figure 2.22E), and it was not significant for independently bound sites (Figure 2.21B,C). In other words, our results suggest that in order for

ETS1 to benefit from cooperativity with RUNX1, a strong RUNX1 affinity site is generally required; however, at certain 'preferred' configurations (identified here as configurations at which we see large fractions of cooperative binding events), lower affinity RUNX1 binding sites can also lead to cooperative binding. This observation is consistent with the fact that the distance and orientation features help improve our model of cooperative vs. independent ETS1-RUNX1 binding, although the RUNX1 binding strength feature is the most important for model accuracy (Figure 2.16A).

Overall, our results are consistent with small-scale studies showing an apparent insensitivity of ETS1 and RUNX1 to the orientation and spacing between their binding sites (Wotton et al., 1994; Kim et al., 1999; Garvie et al., 2002; Goetz et al., 2000). In addition, our high-throughput data and analyses reveal a more nuanced view of ETS1-RUNX1 cooperativity (Figure 2.21D,E), where the precise configuration of the neighboring sites does play a role in the cooperative interaction, although the binding strength of the cooperator (here, RUNX1) is the main determinant.

### 2.3.5 Binding site distance and orientation are predictive of ETS1-ETS1 cooperative binding

For the ETS1-ETS1 system, we found that cooperative binding depends mainly on the distance between the two sites (AUC 0.79; Figure 2.16B). Nevertheless, adding strength and orientation features significantly improved the model performance: AUC=0.87 for distance+orientation and AUC=0.83 for distance+strength. When all three types of features were used, model performance was further improved, reaching an AUC of 0.9. This trend is clearly different from what we observed in the ETS1-RUNX1 system, where binding strength was the dominant feature and distance and orientation has only minor contributions (Figure 2.16B), suggesting a different

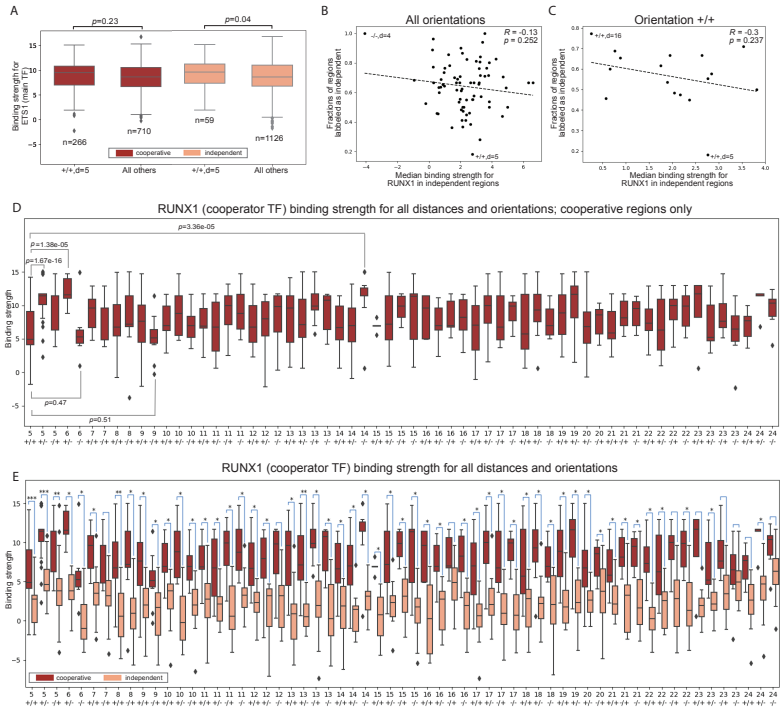



**Figure 2.21**: (A) Binding strength of ETS1 for the sites in the ETS1-RUNX1 experiment for orientation +/+,distance=5 versus all other configurations. The numbers below each box plots show the number of genomic regions in each category. No significant difference was observed for the ETS1 binding strength. (B) RUNX1 binding strength in regions labeled as independent. The x-axis shows the median RUNX1 binding strength over all independent regions with a particular configuration, i.e. distance + orientation. The y-axis shows how "preferred" each configuration is for independent binding, assessed by the fraction of independent binding regions with that configuration. Only configurations with 5 or more regions in our DNA library are shown. (C) Same as (B) but showing only the +/+ orientation. (D) RUNX1 binding strength for all configurations in the cooperative regions. Only configurations with 5 or more regions in our DNA library are shown. Highlighted above the box plots are the top three configurations for which the RUNX1 strength is most different compared to (+/+,d=5). Highlighted below the box plots are the top two configurations for which the RUNX1 strength is most similar compared to (+/+,d=5); these two configurations, (-/-,d=6) and (-/+,d=9) also seem favourable for cooperativity, since cooperative binding occurs even when the cooperator site (here, RUNX1) is of low affinity. (E) Similar to panel (D), but for both cooperative and independent regions. We found that RUNX1 binding strength is always higher at the cooperative regions. The asterisks mark the significance of the difference in RUNX1 binding strength between cooperative and independence regions: *** : $p < 1e-10$; ** : $p < 1e-5$; * : $p < 1e-2$.

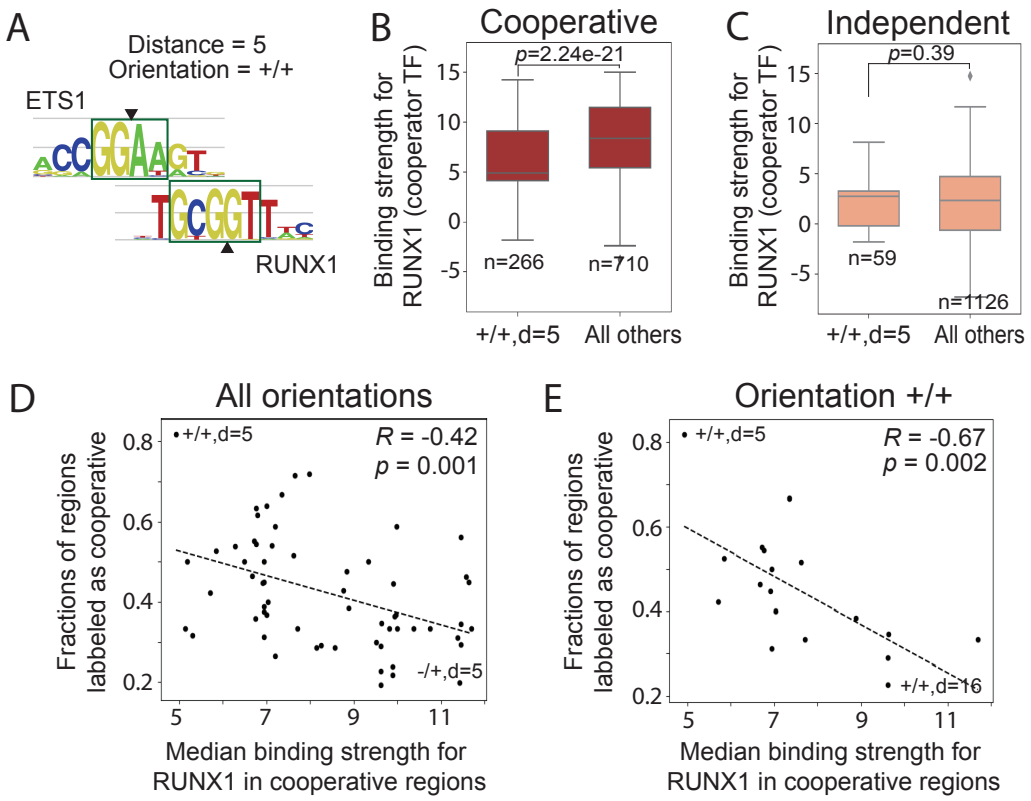



**Figure 2.22**: (A) Alignment of ETS1 and RUNX1 motifs in the +/+,d=5 configuration. (B) Cooperative regions in the +/+,d=5 orientation have lower affinity RUNX1 binding sites compared to cooperative regions in all other configurations. (C) In independent regions, the RUNX1 binding affinity is not different. (D) Each point represents one configuration, i.e. distance + orientation. The x-axis shows the median RUNX1 binding strength over all cooperative regions with a particular configuration. The y-axis shows how 'preferred' each configuration is for cooperative binding, assessed by the fraction of cooperative binding events observed at regions with that configuration. Only configurations with 5 or more regions in our DNA library are shown. (E) Same as panel D, but showing only configurations with the +/+ orientation.

mechanism of cooperativity for ETS1-ETS1 vs. ETS1-RUNX1, as discussed below.

Closer examination of the cooperative vs. independent sites for the ETS1-ETS1 system (Figure 2.23D) revealed a clear pattern: as the distance between the two neighboring ETS1 sites increases, the fraction of cooperative binding events observed at that distance decreases. For distances d=4-9, more than 65% of binding events

were cooperative, for d=10-15 we found between 23% and 51% of neighboring sites to be bound cooperatively, while at larger distances the fraction of cooperative events was typically < 12%. These observations suggests that a physical interaction between structured, rigid domains of the two ETS1 molecules is a likely mechanism for achieving cooperative DNA-binding, as opposed to an interaction through flexible regions that can extend to a wide range of distances, which was the case for ETS1-RUNX1.

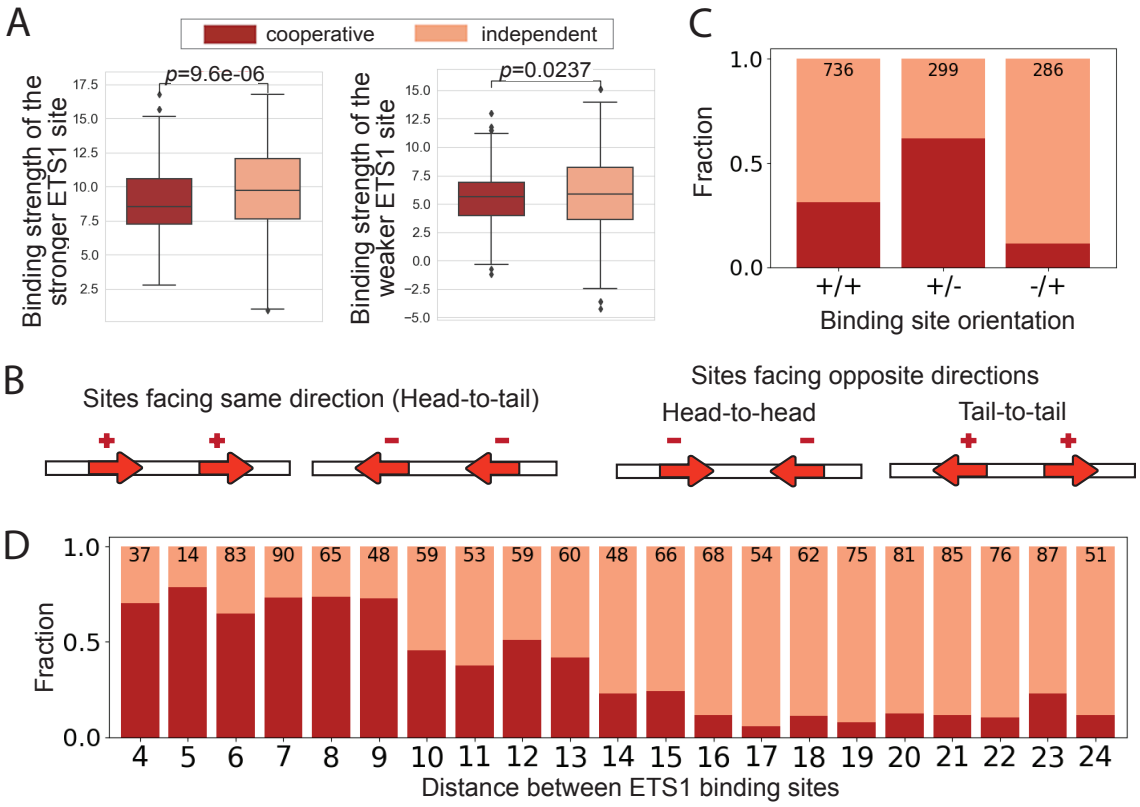



**Figure 2.23**: **Features used to train cooperative binding models on the ETS1-ETS1 data.** (A) Binding site score distributions for sequences with cooperative vs. independent binding. Y-axes show PWM scores. Separate plots are shown for the stronger site (left) and the weaker site (right). (B) Relative orientations for neighboring ETS1 binding sites. The tip of the arrow corresponds to the "head" of the ETS1 binding site. (C,D) Fractions of probes bound cooperatively vs. independently, for different orientations and distances between the neighboring ETS1 binding sites. Numbers above the barplots show the total numbers of sequences in each category.

The orientation between neighboring sites is also an important determinant of ETS1-ETS1 cooperative binding (AUC=0.69; Figure 2.16B), with the +/- orientation showing the largest fraction of cooperative binding events (62%, Figure 2.23C), followed by +/+ with 31% and -/+ with 12%. This trend suggests, yet again, that the cooperative ETS1-ETS1 interactions are likely mediated through direct contacts between specific regions of the two ETS1 molecules, which occur preferentially when the binding sites are in the +/- orientation, i.e. at GGAA-gap-TTCC sequences.

Indeed, co-crystal structures of two ETS1 molecules bound to neighboring DNA sites confirm the cooperativity mechanism suggested by our data. At the human stromelysin-1 promoter, two molecules of ETS1 are known to bind neighboring sites in the +/- orientation (GGAAgcacTTCC) at adjacent major grooves (Lamber et al., 2008; Shiina et al., 2015). Structural data revealed direct contacts (hydrogen bonds and van der Waals) between Gly333 and Pro334 of one ETS1 molecule and Asn380 and Lys379 of the other ETS1 (Lamber et al., 2008; Babayeva et al., 2010). In addition, mutation analyses showed that cooperative binding at this promoter was abolished when residues Gly333, Pro334, and Asn380 were mutated (Lamber et al., 2008; Babayeva et al., 2010). These data indicate a mechanism of cooperative binding facilitated by DNA-mediated homodimerization of ETS1, with the two ETS1 molecules overcoming DNA-binding autoinhibition through direct interactions between well-structured domains of the two ETS1 molecules.

Contrary to the ETS1-RUNX1 system, in the case of ETS1-ETS1 interactions we found that the binding strengths of the two neighboring sites are not strong predictors of cooperative vs. independent binding (AUC=0.63; Figure 2.16B). To include binding site strength features in our models, for each pair of neighboring ETS1 sites we identified the stronger site and the weaker site, and we used their PWM-based binding scores as features. Each feature by itself had even lower prediction accuracy

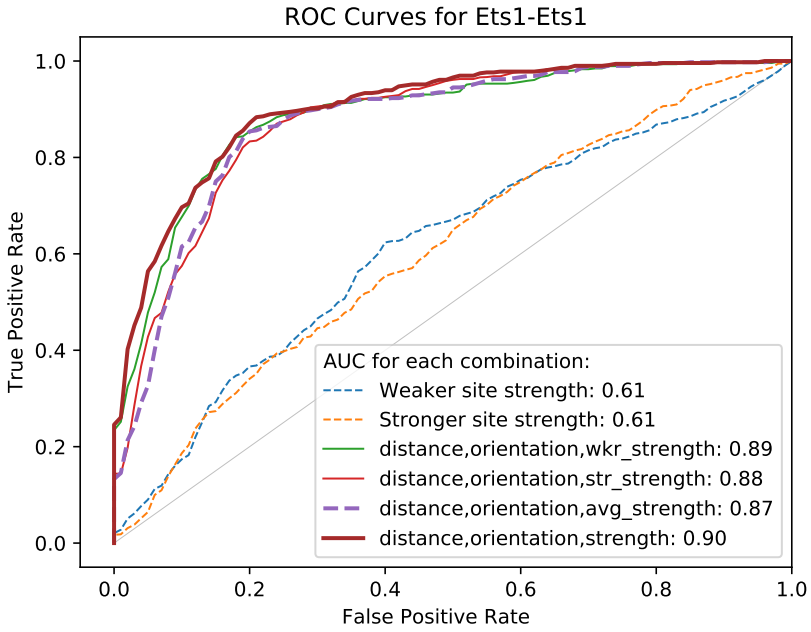



**Figure 2.24**: Performance of Random Forest models trained separately on the binding site strength of either the weaker or the stronger of the two Ets1 sites, for the Ets1-Ets1 system.

than the two features together (AUCs=0.61; Figure 2.24). A closer examination of the two features for regions bound cooperatively vs. independently by ETS1 showed no significant difference in binding strength for the weaker ETS1 site and a moderate difference for the stronger ETS1 site (Figure 2.23A). Overall, these results are consistent with our general expectation that cooperativity occurs more often at weaker TF binding sites. However, in the ETS1-ETS1 system, binding site distance and orientation are the most predictive features of cooperative vs. independent binding, consistent with previous biochemical and structural studies (Lamber et al., 2008; Babayeva et al., 2010; Shiina et al., 2015).

### 2.3.6 Using DNA shape and sequence features to distinguish cooperative versus independent binding

For both the ETS1-RUNX1 and the ETS1-ETS1 system, we found that features reflecting the binding strengths of the two neighboring sites improve the performance of our classification models. Mechanistically, the strength of TF-DNA interactions is determined by direct (base) readout and indirect (shape) eadout (Rohs et al., 2010). Thus, we next asked whether the improvement in classification accuracy when adding binding strength information was due to specific DNA sequence and/or DNA shape features of the binding sites and their flanking regions. To answer this question we trained Random Forest classification models using DNA sequence and shape features, as explained in Materials and Methods.

Our predictions revealed that DNA sequence and shape features are indeed predictive of cooperative vs. independent binding, although they cannot reach the prediction accuracy of binding strength features (Figure 2.25A,B). This could be due to the fact that additional structural information is needed to fully describe sites with cooperative binding events, including potential information on the structural deformations induced by the two proteins, which is not currently available but may be intrinsically captured by the binding strength features.

Focusing on the configurations (i.e. orientations and distances) with the most data in our study, we observed that, as expected, the DNA shape and sequence features do differ around sites with cooperative vs. independent binding, for both the ETS1-RUNX1 (Figure 2.25C,D) and the ETS1-ETS1 (Figure 2.25E,F) systems. Furthermore, the predictive power of DNA shape features, as observed in our study, is consistent with a prior study (Ibarra et al., 2020).

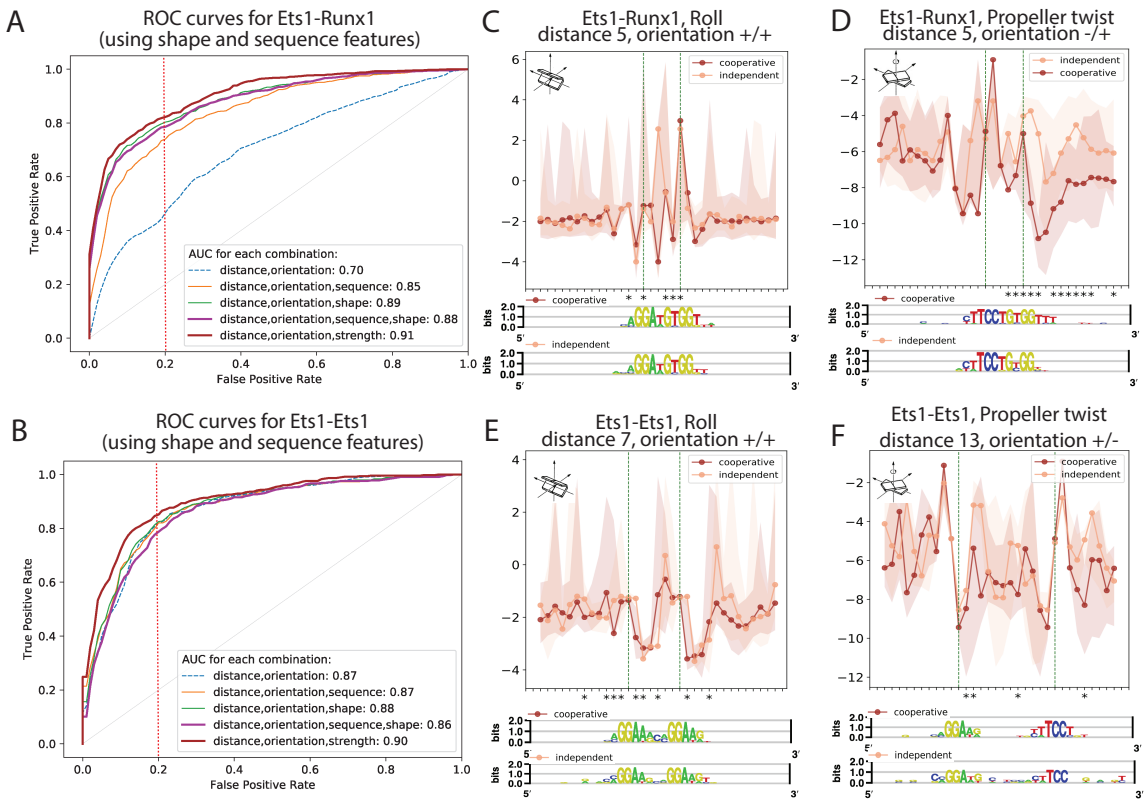



**Figure 2.25**: **DNA sequence and shape features are predictive of cooperative vs. independent binding.** (A, B) ROC curve for Random Forest models using DNA shape and sequence features, compared to models using distance, orientation, and binding site strength. (C,D,E,F) Examples of DNA shape and sequence differences between sites with cooperative vs. independent binding, for probes with various distances between the binding site centers. Stars (*) mark positions with significant differences in shape features between cooperative and independent sites ($p < 0.05$, one-sided Mann-Whitney test). Sequence motifs for the cooperative and independent binding events are shown.

## 2.4 Discussion

Cooperative TF binding is crucial in the regulation of gene expression. Nevertheless, questions surrounding the rules that drive cooperative versus independent binding remain unanswered. Here, we developed the 'cooperative gcPBM' assay to measure cooperative and independent binding of TFs to neighboring DNA sites in high throughput. Taking the cooperative gcPBM data as input, we trained random forest

classification models and we identified genomic features predictive of cooperativity between TFs. Using ETS1-RUNX1 and ETS1-ETS1 as our case studies, we were able to identify distinct mechanisms that ETS1 uses to bind DNA cooperatively.

CAP-SELEX assays (Jolma et al., 2015) have been previously used to assess cooperative binding between TFs, based on over-representation of specific spacings and orientations among binding sites in randomized DNA sequences bound by two TFs. At such spacings/orientations, CAP-SELEX can detect "composite motifs" (similar to the motifs shown in Figure 2.25), which reflect highly preferred binding site configurations. Surprisingly, no ETS1-RUNX1 composite motif was reported in the CAP-SELEX study, despite ETS1-RUNX1 interactions have been extensively investigated in the literature (Wotton et al., 1994; Kim et al., 1999; Goetz et al., 2000; Gu et al., 2000; Shrivastava et al., 2014; Shiina et al., 2015). On the other hand, in a lower throughput co-selection experiment, focused specifically on ETS1 and RUNX1, Wotton et al. (1994) did identify one configuration (+/+ orientation, d=5) as the most highly represented in their data. Overall, our data shows that many other configurations are also compatible with cooperative binding, and further studies need to be conducted to investigate why particular configurations are over-represented among the cooperative binding events in the genome.

In another recent study, Ibarra et al. (2020) combined CAP-SELEX data with data from HT-SELEX experiments (Jolma et al., 2010), which reflect binding of individual TFs. They defined cooperative binding events based on the relative affinity of the TF pair (derived from CAP-SELEX) versus the mean relative affinity of the two individual TFs (derived from HT-SELEX). While this approach revealed new insights into cooperative binding by ETS and FOXO proteins, a major disadvantage was the reliance on accurate measurements of relative affinities for individual TFs tested by HT-SELEX. Being based on exponential enrichment of strongly bound

sites, SELEX-based assays cannot accurately capture the affinities of medium and low affinity binding sites, which are exactly the type of sites where we observed cooperative binding in our study. On the other hand, sequences in our cooperative gcPBM were selected in an unbiased manner, regardless of the enrichment of the binding sites, which enables us to capture medium- to low-affinity binding sites that can be important for cooperative binding.

Cooperative binding has also been studied *in vivo*. For example, Datta et al. (2018) used ChIP-seq data sets for the TF of interest with the co-factor present or knocked-out. However, knocking out a TF in human cells is not trivial and can lead to a wide range of effects, making it hard to deconvolve direct from indirect effects on the binding of the TF of interest. In addition, ChIP-seq is performed on a population of cells, and different sites within a peak may be bound in different cells. Furthermore, due to the low resolution of ChIP-seq data, a peak can contain multiple putative binding sites, and it is unclear which sites are bound in the cell (such as shown in Figure 2.2). Higher resolution ChIP based assays have been developed (Rhee and Pugh, 2012; He et al., 2015), making it possible to study cooperative binding *in vivo* while also capturing the influence of high-order chromatin organization (Avsec et al., 2021). Our method, can complement such *in vivo* studies by indicating important genomic features used by TFs to bind cooperatively to DNA sequences, in the absence of other cellular factors.

In conclusion, here we show how that high-throughput TF-DNA binding assays, specifically designed to detect cooperative binding, can be combined with machine learning modeling to identify potential mechanisms of TF-TF cooperativity on the DNA. In addition, this type of data can pinpoint which pairs of TF binding sites should be investigated further, through structural and functional studies, to characterize the cooperativity between particular TFs and its role in gene regulation.

In the case of proteins such as ETS1 and RUNX1, which have several interacting partners and complex regulatory functions in the cell, the approach described here would shed light on which genomic sites are used in cooperative interactions with different partners. In future work, we plan to apply our approach to other TF pairs to characterize their patterns of cooperative binding.

# Chapter 3

# QBiC-Pred: Quantitative Predictions of Transcription Factor Binding Changes Due to Sequence Variants

## 3.1 Introduction

Genetic variants and mutations play important roles in human disease (Khurana et al., 2016). Most variants occur in non-coding genomic regions, where they can impact gene expression by disrupting interactions between transcription factors (TFs) and DNA. In previous work we have introduced an ordinary least squares (OLS)-based method for assessing the impact of non-coding mutations on TF-DNA interactions (Zhao et al., 2017). Briefly, we used high-throughput *in vitro* TF binding data from universal protein-binding microarray (uPBM) experiments (Berger and Bulyk, 2009)to train regression models of TFDNA binding specificity using OLS estimation. Next, we used the OLS models to predict changes in TF binding due to DNA mutations, and we showed that our binding change predictions correlate well with measured changes in gene expression.

Our approach is novel compared to previous models because, by using OLS, we obtain not only estimates of the model coefficients, but also the variance of these estimates, which allows us to compute normalized binding change scores (z-scores) and significance levels ($P$-values) reflecting our confidence that a mutation affects TF binding. The computed P-values implicitly take into account the quality of the model

and of the training data, so in the case of poor predictive models a large change in binding is required for a mutation to be called significant (Zhao et al., 2017).

Here, we introduce QBiC-Pred (Quantitative Predictions of TF Binding Changes Due to Sequence Variants) (Martin et al., 2019), or QBiC for short, a web service that allows users to run our OLS models through a user-friendly web interface.

***Input***. QBiC takes as input mutation/variant datasets containing single nucleotide variants, in several formats: (i) variant files in the standard variant call format (VCF); (ii) 'simple somatic mutations' files generated by the International Cancer Genome Consortium (ICGC) (Hudson et al., 2010); (iii) tab- or comma separated values files with the columns: chromosome, chromosome_pos, mutated_from and mutated_to; and (iv) text files containing 17-bp DNA sequences with the mutated nucleotide in the center, followed by the 'mutated to' nucleotide, separated by a space character. The first three formats can be used with genomic coordinates from versions hg19 and hg38 of the reference human genome, while the sequence format allows users to input custom DNA sequences. For the sequence format, the context of each variant (8-bp on each side) is needed in order to assess the binding status of each allele, using uPBM 8-mer enrichment scores ($E$-scores) (Berger et al., 2006; Berger and Bulyk, 2009). Examples of input mutation files are described in the 'About' section of the website, and available for download. QBiC also takes as input a list of TF proteins of interest, from a list of 582 human TFs with available OLS models. All TF names are specified using the standard HUGO gene nomenclature (HGNC) (Yates et al., 2014). The list of available TFs and models is available on the QBiC website in the 'Downloads' section.

***Output***. For each input variant, QBiC runs the OLS models for the list of specified human TFs and it computes the predicted TF binding changes, the normalized changes ($z$-scores), the significance of the changes according to each model

($P$-values), as well as the predicted changes in binding status (e.g. from specific binding, or 'bound', to nonspecific binding, or 'unbound') assessed using uPBM 8-mer data. Similarly to our previous work (Shen et al., 2018), we consider a site 'bound' if it contains two consecutive overlapping 8-mers with $E$-scores $> 0.4$, and 'unbound' if it contains only 8-mers with $E$-score $< 0.35$; all other sites are called 'ambiguous'. The $E$-score cutoffs can be modified by the user through the QBiC interface. All computed values are reported as output, in table format. The precise models used by QBiC for each TF protein, as well as the PBM data used to train each model, are reported as part the QBiC results. The user can further process the results using the web interface (e.g. to specify a more stringent $P$-value cutoff for the binding change predictions) and can download the full or filtered results. The web interface also allows users to directly download models or datasets used to obtain individual predictions, and provides links to the HGNC database (Yates et al., 2014) where users can find additional information about individual TFs.

We are not aware of web servers with the same functionality as QBiC. Users interested in evaluating the putative effects of non-coding mutations on TF-DNA binding can certainly use any of the available databases of position weight matrices (PWMs) (e.g. (Hume et al., 2015; Matys et al., 2006; Weirauch et al., 2014; Khan et al., 2018)) or deep learning models Alipanahi et al. (2015) of TF-DNA binding specificity, or search existing databases of annotations for non-coding variants (e.g. (Ward and Kellis, 2015; Boyle et al., 2012)). However, such databases do not provide information on the quality of the binding models, and, as shown in the 'Results' section below, PWM and deep learning models are not as accurate as our OLS models in predicting the *quantitative* effects of DNA variants on TF binding. The OLS models used in QBiC also have the advantage of providing a direct measure of the significance of each predicted TF binding change, given the model and the training data. This

unique feature of our models facilitates interpretation of the results and allows users to prioritize variants for further analysis and validation.

## 3.2 Materials and methods

### 3.2.1 OLS models of TF-DNA binding specificity

The OLS models used by QBiC were trained on curated uPBM data from literature and our laboratory, mapped to 577 human TF proteins. Each uPBM experiment measures the binding specificity of a TF for ∼44,000 60-bp long DNA sequences, each containing a 36-bp variable region followed by a constant 24-bp primer complement (necessary for DNA double-stranding (Berger et al., 2006)). We use as features the number of occurrences of each possible 6-mer within the 60-bp sequences, and as outcomes the log-transformed fluorescence intensity signals, which reflect the levels of TF binding. The entire 60-bp sequence is used to count 6-mer occurrences, despite the fact that part of the sequence is constant, because the TF proteins can bind at any location within the 60-bp DNA molecule. We consider each 6-mer and its reverse complement as the same variable and combined their counts as one feature, resulting in a total of 2,080 features. The relationship between the outcomes $Y$ and the features $X$ is modeled by a multiple linear regression $Y = X\beta + \varepsilon$.

To characterize the TF binding change due to a single nucleotide variant, we define binding scores for the wild-type and the mutant sequences, as the sum of the coefficients for all 6-mers overlapping the variant, in an 11-bp window. The difference between these two scores, which represents the binding change, can be expressed as a linear combination of the regression coefficients: $c^T\beta$, where $\beta$ is a vector containing the coefficients for all 2080 6-mer count features, and $c$ is a vector of the same length

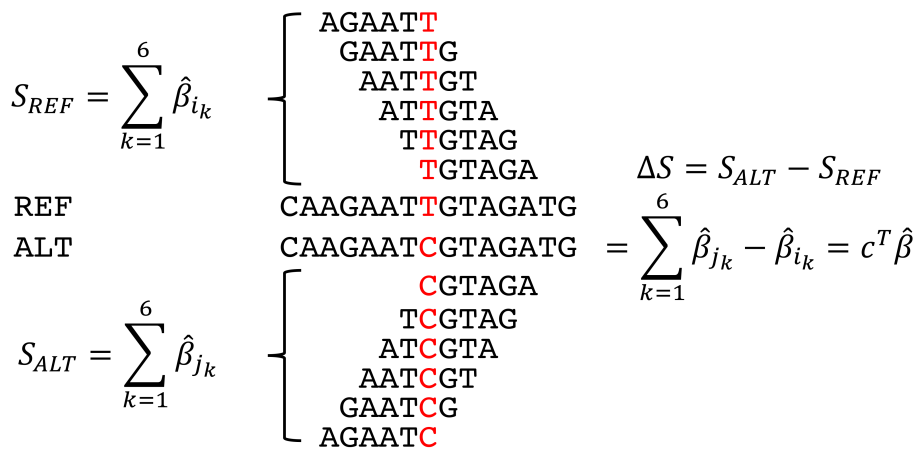



**Figure 3.1**: **TF binding change calculation.** The change in TF binding is computed as a linear combination of the coefficient estimates for all 6-mers overlapping the variant.

containing, for each 6-mer, the difference in counts due to the variant (Figure 3.1). We note that most components of $c$ are 0, as the variant affects the counts for up to twelve 6-mers.

By further assuming normality on the error term of the linear regression model $\varepsilon \sim N(0, \sigma^2 I)$, we are able to leverage the statistical properties of OLS estimation in order to test whether the binding change is statistically significant. The null hypothesis $H_0 : c^T\beta = 0$ can be tested using a t-statistic: $t = c^T\hat{\beta}/\sqrt{c^T\hat{\Sigma}c}$. Here, $\hat{\beta}$ is the OLS estimate for the coefficients vector: $\hat{\beta} = (X^TX)^{-1}X^TY$, and $\hat{\Sigma}$ is an unbiased estimate for the covariance matrix of $\hat{\beta}$: $\hat{\Sigma} = \hat{\sigma}^2(X^TX)^{-1}$, where $\hat{\sigma}^2 = (Y - X\hat{\beta})^T(Y - X\hat{\beta})/(n - p)$, with $n$ being the number of observations and $p$ the number of features. Since the regression contains ∼44,000 observations and 2,080 variables, this t-statistic follows a t-distribution with ∼42,000 degrees of freedom. Thus, we can use a normal approximation to derive the z-score and calculate the $P$-value of the test. For each TF and variant given as input, QBiC calculates and reports the difference in TF binding, the corresponding $z$-score, and the associated

$P$-value.

To select the uPBM data used in QBiC, we started with 3,342 data sets from CIS-BP (Weirauch et al., 2014), 245 data sets from UniPROBE (Hume et al., 2015) that were not included in CIS-BP, and 23 data sets generated in our laboratory (Shen et al., 2018). By using the information in the Human Transcription Factors database (Lambert et al., 2018) for the publicly available uPBM data, and manually curating the data generated in our laboratory, we mapped 1,451 uPBM data sets to 638 human TF proteins, using both uPBM experiments that tested human TFs as well as experiments for homologous TFs with high amino-acid identity in the DNA-binding domain region, similarly to Lambert et al. (2018). Next, to assess the quality of each uPBM data with respect to our task of training accurate quantitative models of TF-DNA binding specificity, we used the cross-validation accuracy of OLS models trained on each uPBM dataset. We removed data sets of poor quality (cross-validation correlation $< 0.2$ computed for the top 10% and top 20% sequences with the highest intensity), and for each TF we selected at most six uPBM data sets, including the top three data sets with the highest cross-validation accuracy, as well as the top three data sets obtained for TFs with the highest amino-acid identify to the human TFs. The final mapping, which includes 667 uPBM data sets and 577 TFs, is available on the QBiC website in the 'About' section.

### 3.2.2 *In vitro* measurements of TF binding changes due to single nucleotide variants

The PBM technology can be used, with custom-designed DNA libraries, to directly measure the *in vitro* effects of single nucleotide variants on TF binding. To build custom DNA libraries we first selected, at random, DNA sequences containing binding

sites for the TFs of interest, and then we introduced all possible single nucleotide variants in the binding site and the immediate flanking regions. Next, we measured the TF binding intensity for all the sequences, and we computed the log ratio of the binding signal between each mutant and the corresponding wild-type sequence to denote the TF binding change due to each variant.

We designed two such DNA libraries and used them to perform custom PBM experiments for six TFs. The DNA library for CREB1, RUNX1, and STAT3 included all single nucleotide variants in the TF binding site (10-12 bp), while the library for ETS1, ELK1, and GATA1 included all single nucleotide variants in the TF binding site and the flanking regions (36 bp). Because several TFs were tested against each DNA library, for each TF we obtain binding data both for variants in their specific binding sites, as well as variants in non-specific regions (which were present in the DNA library because they are specific to other TFs). We used all measurements to evaluate the accuracy of our predictions of TF binding changes (see Chapter 3.3).

### 3.2.3 *In vivo* allele-specific binding data

Allele-specific measurements of TF binding from *in vivo* ChIP-seq data have contributed to the identification of genetic variants that have the potential to change TF binding in the cell (Wagih et al., 2018; Shi et al., 2016). After mapping ChIP-seq reads to each allele of heterozygous variants, allele-specific binding (ASB) events can be identified as the ones with significantly different read counts between the alleles. Here, we used 32,252 ASB events and 79,827 non-ASB events across 81 TFs, as reported in Wagih et al. (2018), to compare the performance of our OLS-based models versus existing models of TF binding specificity (see 'Results' section).

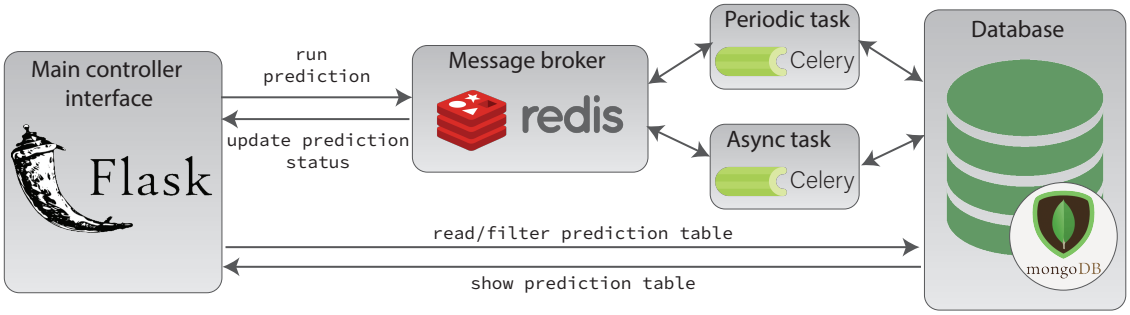



**Figure 3.2**: **QBiC-pred web server architecture** QBiC is built using Flask web framework. The web-server uses Redis as a message broker to relay the prediction job to the Celery task queue. Celery executes the prediction job either as a periodic task (e.g. table initiation and deletion) or as an asynchronous task (e.g. when distributing prediction job between processes). Redis also relays the job status to the user as the percentage of the finished job, which is shown as a progress bar in the website. The resulting prediction table is saved in MongoDB. When user accesses the prediction table, the web-server directly queries MongoDB to obtain the prediction results.

### 3.2.4 QBiC-Pred web server implementation

The QBiC web server was developed using Flask web framework (Figure 3.2) and it runs under Apache 2.4. TF binding change predictions are performed using pre-computed 12-mer tables which contain the predicted TF binding changes, z-scores and P-values for all possible mutations in all possible contexts (please see the http://qbic.genome.duke.edu/about#section4 for details). To perform fast TF binding change predictions for all the input TFs, QBiC uses asynchronous multiprocessing using Celery task queue, where four workers (i.e. processes) are spawned for each request. The input TFs are evenly distributed among the four workers and each worker asynchronously perform binding change predictions for the input mutation data to its assigned TFs.

While the predictions are being calculated, users can leave the QBiC website and return to the job later using the link provided in the 'Recent Jobs' dropdown menu. Importantly, the time needed to execute a prediction job depends mostly on the

QBiC-Pred

Home | Make Predictions | Upload Results | Models | About | Downloads | Recent jobs

**File input:**
QBiC-icgc-example-breast-cancer-mutations-small.tsv

**Running mode:**
top 1 largest absolute z-score

**Genome version:**
hg19

**E-score threshold (specific / nonspecific):**
0.4 / 0.35

TFs used

**Filters:**

TF_gene: has FOXA1 x

binding_status: has unbound>bound x

Show 10 entries

Enter search keyworc | in sequence | Filter

| Index | Ref | Alt | Difference | z score | p value | TF gene |
|---|---|---|---|---|---|---|
| 22 | AAAAAAAAAAAT | AAAAACAAAAT | 0.39951 | 5.5110 | 3.567e-08 | FOXA1 |
| 31 | TTCTTGGTATA | TTCTTTGTATA | 0.51549 | 10.8891 | 1.299e-27 | FOXA1 |
| 547 | TGATACCAAAA | TGATAACAAAA | 0.74275 | 15.7801 | 4.268e-56 | FOXA1 |
| 662 | AATACTTAAGA | AATACATAAGA | 0.43263 | 8.8861 | 6.331e-19 | FOXA1 |
| 668 | CAAATGACGAG | CAAATAACGAG | 0.86226 | 18.0378 | 9.842e-73 | FOXA1 |
| 820 | AGTACCCAGGA | AGTACACAGGA | 0.81829 | 17.4146 | 6.390e-68 | FOXA1 |
| 953 | AGTGTCTTTAA | AGTGTATTTAA | 0.98091 | 20.4499 | 6.025e-93 | FOXA1 |
| 955 | CTTATCATTTC | CTTATTATTTC | 0.46195 | 9.9772 | 1.917e-23 | FOXA1 |

Showing 1 to 8 of 8 entries (filtered from 245 total entries) Previous 1 Next

Download: CSV / TSV

**Figure 3.3**: **QBiC-Pred web server results page** for a sample mutation file containing ICGC breast cancer mutation data (the example use case 'ICGC Breast Cancer Mutations - Small', available in QBiC). Output results were filtered to include only the FOXA1 transcription factor, and only mutations that create TF binding sites, i.e. 'unbound>bound' mutations.

number of TFs selected as input, as each worker needs to read into memory the 12-mer table corresponding to each TF. Adding more variants to the input mutation/variant file will have an almost negligible impact on the processing time. The prediction results are saved in MongoDB for 2 days; during this time the user can access the results using a unique job identifier. The results are displayed in an interactive table format with filtering capabilities (Figure 3.3). Users can also download the prediction results (as csv or tsv files) and re-upload them later, even after the job expired, for further processing within the QBiC website.

## 3.3 Results

In previous work we showed that our OLS model-based predictions of TF binding changes due to DNA mutations correlate well with measured changes in gene expression (Zhao et al., 2017). We also analyzed a large set of pathogenic non-coding variants, showing that these variants lead to more significant differences in TF binding between alleles, compared to common variants, which indicates that there is a strong regulatory component to pathogenic non-coding variants. Here, we complement our previous evaluations of the OLS models by assessing their accuracy in predicting *in vitro* and *in vivo* TF binding changes, and by comparing our OLS models to PWMs and deep learning models of TF binding specificity.

### 3.3.1 OLS performs better than Elastic Net in predicting *in vitro* TF binding changes

OLS is simple and fast to compute since there is no parameter tuning involved in training and testing the models. However, OLS is prone to overfitting, especially when there are correlations among variables and when the number of parameters is larger than the number of samples. In these cases, regularization is often used. Regularization can help in adjusting bias-variance trade-off by reducing variance from known data (i.e. prevent overfit) and increasing bias (i.e. prevent underfit) so the model can generalize better. Using regularized regression, we want to test if we can improve the accuracy of our predictions.

For the regularized regression models, we chose Elastic Net (Zou and Hastie, 2005):

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \alpha(\lambda\|\boldsymbol{\beta}\|_1) + (1-\alpha)(\lambda\|\boldsymbol{\beta}\|_2^2),\ \alpha \in [0,1], \lambda \geq 0 \tag{3.1}$$

where $\alpha$ is the Elastic Net coefficient. Elastic Net combines the characteristic of L1 and L2 regularizations. When $\alpha = 1$, it performs L1 (i.e. lasso) regression and when setting $\alpha = 0$, it performs L2 (i.e. ridge) regression. Thus, by carefully tuning $\lambda$ and $\alpha$ we can get a penalty that takes into account both L1 and L2 regularizations. It is also important to note that Elastic Net with $\lambda = 0$ is the same as OLS.

We tuned Elastic Net for parameters $\lambda$ and $\alpha$ by using 90% data to train and the rest 10% to test. We did not use cross validation since cross validation takes much longer to run, especially since we have hundreds of TFs to test. We believe that the simple train-test split can be used to get insights on whether the model tested can possibly perform better. We also tested for the larger value of $k - mer$ using Elastic Net. In our previous study (Zhao et al., 2017), we showed that $k = 7$ results in overfitting for OLS. Using Elastic net, we want to test whether regularization can help in solving the overfitting we encountered in OLS and whether it can result in better prediction accuracy than OLS.

We found that OLS performs as well as the Elastic Net models (Figure 3.4). Using 7-mer features, Elastic Net is able to eliminate the overfitting we encountered in OLS models (Zhao et al., 2017), however the testing correlation is similar with the OLS models. The Elastic Net 8-mer shows lower testing correlation (despite higher training correlation) than OLS, which suggests overfitting. Taking into account the complexity required to tune the parameters for Elastic Net and its similar accuracy with OLS, we concluded that using simple OLS models is still the best method in predicting *in vitro* TF binding changes.

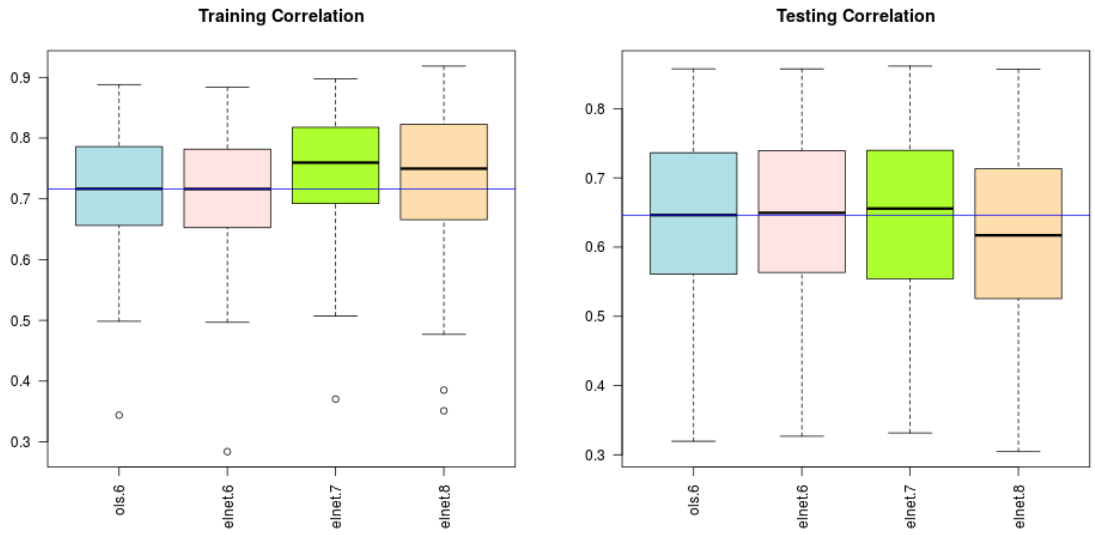



**Figure 3.4**: Pearson correlation for training (left) and testing (right) comparing OLS 6-mer and Elastic Net models. The number $k$ after *elnet* denotes the value of $k-mer$ used to train the Elastic Net models.

### 3.3.2 OLS models of TF binding specificity outperform PWMs and DeepBind models in predicting *in vitro* TF binding changes

As described in Materials and Methods, we designed custom DNA libraries for PBM experiments to test the effects of all single nucleotide variants within binding sites of six human TFs. We used the log ratio of the binding intensity between a mutant and its corresponding wild-type site to represent the TF binding change. Next, we made predictions of these binding changes using six types of models: OLS models, PWM models used in Wagih et al. (2018), PWM-based sTRAP models (Thomas-Chollier et al., 2011), and DeepBind models (Alipanahi et al., 2015) trained on *in vivo* ChIP-seq data, *in vitro* HT-SELEX data, and *in vitro* uPBM data. The uPBM data sets used to train DeepBind and OLS models were the same. The PWMs were obtained from the JASPAR (Khan et al., 2018)

and HOCOMOCO (Kulakovskiy et al., 2017) databases. For TFs with multiple PWMs available, the results we report below are for the PWM that performed best in our evaluation (ETS1: HOCOMOCO ETS1_HUMAN.H11MO.0.A, ELK1: HOCOMOCO ELK1_HUMAN.H11MO.0.B, GATA1: JASPAR MA0035.2, CREB1: JASPAR MA0018.2, RUNX1: JASPAR MA0002.2, STAT3: HOCOMOCO STAT3_HUMAN.H11MO.0.A). For DeepBind ChIP-seq and SELEX models, we used the v0.11 tools made available for download by the authors (Alipanahi et al., 2015). For DeepBind PBM models, the authors kindly provided assistance training the models on our uPBM data.

OLS models can directly predict the TF binding change due to a variant in a fixed-length or variable-length sequence. In contrast, for PWM and DeepBind models we computed likelihood scores for the wild-type and mutant sequences, based on fixed-length window scores. For these models, we predicted the binding change as the difference between the maximum of all wild-type window scores and the maximum of all mutant window scores. This definition is the same as delta track metric defined in Wagih et al. (2018), which performed best in their study.

The correlations between model predictions and the TF binding changes measured using custom PBM experiments across the six TFs are shown in Figure 3.5. Except for RUNX1, for which the DeepBind SELEX model was slightly better than the rest of the models, DeepBind PBM models and our OLS models outperformed the other models in predicting TF binding changes *in vitro*. Compared to DeepBind PBM models, our OLS models are simpler and much faster for training and for predictions. In addition, OLS models can be used to assess the statistical significance of the TF binding changes predicted for each variant.

Figure 3.6 shows a detailed comparison of five models (OLS, PWM, sTRAP, DeepBind SELEX, and DeepBind PBM) for a binding site of TF ELK1. The input

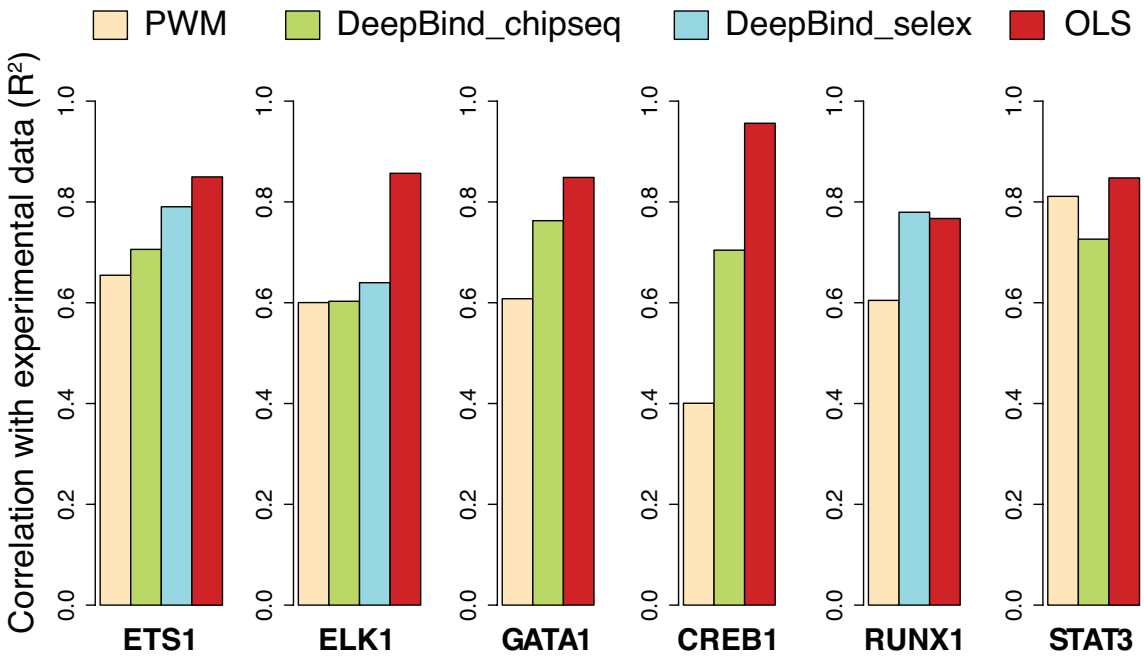



**Figure 3.5**: **Performance of OLS models in predicting *in vitro* TF binding changes, compared to PWM and DeepBind models.** When multiple models are available for a TF, we choose the one that gives the best prediction result. We note that DeepBind ChIP-seq models are not available for RUNX1, and DeepBind SELEX models are not available for GATA1, CREB1 and STAT3.

mutation file used in QBiC to generate the ELK1 binding change predictions shown in Figure 3.6 can be downloaded from the QBiC website as the sample input file in sequence format.

### 3.3.3 The cross-validation accuracy of OLS models correlates with their accuracy in predicting *in vitro* TF binding changes

A TF can have multiple PWM models and DeepBind models available, and it is often difficult to choose which model to use for prediction. In contrast, for our OLS-based approach, we are able to rank the models based on cross-validation accuracy

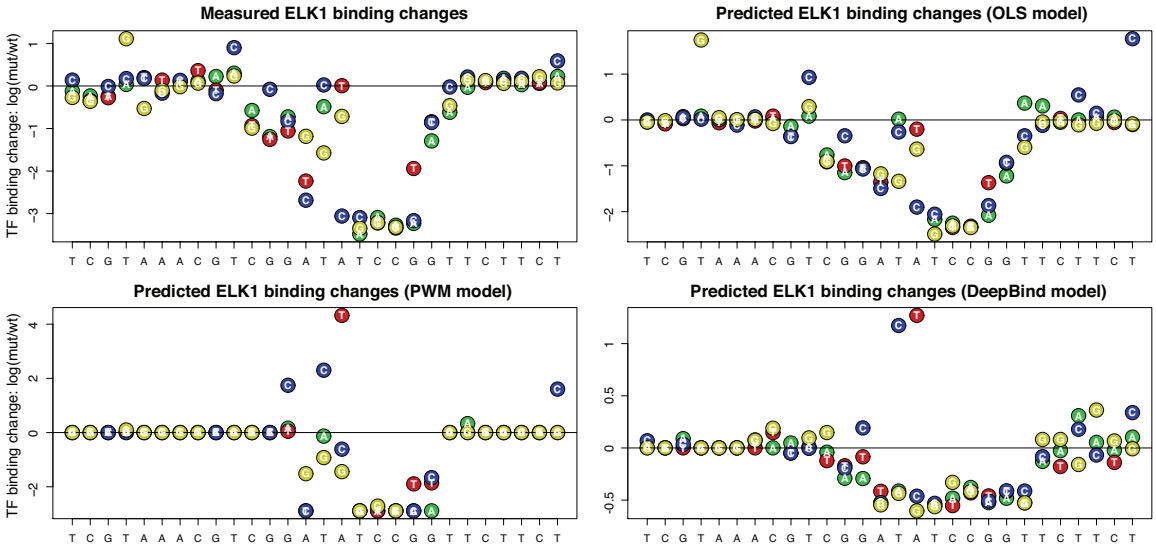



**Figure 3.6**: **Measured and predicted effects of single nucleotide mutations in an ELK1 binding site and its flanking regions.** Since the wild-type sequence contains an ELK1 binding site, most of the variants decrease binding. The A to T mutation in the middle generates a perfect match to core ELK1 motif TTCC. This, however, does not increase the binding signal, likely because the flanking regions already made the ATCC site very strong. Both the PWM and DeepBind models incorrectly predict a dramatic increase in binding due to the A to T mutation. The OLS model, however, correctly predicts the TF binding to be nearly unchanged.

on the uPBM training data set. As expected, we found that there is a positive relationship between the in-sample cross-validation accuracy and the TF binding change prediction accuracy on independent *in vitro* data (Figure 3.7). Thus, when a TF has multiple OLS models, we recommend choosing the model with the highest cross-validation accuracy. Detailed information on the available OLS models for each human TF can be found in the About section of the QBiC website.

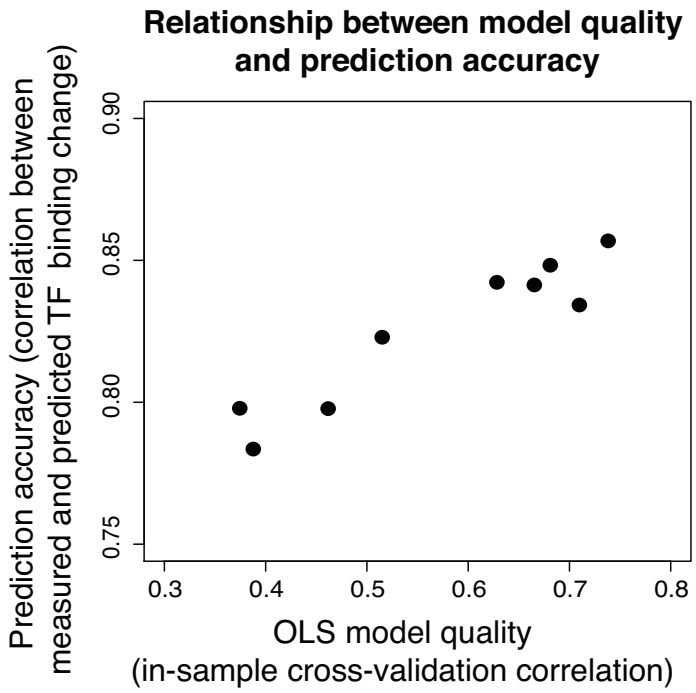



**Figure 3.7**: **Relationship between OLS model quality and the prediction accuracy on independent *in vitro* mutation data.** The OLS model quality is assessed as the in-sample cross-validation correlation. Figure shows the performance of OLS models trained on nine different uPBM data sets for TF ELK1.

### 3.3.4 OLS models of TF binding specificity outperform PWMs and DeepBind models in predicting *in vivo* allele-specific binding variants

To test the performance of OLS models on *in vivo* data, we used the allele-specific binding (ASB) and non-ASB variants in Wagih et al. (2018). We compared the performance of OLS models, PWM models, and DeepBind models in distinguishing ASB variants from non-ASB variants. The performance of each model was assessed using the area under the Receiver Operating Characteristic curve (AUROC) measure. For PWMs and DeepBind ChIP-seq models, we used the binding change scores reported by Wagih et al. (2018). For DeepBind SELEX and PBM models we derived the

binding change scores similarly to Wagih et al. (2018), and used them for the classification. For OLS models we used the z-score outputs to classify the variants. The DeepBind PBM and OLS models were trained on the same sets of PBM data. To illustrate how QBiC can be used to analyze ASB and non-ASB variants, we provide the input mutation file corresponding to the ASB data on the QBiC website, as the sample input file for the VCF format.

A total of 14 human TFs have PWM models, OLS models, and DeepBind models available. For these TFs we divided their ASB variants into gain-of-binding and loss-of-binding variants (for which the TF binding changes have opposite signs), and for each set we used the different TF binding models to distinguish between ASB and non-ASB variants. OLS models clearly outperformed PWMs (Figure 3.8A), which was expected given the limitations of PWM models in capturing TF binding specificity (Shen et al., 2018; Zhou et al., 2015; Siggers and Gordân, 2014). OLS models also outperformed DeepBind SELEX models trained on *in vitro* binding data from HT-SELEX experiments (Figure 3.8B) and DeepBind PBM models trained on *in vitro* data from PBM experiments (Figure 3.8C) demonstrating that, when using only DNA sequence information for training, OLS models perform best in predicting *in vivo* allele-specific binding variants.

We also compared the performance of OLS models to DeepBind models trained on *in vivo* ChIP-seq data (Figure 3.8D). Using OLS models we obtained larger AUROC values for about half of the TFs, and overall the two models had similar power in distinguishing ASB from non-ASB variants. Nevertheless, we note that the DeepBind ChIP-seq models were trained on ChIP-seq data from the same cell type as the ChIP-seq data from which the ASB variants were called. Therefore, OLS models managed to reach similar performance to models trained on the ChIP-seq data itself, despite the fact that OLS models do not use any cell type specific information.

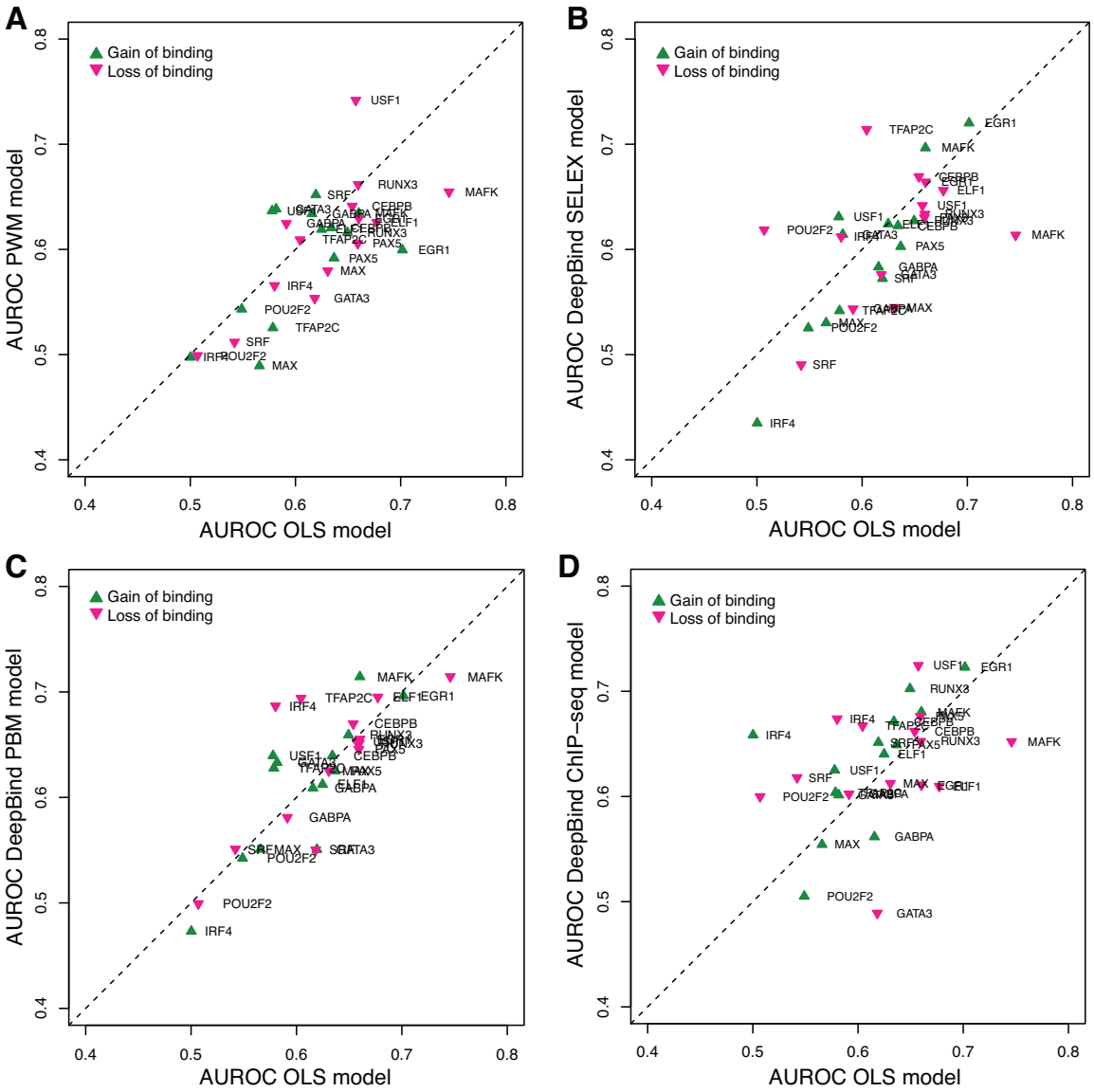



**Figure 3.8**: **Performance of OLS, DeepBind, and PWM models in distinguishing between ASB and non-ASB variants**. The variants were identified from *in vivo* ChIP-seq data.

## 3.4 Discussion

Quantitative predictions of TF binding changes can help us understand the functional roles of genetic variants, and prioritize variants that are likely to have regulatory effects. QBiC-Pred provides a fast and accurate approach to predict TF binding

changes due to genetic variants, based solely on their sequence context. QBiC-Pred models are trained on *in vitro* high-throughput universal PBM data, and they outperform current PWM-based models and DeepBind models, which are also based mainly on DNA sequence information. In addition, QBiC-Pred offers a way to statistically test the significance of each variant, taking the quality of the predictive models into account. The quality measure of the models also helps circumvent the problem of deciding which model to use when multiple models are available, which is often encountered when making predictions using PWMs.

Several recent methods, including Sasquatch (Schwessinger et al., 2017), DeepSEA (Zhou and Troyanskaya, 2015), and deltaSVM (Lee et al., 2015), predict the impact of non-coding variants by taking advantage of cell- and tissue-specific information, oftentimes beyond TF binding data. These methods are complementary to ours: they focus on overall functional changes caused by non-coding variants, while we examine more specifically the potential binding changes for each individual TF. For example, Sasquatch predicts the change in the DNase footprint due to a variant, but does not directly pinpoint the binding of which TF(s) is affected by the variant (unless one post-processes the results using specific TF binding models). In contrast, QBiC-Pred can make quantitative predictions in a TF-specific manner, for a large number of TFs, although it cannot predict the effect of the variant in any specific cell type. Using these methods together would give us a better understanding of the functional impact of non-coding variants in the cell.

Annotation-based methods such as rVarBase Guo et al. (2016), INFERNO Amlie-Wolf et al. (2018), HaploReg Ward and Kellis (2015), and RegulomeDB Boyle et al. (2012) can also be used to investigate potential regulatory variants. These methods test whether the input variants fall within known regulatory regions annotated, for example, using PWM models and cell type-specific data. Thus, predictions made by

annotation-based methods depend on the quality of the existing annotations, and, in the case of TF binding sites, these methods are unlikely to detect variants that lead to the creation of new binding sites in the genome. In addition, we note that none of the methods mentioned above provides a direct measure of the confidence in the predicted changes in TF binding, based on the quality of the binding data and model, which is a distinguishing feature of QBiCPred.

In summary, QBiC-Pred (Martin et al., 2019) uses OLS models of TF-DNA binding specificity to make accurate predictions of TF binding changes due to single nucleotide variants. In addition to the current functionalities of QBiC-Pred, a natural extension would be to allow input sequences containing multiple variants. As shown in our previous work, OLS models perform very well on data containing multiple variants, being able to predict $\sim 50\%$ of the resulting variation in gene expression Zhao et al. (2017). Another extension would be to include models trained on other types of high-throughput *in vitro* TF binding data, such as HT-SELEX data Jolma et al. (2010, 2013). This would extend the list of human TFs that can be analyzed using QBiC-Pred beyond the 582 TFs with available high-quality uPBM data. This extension, however, will require the development of new methodology that takes into account the statistical properties of the HT-SELEX data, in order to allow us to use the data directly to compute significance levels ($P$-values) reflecting our confidence in the predicted effects of mutations on TF binding.

# Chapter 4

# Transcription Factor-Centric Approach to Identify Non-Recurring Putative Regulatory Drivers in Cancer

## 4.1 Introduction

Studies of somatic mutations in cancer genomes are generally focused on mutations that alter the amino acid sequences of protein-coding genes. However, whole-genome sequencing of human tumors has revealed that the vast majority of somatic mutations in cancer are non-coding (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), suggesting that they could play a role in cancer initiation and development. Tumorigenesis is thought to be due to the accumulation of multiple driver mutations that confer a growth advantage to the tumor cells; some of these driver mutations may be non-coding (Khurana et al., 2016). But only a small proportion of the mutations present in cancer cells are drivers, so it is important to accurately identify them and distinguish them from the much larger number of passenger mutations (Elliott and Larsson, 2021).

Given that driver mutations are expected to be under positive selection, their identification is generally based on patterns of recurrence among tumor samples. Several recent studies have attempted to discover non-coding driver mutations in regulatory DNA sites using recurrence information (e.g. Lochovsky et al., 2015, 2018; Rheinbay et al., 2017; Weinhold et al., 2014). Such studies usually involve the identification

of genomic regions with high mutational frequency (i.e. hotspots) by comparing the mutation rate within a DNA window to a background distribution. However, it is generally challenging to precisely estimate the background mutation rate in small genomic regions, given the heterogeneity across different patients and across the genome (Lawrence et al., 2013). A recent meta-analysis of methods for predicting regulatory driver mutations reported that hotspot-based methods can generate large sets of candidate drivers, many of which are false positives (Rheinbay et al., 2020). To narrow down the list of candidates, one can also incorporate information on the functional impacts of putative non-coding driver mutations, in particular their effect on transcription factor (TF) binding. One of the most widely used approaches to prioritize mutations in regulatory regions involves the identification of TF binding sites created or disrupted by the mutations, which can be predicted using position weight matrices (PWMs) and motif prediction algorithms (Link et al., 2018; Shen et al., 2020). However, such methods are limited by the high false positive and false negative rates of binding site prediction algorithms.

In addition, Rheinbay et al. (2020) have recently reported that non-coding regulatory driver mutations are much less frequent than protein-coding drivers, with the only notable exception being driver mutations in the *TERT* gene promoter (Horn et al., 2013; Huang et al., 2013). Moreover, some non-coding drivers identified in previous studies were found to be the result of poorly-understood localized hypermutation processes such as mutations originating from differential DNA damage (Buisson et al., 2019) or differential DNA repair (Mas-Ponte and Supek, 2020; Sabarinathan et al., 2016; Perera et al., 2016). On the other hand, recent studies of cancer drivers have shown that mutations do *not* have to be highly recurrent in order to be true drivers (Kim et al., 2016); in fact, even mutations that occur in individual tumor samples can drive tumorigenesis.

Here, we describe a new method for analyzing non-coding cancer mutations in regulatory genomic regions (i.e. promoters and enhancers) with the goal of prioritizing mutations based on their effects on TF binding. Unlike previous methods for prioritizing putative non-coding drivers, our method does not rely on the recurrence of mutations across tumor samples. Instead, we consider mutations to be potentially 'significant' if they lead to larger changes in TF binding affinity than expected by chance in that particular genomic region. Thus, the magnitude of the mutations' effects, rather than their recurrence, is the basis for prioritizing mutations and regulatory regions for further studies.

To predict the quantitative effects of non-coding variants on TF binding we use QBiC-Pred (Martin et al., 2019), a computational method based on regression models of TF-DNA binding specificity trained on high-throughput *in vitro* data (Zhao et al., 2017). We focus on single-nucleotide mutations, since they are the dominant type of somatic mutation identified in cancer genomes (Khurana et al., 2016). Importantly, our method links enhancers and promoters to the genes they are likely to regulate, and it combines evidence from all regulatory regions of each gene in order to infer whether a gene is potentially dysregulated due to non-coding mutations. Finally, we use gene expression data from donors with versus without mutations in promoters and enhancers in order to validate that our method prioritizes biologically relevant mutations and regulatory regions.

## 4.2 Materials and methods

### 4.2.1 ICGC simple somatic mutations and gene expression data

To develop and test our new method for prioritizing putative regulatory driver mutations, we used the Liver Cancer-RIKEN, Japan (LIRI-JP) project from the International Cancer Genome Consortium (ICGC) (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). We chose this project because it has a large number of donors with whole-genome simple somatic mutation (SSM) data (258 donors), as well as gene expression data (RNA-seq) for 230 out of the 258 donors with SSM data. The study reported a total of ∼3.8 million mutations, most of which are single nucleotide mutations (∼3.5 million). We used these mutations in our analyses.

### 4.2.2 Promoter and enhancer data

We focused our analyses on mutations within promoter and enhancer regions, as TF binding sites are located in these regions. We defined promoters as the genomic sequences within +/- 1000 bp of each RefSeq (O'Leary et al., 2016) transcription start site (TSS), excluding any RefSeq exon sequences. We focused on promoters of protein-coding genes, using only TSSs that map to genes within the HUGO gene nomenclature (HGNC) (Tweedie et al., 2021). These criteria resulted in a set of 21,543 promoters.

For enhancers, we used the experimentally determined enhancers from the FANTOM5 project (Andersson et al., 2014; Lizio et al., 2015), which are frequently used in studies of non-coding mutations (e.g. Weinhold et al., 2014; Khurana et al., 2016). Importantly, the FANTOM consortium provides information about the linkage be-

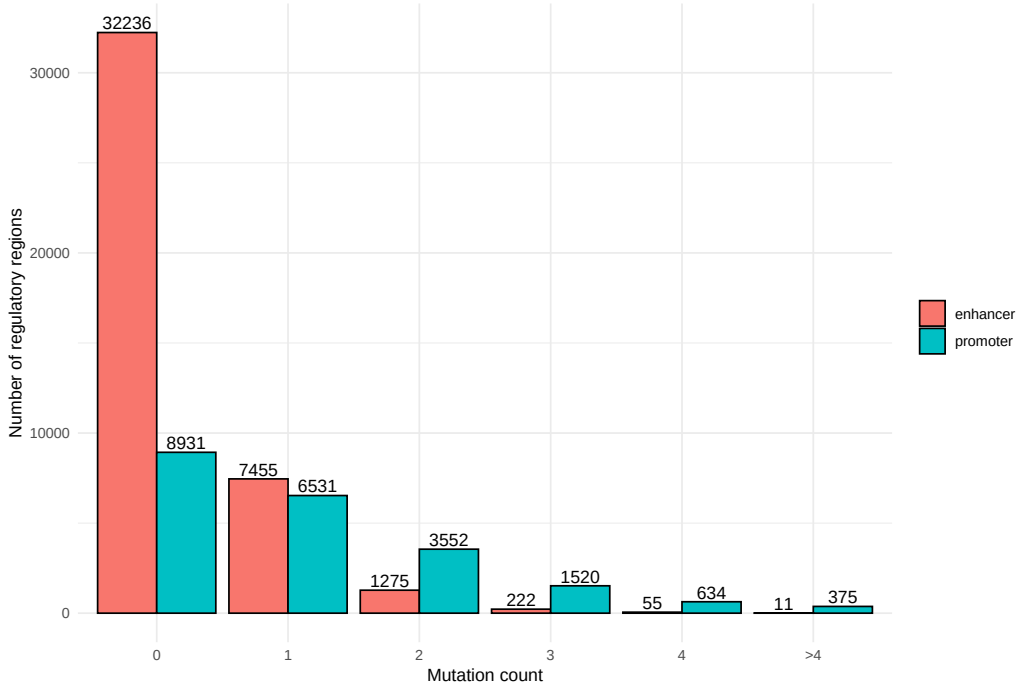



**Figure 4.1**: **Number of mutations in regulatory regions.** The average mutation rate is $3.6 \times 10^{-6}$ in enhancers and $2.1 \times 10^{-6}$ in promoters.

tween enhancers and associated TSSs, which is critical for being able to connect our enhancer results to gene expression data. After removing enhancers that overlapped with promoter regions, we obtained a total of 41,254 enhancers.

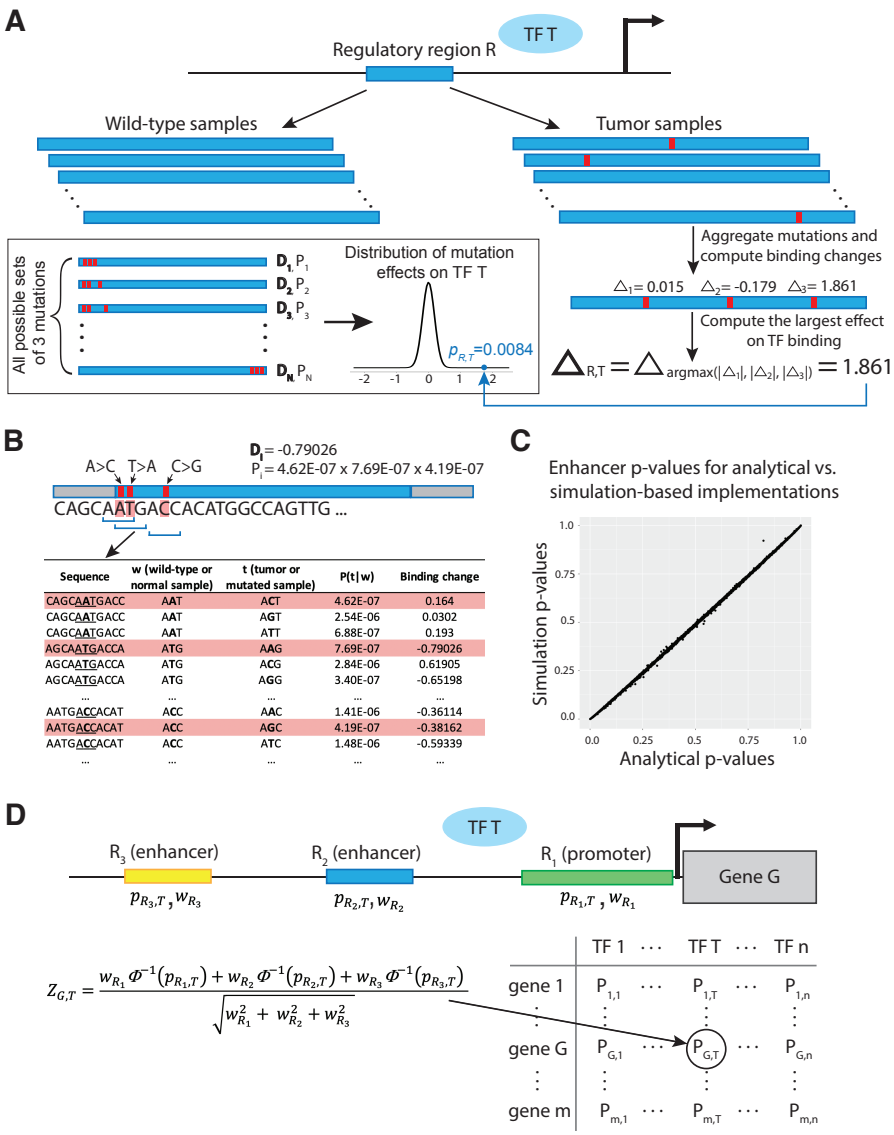
We further filtered the 21,543 promoters and the 41,254 enhancers to keep only those that contain mutations. Since most enhancers are hundreds of base pairs long (median=254bp) and promoters are ∼2,000 bp, the majority of enhancers and a fair number of promoters do not contain any mutations (Figure 4.1). Therefore, after removing these regulatory regions without mutations, we obtained a set of 12,612 promoters and 9,018 enhancers with mutations in the LIRI-JP study.

### 4.2.3 Defining the effects of mutations on TF binding, and the significance of these effects

We use TF binding changes to prioritize mutations that might act as drivers within each regulatory region (enhancer or promoter). For each region, we ask whether the mutations detected in that region lead to *larger* TF binding changes than expected by chance, i.e. according to a background model of random mutations (Figure 4.2A). To assess the effect of non-coding mutations on TF-DNA binding we use QBiC-Pred (Martin et al., 2019), a method we recently developed to quantify TF binding changes based on regression models trained on high-throughput *in vitro* binding data. While other binding specificity models can be used to predict the effects of mutations on TF binding, here we use QBiC-Pred because it performed better than methods based on position weight matrix or deep learning models of specificity (Martin et al., 2019; Zhao et al., 2017).

For each mutation of interest, QBiC-Pred reports its predicted effect on the binding specificity of 582 human TFs, based on models derived from 667 universal protein binding microarray (PBM) data sets. The effect of a mutation $m$ on TF $T$ is reported in terms of the difference ($\Delta_m$) in the logarithm of the PBM binding intensity signal for the mutated sequence relative to the wild-type sequence according to the binding model for TF $T$, as described in detail in (Martin et al., 2019). Positive values represent increased TF binding, while negative values represent decreased binding. Although here we focus on binding changes predicted with QBiC-Pred, our approach can directly use other binding specificity models, as long as they accurately reflect the quantitative TF binding changes induced by DNA mutations.

For a transcription factor $T$ and a regulatory region $R$ that has one or more mutations in the data set of interest, we compute the largest effect on TF binding




| Sequence | w (wild-type or normal sample) | t (tumor or mutated sample) | P(t\|w) | Binding change |
|---|---|---|---|---|
| CAGCAATGACC | AAT | ACT | 4.62E-07 | 0.164 |
| CAGCAATGACC | AAT | AGT | 2.54E-06 | 0.0302 |
| CAGCAATGACC | AAT | ATT | 6.88E-07 | 0.193 |
| AGCAATGACCA | ATG | AAG | 7.69E-07 | -0.79026 |
| AGCAATGACCA | ATG | ACG | 2.84E-06 | 0.61905 |
| AGCAATGACCA | ATG | AGG | 3.40E-07 | -0.65198 |
| ... | ... | ... | ... | ... |
| AATGACCACAT | ACC | AAC | 1.41E-06 | -0.36114 |
| AATGACCACAT | ACC | AGC | 4.19E-07 | -0.38162 |
| AATGACCACAT | ACC | ATC | 1.48E-06 | -0.59339 |
| ... | ... | ... | ... | ... |

**Figure 4.2**: **TF-centric approach to prioritize genes based on mutations in their regulatory regions.** (A) For each regulatory region $R$ and TF $T$, we aggregate the mutations across all patients in the cohort of interest, and compute the largest effect (either positive or negative) across all mutations. This effect, $\mathbf{\Delta}_{R,T}$, is then compared to the distribution of effects computed for all possible sets of $k$ mutations, where $k$ is the number of actual mutations observed in region $R$ (here, $k = 3$). (B) For any set $i$ of $k$ mutations, we compute the binding change with the largest magnitude among these $k$ mutations ($\mathbf{D}_i$), and the probability of that set of $k$ mutations ($P_i$), as described in Methods. (C) Comparison of p-values computed for mutation effects on MYC binding, for 9,018 enhancers. Plot shows the high correlation between p-values calculated using the analytical versus the simulation-based approach. (D) For genes with multiple regulatory regions, we compute the combined significance of the TF binding changes for TF $T$ across all regions $R_i$ by combining their p-values $p_{R_i,T}$ using Liptak's method, also known as weighted Stouffer's method, as described in Methods. We then use the combined p-values, adjusted for multiple hypothesis testing, to rank genes according to the smallest p-value across all TFs.

(either positive or negative) over all mutations in $R$ ($\mathbf{\Delta}_{R,T}$, Figure 4.2A). Next, to determine if this effect is significant, we compare it against the distribution of effects expected by chance, according to a background model that takes into account: 1) the mutation spectra in that particular cohort, and 2) the particular DNA sequence of regulatory region $R$. Conceptually, if there are $k$ total mutations in region $R$ ($k = 3$ in Figure 4.2A), the full distribution of possible binding effects will be computed taking into account all possible sets of $k$ mutations across the region. Each set $i$ of $k$ mutations will have a particular effect on TF binding ($\mathbf{D_i}$) and will occur with a particular probability ($P_i$) (Figure 4.2B). The effect $\mathbf{D_i}$ is computed by taking the largest effect over the $k$ mutations, similarly to the case of the real mutations. The probability $P_i$ of a particular set of $k$ mutations is computed as described below.

Using all single-nucleotide somatic mutations reported in the LIRI-JP study, we computed the mutation spectra for this cohort, analyzing mutations in their trinucleotide contexts (Alexandrov et al., 2020). As in previous studies (Alexandrov et al., 2020; Jusakul et al., 2017), we consider 6 mutation types (C>A, C>G, C>T, T>A, T>C, T>G) each in 16 possible contexts, based on the nucleotide right before and right after the mutated position. Mutations and their reverse complements (e.g. A**T**G>A**C**G and C**A**T>C**G**T) are counted together. For each trinucleotide, we estimate the probability of mutating the central base by taking the ratio between how many times that base was mutated in that context and how many times the wild-type trinucleotide occurs in regulatory regions (enhancers or promoters) across all normal samples. For example, the probability of mutating T in an A**T**G context in enhancers is:

$$P(t \neq w | w = A\mathbf{T}G) = \frac{\#_of_A\mathbf{T}G_to_A\mathbf{V}G_mutations_in_enhancers}{\#_of_A\mathbf{T}G_in_enhancers_in_all_normal_samples} \quad (4.1)$$

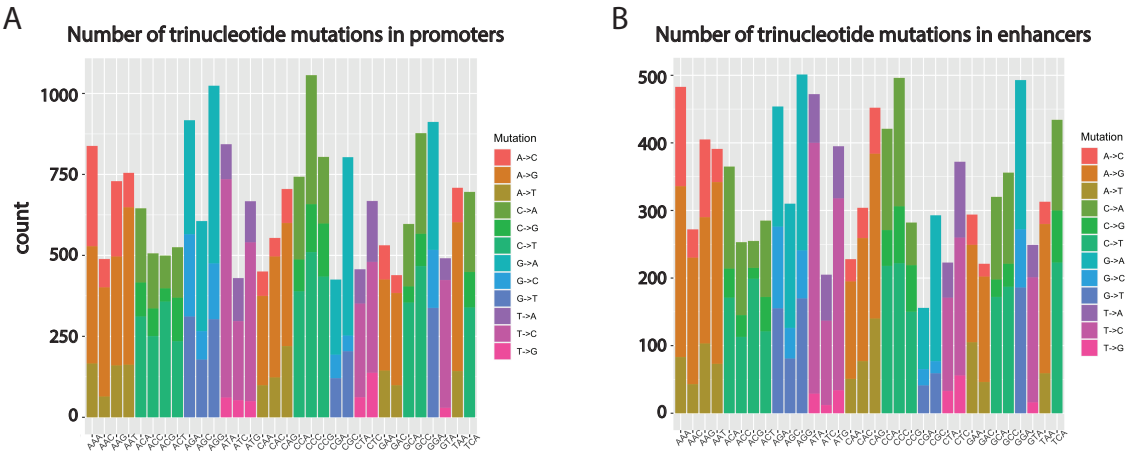



**Figure 4.3**: **Number of mutations in regulatory elements.** (A) promoters and (B) enhancers, computed for each mutation type in each trinucleotide context.

where $t$ is the trinucleotide in the tumor sample, $w$ is the trinucleotide in the corresponding normal sample, and V is the IUPAC code for {A, C, G} (i.e. not T). We calculate mutation probabilities separately for enhancers versus promoters, since they may be affected differently by mutagenic processes; indeed, we saw significant differences between the mutation spectra at enhancers versus promoters (Figure 4.3).

When the central nucleotide of a trimer is mutated, i.e. $t \neq w$, there are three possible mutations, e.g. A**T**G to A**A**G, A**C**G, or A**G**G. We estimated the probability of each mutation type (e.g. A**T**G to A**A**G), given that a mutation exists at the central nucleotide, as the number of times we observed that particular mutation type divided by the total number of times the central nucleotide was mutated in that context, in the regulatory regions of interest. For example, focusing on enhancer regions, we compute:

$$P(t = A\mathbf{A}G | w = A\mathbf{T}G, w \neq t) = \frac{\#_of_A\mathbf{T}G_to_A\mathbf{A}G_mutations_in_enhancers}{\#_of_A\mathbf{T}G_to_A\mathbf{V}G_mutations_in_enhancers} \tag{4.2}$$

Next, to calculate the probability of a particular mutation in a particular trinucleotide context, we multiply the probability that the trinucleotide is mutated with the probability of the specific mutation, e.g.:

$$P(t = A\mathbf{A}G|w = A\mathbf{T}G) = P(w \neq t|w = A\mathbf{T}G) \times P(t = A\mathbf{A}G|w = A\mathbf{T}G, w \neq t) \tag{4.3}$$

which can be simplified to the number of A**T**G to A**A**G mutations in enhancers divided by the total number of A**T**G trinucleotides in enhancers across all normal samples.

We note that the $k$ mutations aggregated over region $R$ are typically from different samples in our cohort, and can thus be considered independent. Therefore, for a set $i$ of $k$ mutations we compute the overall probability $P_i$ of that particular set by multiplying the individual probabilities $P(t|w)$ for each of the $k$ mutations, as illustrated in Figure 4.2B. Finally, to assess the significance of $\mathbf{\Delta}_{R,T}$, we compare this value against the distribution of effects for random sets of $k$ mutations in region $R$, with the p-values being computed efficiently from $P_i$ and $\mathbf{D}_i$ values using either an analytical or a simulation-based approach, as described below.

### 4.2.4 Analytical and simulation-based approaches to compute the significance of mutation effects on TF binding

Our simulation-based approach uses the mutation probabilities described above to repeatedly sample mutations in the regulatory region of interest. Let us consider a regulatory region $R$ of length $L$ that contains $k$ mutations across all patients in our cohort. Since there are 3 possible mutations for each position in $R$, we have a total

of $3L$ mutations to consider in this region. At each iteration of our simulation-based approach, we randomly sample $k$ out of the $3L$ possible mutations, with replacement. We do the sampling with replacement because the exact same mutation can occur in two or more patients. For this sampling process, the probability of selecting a particular set $i$ of $k$ mutations ($P_i$) is computed by multiplying the probabilities of the $k$ mutations (Figure 4.2B). The TF binding change for this set of mutations ($\mathbf{D}_i$) is computed by taking the maximum effect over the $k$ randomly chosen mutations, as described above and illustrated in Figure 4.2A,B. By repeatedly sampling sets of $k$ mutations using this procedure, we can approximate the distribution of mutation effects on TF binding (Figure 4.2A), and use it to compute empirical p-values for $\mathbf{\Delta}_{R,T}$, taking the sign of the effect into account. This simulation-based approach is simple to understand and implement. However, simulations are time consuming and unfeasible for generating background distributions of mutations effects for all regulatory regions (totalling 12,612 promoters and 9,018 enhancers) and all TFs (totalling 582 TFs with 667 binding models available).

Alternatively, we can use an analytical approach to directly compute the p-value for the effect $\mathbf{\Delta}_{R,T}$. Conceptually, the p-value of interest is the probability of obtaining an effect on TF binding at least as large as $\mathbf{\Delta}_{R,T}$ when we randomly choose $k$ of the $3L$ possible mutations in the regulatory region $R$. For simplicity, let us consider these effects in absolute value. For a set $i$ of $k$ mutations, if *at least one* of the mutations leads to an absolute change in binding of TF $T$ that is $\geq |\mathbf{\Delta}_{R,T}|$, then $|\mathbf{D}_i| \geq |\mathbf{\Delta}_{R,T}|$. On the other hand, if *all* the $k$ mutations lead to absolute binding changes $< |\mathbf{\Delta}_{R,T}|$, then we have $|\mathbf{D}_i| < |\mathbf{\Delta}_{R,T}|$. Thus, focusing on the absolute values

of the effects of mutations on TF binding, we can compute our p-value of interest as:

$$P(|\text{effect of set of } k \text{ mutations}| \geq |\mathbf{\Delta}_{R,T}|) = \tag{4.4}$$
$$1 - P(|\text{effect of random set of } k \text{ mutations}| < |\mathbf{\Delta}_{R,T}|) = 1 - \sum_{\substack{\text{Sets } i \text{ of } k \text{ mutations} \\ s.t. |\mathbf{D_i}| < |\mathbf{\Delta}_{R,T}|}} P_i$$

The total number of possible sets $i$ of $k$ mutations in regulatory region $R$ is $\binom{3L+k-1}{k}$, which is the number of possible unordered outcomes when sampling $k$ out of $3L$ mutations with replacement. Even when choosing only the sets for which all $k$ mutations have absolute binding changes $< |\mathbf{\Delta}_{R,T}|$, the number of possibilities can be very large and not feasible to compute explicitly. To overcome this problem, we compute a vector $\pi = (\pi_1, \pi_2, \ldots, \pi_l)$, $0 \leq l < 3L$, with the probabilities of all individual mutations $m$ in region $R$ for which $|\Delta_m| < |\mathbf{\Delta}_{R,T}|$. The sum in Equation 4.4 can then be written in terms of the vector $\pi$, as the sum of the all elements in the outer product of $\pi$ with itself, taken $k$ times. For example, for $k = 3$, each element $P_i$ in Equation 4.4 is an element of $\pi \otimes \pi \otimes \pi$. Importantly, we do not need to compute this outer product explicitly, as we are only interested in the sum of all elements in the product, which can be written as $(\pi_1 + \pi_2 + \ldots + \pi_l)^k$. Thus, our p-value of interest can be calculated as:

$$P(|\text{effect of a set of } k \text{ mutations}| \geq |\mathbf{\Delta}_{R,T}|) = 1 - \left(\sum_{j=1}^{l} \pi_j\right)^k \tag{4.5}$$

Finally, since we want to take into account the directionality of the effect of mutations on TF binding, i.e. decreased binding ($\mathbf{\Delta}_{R,T} < 0$) or increased binding ($\mathbf{\Delta}_{R,T} > 0$), we calculate the one-sided p-value as $p_{R,T} = (1 - (\sum_{j=1}^{l} \pi_j)^k)/2$.

We confirmed that the results from our analytical and simulation-based approaches agree with each other (Figure 4.2C). For this test, we randomly picked one of the

TFs, MYC, and we calculated the p-values for changes in MYC binding specificity for all 9018 mutated enhancers, using a simulation process with one million iterations. Since the simulation-based approach is more time consuming and has limited precision in estimating the p-values of interest (due to the limited number of iterations), we used the analytical approach for all subsequent analyses.

### 4.2.5 Integrating results across all regulatory regions of a gene

Genes encoded in the human genome typically have multiple regulatory regions (enhancers and promoters); mutations in either of these regions could affect a gene's expression. Thus, it is of interest to integrate the effects of mutations across all regulatory regions of each gene. As detailed above, we define gene promoters based on TSS coordinates in the RefSeq database (O'Leary et al., 2016), and we leverage TSS-enhancer links from the FANTOM consortium (Lizio et al., 2015), considering all the cell types and tissues with available data. In other words, if a genomic region has been identified as an enhancer for gene $G$ in one tissue, then we consider that region as part of the regulatory landscape of gene $G$, in order to be as inclusive as possible. On average, each TSS is associated with 4.9 enhancers according to the FANTOM data. For genes that have multiple TSSs in RefSeq, we consider each TSS separately. Thus, some genes may appear multiple times in our final results, with different p-values that correspond to its different TSSs.

Given a transcription factor $T$ and a gene $G$ with $r$ regulatory regions containing at least one mutation in our cohort of interest, we calculate the significance $p_{R_i,T}$ of the effects ($\mathbf{\Delta}_{R_i,T}$) of mutations in each regulatory region $R_i$ on the binding specificity of TF $T$. Next, we want to integrate these effects over all $r$ regulatory regions by

combining their p-values. Intuitively, our null hypothesis ($H_0$) is that the effects for all regulatory regions ($\mathbf{\Delta}_{R_i,T}$) come simply from the background distribution of effects due to random mutations. The alternative hypothesis ($H_1$) is that at least one regulatory region of gene $G$ has an effect $\mathbf{\Delta}_{R_i,T}$ significantly larger than expected by chance according to the background model of mutations in regulatory regions. Importantly, as the promoter and enhancer regions used in our analysis do not overlap, we can consider the p-values $p_{R_i,T}$ as coming from independent tests.

One approach to combine the $r$ p-values computed for gene $G$ is Fisher's method (Fisher, 1934), which is often used in meta-analyses, including analyses of non-coding mutations in cancer (Rheinbay et al., 2020; Lawrence et al., 2014; Araya et al., 2016). However, Fisher's method would not take into account the fact that different regulatory regions have different lengths and different probabilities of harboring mutations. If a regulatory region $R_i$ is very long, then the number of possible mutations, and thus the number of possible effects on TF $T$, is also large. In comparison, a shorter regulatory region, $R_j$, will have fewer possible mutations and fewer possible effects on TF $T$. If the two regions have the same p-value, i.e. $p_{R_i,T} = p_{R_j,T}$, we might still consider, intuitively, that the mutation effect in $R_i$ is more significant because for $R_i$ it would be easier to achieve a large effect on TF binding, since more mutations can occur in $R_i$ than $R_j$. This is similar to meta-analyses of studies with very different sample sizes, where a weighted version of Stouffer's method, developed by Liptak (Lipták, 1958) and also known as the weighted Z-method or weighted Z-test, was found to be superior to Fisher's method when combining p-values from independent tests (Whitlock, 2005; Zaykin, 2011).

Here, we use Liptak's method (Lipták, 1958) to combine the p-values of all regulatory regions of a gene, with the weights computed based on the mutations probabilities in each region, combined using Shannon's entropy. Specifically, for a regulatory

region $R$ of length $L$ we compute its weight as $w_R = -\sum_{m=1}^{3L} p_m log(p_m)$, where $p_m$ is the probability of the $m^{th}$ possible mutation, computed according to the trinucleotide centered at that position (see Figure 4.2B). We note that weights computed in this manner are overall correlated with the length of the regulatory regions, but avoid giving an out-sized importance to very long regions (Figure 4.4). Thus, for TF $T$ and gene $G$ with $r$ regulatory regions, we compute the weighted test statistic $\sum_{i=1}^{r} w_{R_i} z_{R_i,T}$, where $z_{R_i,T} = \Phi^{-1}(p_{R_i,T})$ and $\Phi^{-1}$ is the inverse of the standard normal cumulative distribution function, as initially proposed by Liptak (Lipták, 1958; van Zwet and Oosterhoff, 1967). Under the null hypothesis ($H_0$), this test statistic follows a normal distribution N(0,$\sum_{i=1}^{r} w_{R_i}$) for any choice of weights (van Zwet and Oosterhoff, 1967; Heard and Rubin-Delanchy, 2018). This allows us to compute the p-value of the combined test, $P_{G,T}$, as follows:

$$Z_{G,T} = \frac{\sum_{i=1}^{r} w_{R_i} z_{R_i,T}}{\sqrt{\sum_{i=1}^{r} w_{R_i}^2}} \text{ and } P_{G,T} = 2 \times (1 - \Phi(|Z_{G,T}|)) \tag{4.6}$$

where $\Phi$ is standard normal cumulative distribution function. Finally, we used the Benjamini-Hochberg correction (Hochberg, 1988) to adjust for multiple hypothesis testing across all genes and all TFs. We then ranked genes according to the smallest p-value across all TFs, and we analyzed the top genes for differences in gene expression. In total, we analyzed 5,336 genes with mutations in enhancers, 11,721 genes with mutations in promoters, and 13,982 genes with mutations in at least one regulatory region (enhancer or promoter).

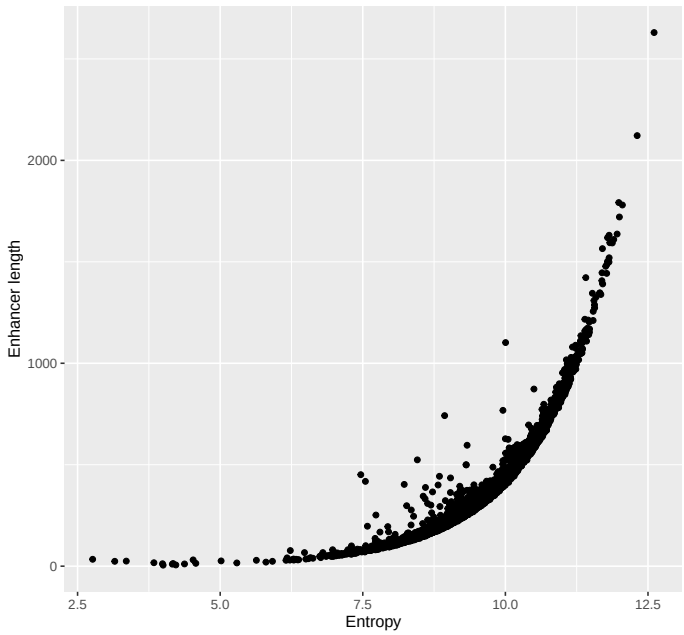



**Figure 4.4**: **Comparison between the entropy computed for each enhancer**, based on the probabilities of all possible mutations, versus the enhancer's length.

## 4.3 Results

### 4.3.1 Integrated analysis across regulatory regions identifies 54 genes with significant TF binding changes due to mutations in regulatory DNA

Using the single nucleotide mutation data from the LIRI-JP study, we identified 13,982 genes with mutations in either the promoter or the enhancer regions. For each of these genes, we analyzed all 582 human TFs for which QBiC-Pred models are available (Martin et al., 2019), and we computed the smallest p-value $P_{G,T}$ across all TFs, adjusted for multiple testing, as described in the Methods section (Figure 4.2). We chose to take the minimum p-value, rather than integrate the p-values across factors, because TF binding specificities are oftentimes highly correlated, especially for closely-related paralogous TFs. Next, we ranked genes according to the smallest

**A**

| gene | count | min_adjusted_p | TF |
|---|---|---|---|
| FDPS | 2 | 3.17E-04 | HBP1 |
| C1R | 4 | 6.00E-04 | CXXC1 |
| KANK4 | 9 | 7.83E-04 | FOXM1 |
| CENPA | 5 | 1.10E-03 | DMRTA2 |
| WFIKKN2 | 2 | 1.21E-03 | LBX1 |
| CCNH | 2 | 1.61E-03 | ELK1 |
| TECR | 2 | 2.05E-03 | HOXA13 |
| ANGPTL6 | 4 | 2.69E-03 | EGR1 |
| A4GALT | 2 | 2.92E-03 | RORA |
| EHBP1 | 3 | 3.14E-03 | TCF12 |
| C10orf67 | 2 | 3.33E-03 | RORA |
| USP45 | 2 | 3.70E-03 | ESR1 |
| C10orf82 | 6 | 3.82E-03 | NR5A1 |
| CHRDL1 | 2 | 3.96E-03 | TERF2 |
| GPN3 | 3 | 3.98E-03 | BARHL1 |
| PPP1R18 | 6 | 4.73E-03 | PAX6 |
| DPM3 | 3 | 5.23E-03 | IRF9 |
| POLM | 3 | 5.83E-03 | KMT2A |
| ⋮ | ⋮ | ⋮ | ⋮ |

**B**

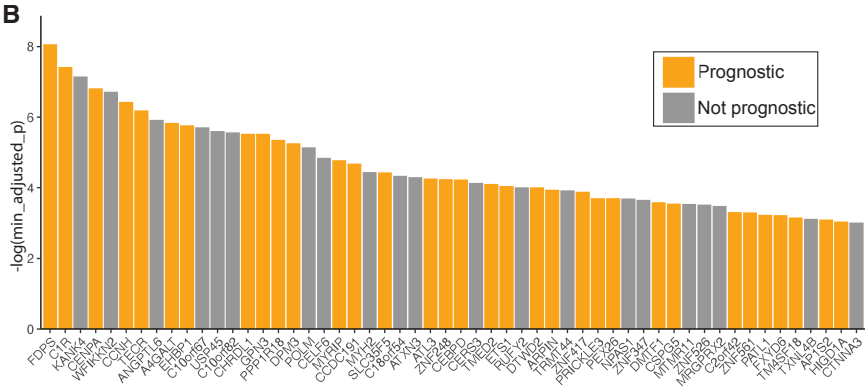



**Figure 4.5**: **Genes prioritized based on mutations in their regulatory regions.** (A) Top genes with the smallest adjusted p-values. The table shows each gene's name, the number of mutations in all its regulatory regions, the minimum p-value after adjusting for multiple testing, and the TF with the most significant binding changes for that gene. (B) Barplot showing the top 54 prioritized genes from our combined promoter and enhancer analysis. Y-axis shows the negative logarithm of the minimum adjusted p-value. Cancer prognostic markers, as reported in The Human Protein Atlas, are shown in yellow.

adjusted p-value across all TFs, and we identified 54 genes for which this p-value was $< 0.05$ (Figure 4.5, Table 4.1). In other words, for each of these 54 genes, our analysis revealed at least one TF for which the mutations in the regulatory regions of the gene have larger effects on TF binding specificity than expected by chance based on the mutation spectra in our cohort.

To determine if these 54 prioritized genes are potentially relevant in cancer, we first asked whether this set is enriched for cancer prognostic genes. Using pathology data from The Human Protein Atlas (Uhlen et al., 2017), we found that 33 out of the 54 prioritized genes are indeed prognostic markers in at least one cancer type (Figure 4.5B). This represents a significant enrichment ($p = 0.095$, Fisher's exact test) when compared to genes ranked in the bottom half of our ranked list. In addition, the top gene on our prioritized list (FDPS), as well as four other genes on the list (CENPA, PEX26, FXYD6, and TM4SF18) are prognostic markers in liver cancer.

We performed similar analyses focusing only on promoters or only on enhancer re-

**Table 4.1**: **Genes with significant TF binding changes due to mutations in regulatory DNA.** The column *min_adjusted_p* shows the minimum p-value across all TFs for each gene. Column *expr_p* shows the p-values for the expression differences between mutated (i.e. case) versus non-mutated (i.e. control) samples. NA signifies that the expression data is not available. The *n_case* and *n_ctrl* show the number of samples for each group respectively.

| gene | count | min_adjusted_p | TF | expr_p | n_case | n_ctrl |
|---|---|---|---|---|---|---|
| FDPS | 2 | 3.17E-04 | HBP1 | NA | 0 | 235 |
| C1R | 4 | 6.00E-04 | CXXC1 | 1.15E-02 | 3 | 232 |
| KANK4 | 9 | 7.83E-04 | FOXM1 | 1.10E-01 | 2 | 203 |
| CENPA | 5 | 1.10E-03 | DMRTA2 | 1.99E-02 | 3 | 232 |
| WFIKKN2 | 2 | 1.21E-03 | LBX1 | NA | 0 | 143 |
| CCNH | 2 | 1.61E-03 | ELK1 | 1.51E-01 | 1 | 234 |
| TECR | 2 | 2.05E-03 | HOXA13 | 6.01E-01 | 1 | 234 |
| ANGPTL6 | 4 | 2.69E-03 | EGR1 | 2.32E-01 | 2 | 233 |
| A4GALT | 2 | 2.92E-03 | RORA | NA | 0 | 235 |
| EHBP1 | 3 | 3.14E-03 | TCF12 | 1.86E-01 | 2 | 233 |
| C10orf67 | 2 | 3.33E-03 | RORA | NA | 0 | 108 |
| USP45 | 2 | 3.70E-03 | ESR1 | 2.72E-01 | 1 | 234 |
| C10orf82 | 6 | 3.82E-03 | NR5A1 | NA | 0 | 65 |
| CHRDL1 | 2 | 3.96E-03 | TERF2 | 9.84E-01 | 1 | 167 |
| GPN3 | 3 | 3.98E-03 | BARHL1 | 9.63E-01 | 2 | 233 |
| PPP1R18 | 6 | 4.73E-03 | PAX6 | NA | 0 | 0 |
| DPM3 | 3 | 5.23E-03 | IRF9 | 8.59E-02 | 1 | 234 |
| POLM | 3 | 5.83E-03 | KMT2A | 4.44E-01 | 3 | 232 |
| CELF6 | 6 | 7.84E-03 | NR4A1 | 7.72E-01 | 5 | 230 |
| MYRIP | 2 | 8.46E-03 | SP1 | 7.61E-01 | 1 | 232 |
| CCDC191 | 2 | 9.26E-03 | KLF1 | NA | 0 | 0 |
| MYH2 | 5 | 1.18E-02 | HSF2 | 1.03E-01 | 1 | 67 |
| SLC35F5 | 2 | 1.19E-02 | ZBTB3 | 2.53E-01 | 1 | 234 |
| C18orf54 | 7 | 1.32E-02 | TBX4 | 3.74E-01 | 7 | 228 |
| ATXN3 | 9 | 1.36E-02 | RUNX1 | 4.41E-02 | 8 | 227 |

| | | | | | | |
|---|---|---|---|---|---|---|
| ATL3 | 2 | 1.41E-02 | CEBPA | 1.82E-01 | 1 | 234 |
| ZNF248 | 4 | 1.44E-02 | ATF1 | 6.54E-01 | 3 | 232 |
| CEBPD | 3 | 1.46E-02 | TFAP4 | 5.61E-02 | 3 | 232 |
| CERS3 | 2 | 1.60E-02 | KLF1 | NA | 0 | 0 |
| TMED2 | 2 | 1.65E-02 | SOHLH2 | 1.59E-01 | 1 | 234 |
| ETS1 | 22 | 1.75E-02 | E2F2 | 1.66E-01 | 20 | 215 |
| RUFY2 | 2 | 1.81E-02 | LMX1A | NA | 0 | 235 |
| DTWD2 | 2 | 1.82E-02 | NKX2-2 | 1.42E-01 | 1 | 234 |
| ARPIN | 2 | 1.93E-02 | ARID5A | NA | 0 | 0 |
| TRMT44 | 2 | 1.97E-02 | POU4F1 | NA | 0 | 0 |
| ZNF417 | 2 | 2.06E-02 | OSR1 | 9.71E-01 | 1 | 234 |
| PRICKLE3 | 3 | 2.46E-02 | SNAI1 | 1.16E-01 | 1 | 234 |
| PEX26 | 2 | 2.48E-02 | RFX1 | 1.13E-01 | 1 | 234 |
| NPAS1 | 2 | 2.48E-02 | FOXJ2 | 2.22E-01 | 2 | 228 |
| ZNF347 | 4 | 2.60E-02 | SIX1 | 1.00E-01 | 1 | 234 |
| DMTF1 | 2 | 2.78E-02 | SOX1 | 2.13E-01 | 1 | 234 |
| CSPG5 | 3 | 2.88E-02 | NR2E3 | 1.03E-01 | 1 | 233 |
| MTMR11 | 5 | 2.92E-02 | SP1 | 6.97E-01 | 3 | 232 |
| ZNF526 | 4 | 2.96E-02 | EGR2 | 6.57E-01 | 2 | 233 |
| MRGPRX2 | 7 | 3.09E-02 | SOX10 | 8.33E-01 | 1 | 47 |
| C2orf42 | 3 | 3.65E-02 | ARID5A | 3.44E-01 | 2 | 233 |
| ZNF561 | 2 | 3.68E-02 | EGR1 | 9.14E-02 | 1 | 234 |
| PATL1 | 4 | 3.93E-02 | PBX1 | 5.30E-01 | 3 | 232 |
| FXYD6 | 6 | 3.99E-02 | HSF2 | 8.76E-01 | 5 | 230 |
| TM4SF18 | 4 | 4.28E-02 | RFX1 | 4.89E-02 | 2 | 233 |
| TXNL4B | 2 | 4.43E-02 | MITF | 3.42E-01 | 1 | 234 |
| AP1S2 | 3 | 4.50E-02 | GCM1 | 2.40E-01 | 3 | 232 |
| HIGD1A | 4 | 4.77E-02 | MYPOP | 5.41E-01 | 2 | 233 |
| CTNNA3 | 3 | 4.93E-02 | FOXA1 | 8.62E-02 | 1 | 209 |

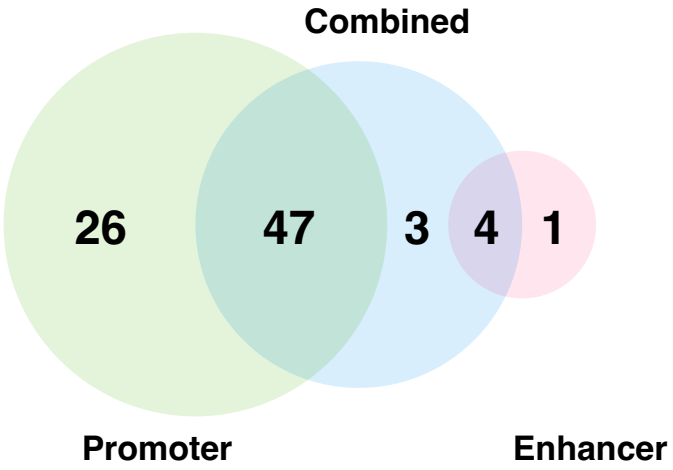



**Figure 4.6**: **Number of prioritized genes** in the promoter-only (green), combined (blue), and enhancer-only (pink) analyses.

gions. For the enhancer-only analysis, we identified 5 significant genes (at a minimum adjusted p-value cutoff of 0.05) among the 5,336 genes with enhancer mutations. For the promoter-only analysis, we identified 73 significant genes among the 11,721 genes with promoter mutations, none of which were also prioritized in the enhancer-only analysis. These results are not surprising, given that some genes only have either enhancer or promoter mutations, but not both. In addition, here we use a stringent set of enhancers, as reported by the FANTOM consortium, in order to limit the number of false positive enhancer calls; however, there are likely a large number of false negatives, i.e. enhancers that are missing from our data. Among the 54 genes identified in the combined analysis of promoters and enhancers, 4 of them are also prioritized in the enhancer-only analysis, and 47 of them are prioritized in the promoter-only analysis (Figure 4.6). Three genes, ETS1, CELF6, and PALT1 were identified only in the combined analysis (Figure 4.5B).

We also found a significant enrichment of cancer prognostic genes in the set of 73 genes prioritized in the promoter-only analysis (44 of the 73 prioritized genes are prognostic markers, Fisher's exact test $p = 0.084$). For the enhancer-only analysis,

we found that 3 of the 5 prioritized genes are prognostic markers. However, given the small number of genes prioritized in this analysis, the enrichment in prognostic markers was not significant.

### 4.3.2 Genes with significant mutations in their regulatory regions show large expression differences in mutated versus non-mutated samples

For the 54 genes prioritized based on mutations in enhancers and promoters, we also asked whether the mutations are likely to affect gene expression. To test this, we leveraged the gene expression data (EXP-S) available in ICGC for our cohort of interest, LIRI-JP (Methods). For each gene, we compared its expression level (i.e. normalized read counts, or normalized TPM values) for donors with versus without mutations in the regulatory regions of that gene, and we used a Wilcoxon rank-sum test to assess the significance of the observed gene expression differences (Table 4.1). Our analysis revealed that the difference in gene expression between donors with vs. without mutations in regulatory regions are much more significant for the set of 54 prioritized gene (i.e. those with minimum adjusted Liptak's test $p < 0.05$) compared to a control set of genes (i.e. those with $p \geq 0.1$) (Figure 4.7A). Gene with intermediate p-values ($0.05 \leq p < 0.1$) also showed significant gene expression differences, although less so than the top 54 prioritized genes (Figure 4.7A).

We note that gene expression analyses could not be performed for all genes, as for some genes with mutations in regulatory regions there was no expression data available from the donors where the mutations were observed. Among the 54 prioritized genes, 43 genes had expression data for one or more donors with mutations in enhancers or promoters, and in most of those cases the number of such donors

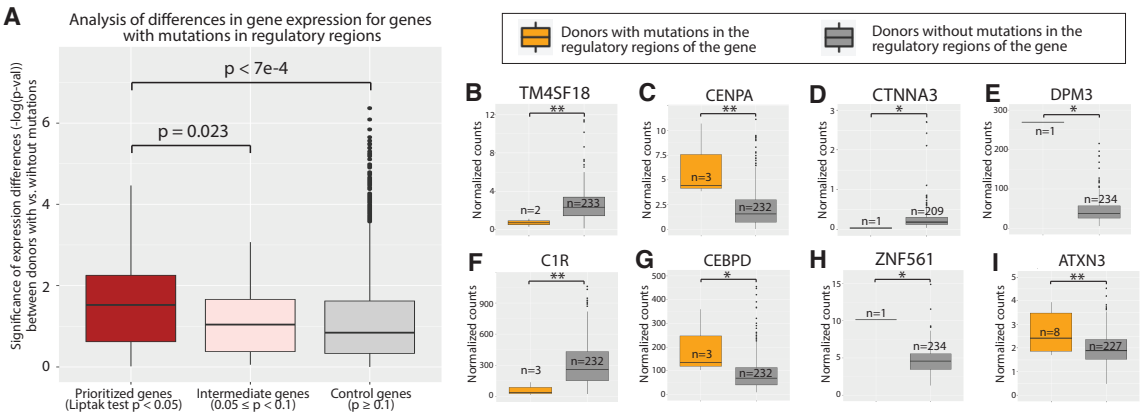



**Figure 4.7**: **Analysis of differences in gene expression for genes with mutations in enhancer and promoter regions, comparing donors with vs. without mutations in regulatory DNA.** (A) Genes prioritized by our analysis (minimum adjusted Liptak's test $p < 0.05$) have larger expression differences due to mutations in regulatory regions compared to intermediate ($0.05 \leq p \leq 0.1$) and control ($p \geq 0.1$) genes. To represent the expression differences (y-axis), we show the negative logarithm of the Wilcoxon rank-sum test p-values between the gene's expression level in the two donor groups, i.e. with vs. without mutations in regulatory regions. (B-H) Boxplots showing the 8 prioritized genes with significant gene expression differences. The y-axes show the gene expression levels, as normalized read counts, as reported in ICGC. P-values are calculated using a two-sided Wilcoxon rank-sum test, with ** denoting $p < 0.05$ and * denoting $0.05 \leq p < 0.1$.

was one, making it difficult to reach statistical significance. Nevertheless, we found significant gene expression differences for 8 of the 43 prioritized genes with available data: TM4SF18, CENPA, CTNNA3, DPM3, C1R, CEBPD, ZNF561 and ATXN3 (Figure 4.7B-H).

Among these genes, TM4SF18 and CENPA (Figure 4.7B,C) are prognostic markers in liver cancers according to The Human Protein Atlas (Uhlen et al., 2017). In addition, CENPA has been shown to be aberrantly expressed in hepatocellular carcinomas (HCCs) compared to non-tumor tissues (Li et al., 2011). Another one of the significant genes, CTNNA3, is a tumor suppressor in HCCs (He et al., 2016); according to our analysis, its expression is very low in the only donor with mutations in the regulatory regions of CTNNA3 (Figure 4.7D), which is consistent with the gene's

role as a tumor suppressor. Gene DPM3 is part of the DPM family, whose members were found to be significantly correlated with shorter overall survival in liver cancer patients (Li et al., 2020); in our analysis, the only one donor with mutations in the regulatory regions of DPM3 had a very high DPM3 expression level (Figure 4.7E). Genes C1R, CEBPD, and ZNF561 (Figure 4.7F,G,H) are prognostic markers in renal cancers, which have been shown to metastasize to liver (Bianchi et al., 2012).

The remaining prioritized gene with significant expression differences, ATXN3 (Figure 4.7I), was interestingly not among the genes characterized as prognostic markers in cancer.ATXN3 is a member of the deubiquitinating enzymes family (DUBs). These enzymes catalyze the removal of ubiquitin from protein substrates and regulate several aspects of protein fate. ATXN3 was found to play important roles in several tumours, for example by deubiquitinating PTEN in lung cancer (Sacco et al., 2014) and KLF4 in breast cancer (Zou et al., 2019), although a role for ATXN3 in liver cancer has not yet been reported. According to our analysis, mutations in the ATXN3 enhancer regions result in significant gain-of-binding mutations for TF RUNX1 (Figure 4.8), a protein involved in tumour initiation and development in hematopoietic cells and in several tissues (Taniuchi et al., 2012; Otálora-Otálora et al., 2019; Lie-A-Ling et al., 2020). Among the eight donors with mutations in ATXN3 enhancers, five have mutations that lead to significant changes in RUNX1 binding specificity, and for these five donors the expression level of ATXN3 is significantly higher than for all other donors ($p = 0.011$, Wilcoxon rank-sum test), despite the small sample size. Overall, these results show that our method is able to identify and prioritize relevant genes that are likely to be significantly affected by mutations in their regulatory regions.

We also analyzed the expression levels of genes prioritized in the promoter-only and enhancer-only analyses. For the 5 genes prioritized according to mutations in

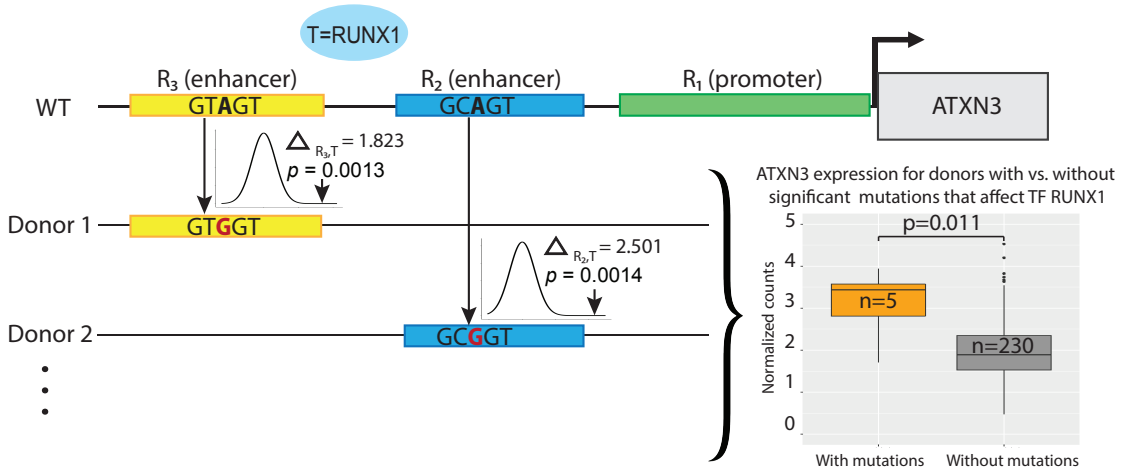



**Figure 4.8**: **Mutations in the ATXN3 enhancers that alter the binding specificity of TF RUNX1.** Out of 8 patients with ATXN3 regulatory mutations, 5 patients have significant ($p < 0.1$, one-sided Wilcoxon rank-sum test) mutations in enhancers which result in the gain of binding of TF RUNX1. Two enhancers where RUNX1 have the most significant binding changes among all TFs are shown. Donors with significant gain-of-binding RUNX1 mutations (orange boxplot) have higher ATXN3 expression compared to donors without ATXN3 regulatory mutations (grey boxplot), $p < 0.011$ two-sided Wilcoxon rank-sum test.

enhancer regions, we found more significant expression differences (between donors with vs. without regulatory mutations) than in the control gene set ($p = 0.03$, one-sided Wilcoxon rank-sum test; Figure 4.10B). Interestingly, the set of 73 genes prioritized in the promoter-only analysis (Figure 4.9) did not differ significantly from a control set of genes ($p = 0.231$; Figure 4.10A). Of these 73 genes, 47 were also prioritized in the combined analysis of promoters and enhancers, while the remaining 26 genes appeared only in the promoter-only analysis, suggesting that the promoter-only analysis may be more prone to false positive, i.e. prone to prioritizing genes for which promoter mutations do not correspond to changes in gene expression. To further investigate this, we focused on the 26 genes, and found that only one of these genes (DDX21) has significant expression differences between donors with vs. without mutations (Figure 4.10C). This suggests that integrating regulatory mutations

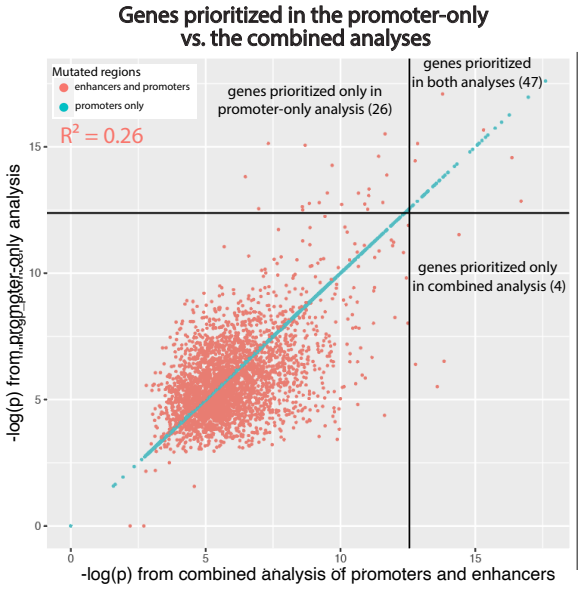



**Figure 4.9**: **Comparison between the genes' significance in the promoter-only versus the combined analysis**, shown as negative logarithm of the p-value. The $R^2$ shows the correlation for genes with mutations in both promoters and enhancers (shown in orange). Only genes with at least one promoter mutation were considered in this analysis.

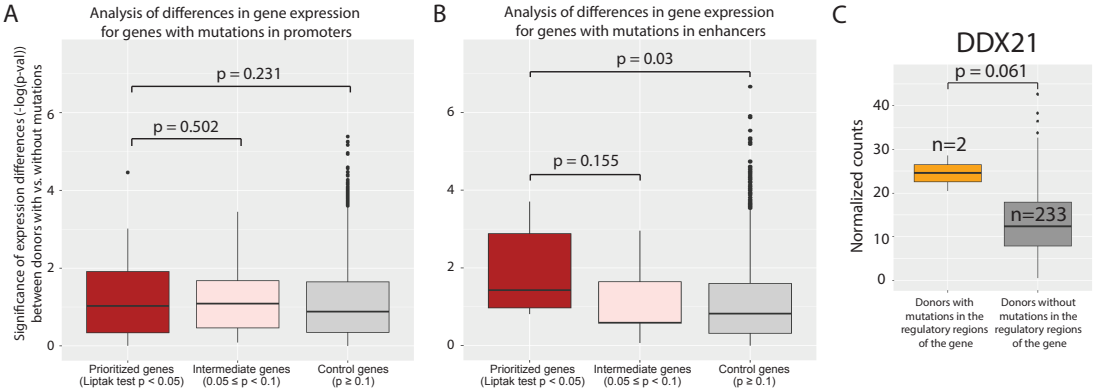



**Figure 4.10**: **Significance of gene expression differences between donors with vs. without mutations.** Shown for prioritized genes (minimum adjusted Liptak's test $p < 0.05$), intermediate genes ($0.05 \leq p < 0.1$) and control genes ($p \geq 0.1$). For (A) promoters and (B) enhancers. (C) Only one gene (DDX21) that was prioritized in the promoter-only analysis but not in the combined analysis, has a significant expression difference between donors with vs. without mutations.

over enhancers and promoters, rather than promoters alone, has the best potential to prioritize genes that are dys-regulated in cancer cells, likely due to non-coding mutations in the regulatory regions of the genes.

# Chapter 5

# Conclusions

Cooperative binding and mutations in the TF binding sites can influence how TFs recognize their DNA targets *in vivo*. Both factors require quantitative and rigorous approaches to distinguish real effects from random noise. In this study, we developed novel methods using protein binding microarray (PBM) data (Berger et al., 2006; Berger and Bulyk, 2009), which enabled us to quantify the effects of both factors on the TF-DNA recognition.

First, we developed cooperative genomic-context protein-binding microarray (cooperative gcPBM) assay to measure cooperative and independent binding of TFs to neighboring DNA sites in high throughput (Chapter 2). Using data obtained from cooperative gcPBM, we trained machine learning models to distinguish between cooperative versus independent TF binding events, based on the DNA features that are important for DNA recognition by TFs. This method enabled us to identify distinct cooperative DNA binding mechanisms between TFs ETS1-ETS1 and ETS1-RUNX1. We found that the distance and orientation between ETS1 sites are critical determinants of cooperative ETS1-ETS1 binding, while cooperative ETS1-RUNX1 interactions show more flexibility in distance and orientation, and can be predicted based on the sequence/shape features of the binding sites. The approach we described here, combining cooperative gcPBM assay with machine learning modeling, can be easily applied to study the cooperative DNA-binding patterns of any TFs.

Currently, we framed the cooperative binding modeling as a classification problem, but our methods can be expanded to predict the level of cooperative binding

to neighboring sites (e.g. as a regression problem). Since the cooperative gcPBM experiments output the binding intensity representing TF binding level to each DNA sequence, we can use the intensity values (as the dependent variable) and the DNA features of the binding sites (as the independent variables) to train regression models. Importantly, for the regression models, using different subsets of inputs for different cooperative binding systems can possibly affect their performance. In the binary systems, where there is a discrepancy in the DNA features used when TFs bind cooperatively versus independently, such as in the case of TFs ETS1-ETS1, using only the cooperative sequences might yield better performance for the models compared to using all cooperative and independent sequences. On the other hand, in a continuum systems where cooperative binding is flexible and occurs in most instances, as shown for TFs ETS1-RUNX1, using all cooperative and independent sequences might yield a better performance compared to using cooperative sequences alone. Furthermore, using complex models such as Deep Learning (as used in Avsec et al., 2021) to predict cooperative binding level can be beneficial. Compared to random forest, deep learning can learn flexible models using hierarchical layers of nonlinear features that can better capture complex DNA features such as sequence (i.e. kmers) and shape features. Nevertheless, the regression models can complement our random forest classification models by quantifying the levels of cooperative binding between TFs.

Second, to study the impact of DNA mutations on TF binding, we developed QBiC-Pred, an OLS-regression based method to quantitatively predict TF binding changes due to single nucleotide mutations; and implemented it as a webservice: qbic.genome.duke.edu (Chapter 3). We used this method to identify non-recurring putative regulatory drivers in cancers (Chapter 4). Using ICGC LIRI-JP liver cancer mutation data as input, we generated TF binding change predictions for mutations

found in promoters and enhancers. To test the significance of the predictions, we compared each prediction against a background model of random mutation effects that takes into account the mutation spectra in that particular cohort and the DNA sequence of the regulatory region. To infer whether a gene is potentially dysregulated due to non-coding mutations, we linked regulatory regions to the genes they regulate. Then, we combined the effects of mutations across all regulatory regions of each gene, using a statistical approach based on Liptak's weighted Z-score method (Lipták, 1958) and identified 54 genes whose regulation in tumor cells is likely to be significantly perturbed by non-coding mutations. We showed that these genes are enriched for cancer prognostic markers and have higher differences in expression levels between donors with vs. without mutations in regulatory regions, compared to control gene sets. This suggest that focusing on the functional effects of non-coding mutations, rather than their recurrence, has the potential to identify putative regulatory drivers and the genes they dysregulate in tumor cells.

Our study performed here on the liver cancer, can be applied to any other caner types or potentially to other types of diseases. However, for other types of diseases, different background mutation models might need to be used. For cancer, mutational signatures are commonly derived from the frequency of mutated trinucleotides and has been used as background mutation models in cancer studies (Poulos et al., 2018; Alexandrov et al., 2020). The trinucleotide mutational signatures can possibly be applied to other diseases, but further study needs to be conducted to determine the best background models for a particular type of disease. Once the background models are designed, the same method to identiy putative driver mutations in regulatory regions and prioritize genes from the mutations can be applied.

Furthermore, since our driver identification method does not rely on the recurrence of the mutations, using the simulation-based approach, we can apply our

method for patient-level analyses. Briefly, we would perform sampling *without replacement*, by randomly sampling $k$ out of the $3L$ possible mutations. We perform sampling without replacement for patient-level analyses since a patient cannot have same mutation more than once (details on the resampling process is explained in Chapter 4.2.4). However, to have enough power to build the background mutation models, we need to perform millions of resampling processes for each set of mutations in each regulatory region. This process will take a very long time to perform over the millions of input mutations and the ∼600 TFs in our data sets. Therefore, efficient parallel processing, which might be performed using cloud computing, can be designed to distribute the simulation tasks over many nodes to speed up the computation.

In this study, we consider cooperative binding and mutations in a TF binding site as separate problems. However, when mutations occur in the location for which TFs bind cooperatively, this raises a new question: how do mutations in one of the sites change the level of cooperative binding? To answer this question, we can use the cooperative binding models to predict the effect of the mutations. Briefly, when a mutation happens in one of the two adjacent sites, we first predict whether the TFs bind cooperatively to these sites using the cooperative binding classification models. If the binding is cooperative, then we can use the cooperative binding regression models as previously described in this chapter, to predict the change in the cooperative binding levels due to mutations.

Overall, our studies on quantifying the effects of cooperativity and mutations on TF binding have revealed how each factor modulates TF binding to its target sites. The approaches developed here can be used to investigate the effects of each factor to TF-DNA recognition in the cell, along with other factors modulating TF-DNA interactions (explained in Chapter 1). Such investigations will eventually lead to a

better understanding of the different roles that TFs play in the regulation of gene expression, which enable cells to carry out their diverse functions.

0.5

# Bibliography


Adams, C. C. and Workman, J. L. (1995). Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. Mol Cell Biol, 15(3):1405–1421.

Afek, A., Shi, H., Rangadurai, A., Sahay, H., Senitzki, A., Xhani, S., Fang, M., Salinas, R., Mielko, Z., Pufall, M. A., Poon, G. M. K., Haran, T. E., Schumacher, M. A., Al-Hashimi, H. M., and Gordân, R. (2020). DNA mismatches reveal conformational penalties in protein-DNA recognition. Nature, 587(7833):291–296.

Alexandrov, L. B. et al. (2020). The repertoire of mutational signatures in human cancer. Nature, 578(7793):94–101.

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol, 33(8):831–838.

Amlie-Wolf, A., Tang, M., Mlynarski, E. E., Kuksa, P. P., Valladares, O., Katanic, Z., Tsuang, D., Brown, C. D., Schellenberg, G. D., and Wang, L. S. (2018). INFERNO: inferring the molecular mechanisms of noncoding genetic variants. Nucleic Acids Res., 46(17):8740–8753.

Andersson, R. et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature, 507(7493):455–461.

Annala, M., Laurila, K., Lähdesmäki, H., and Nykter, M. (2011). A linear model for transcription factor binding affinity prediction in protein binding microarrays. PLoS One, 6(5):e20059.

Araya, C. L., Cenik, C., Reuter, J. A., Kiss, G., Pande, V. S., Snyder, M. P., and Greenleaf, W. J. (2016). Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. Nat Genet, 48(2):117–125.

Avsec, , Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat Genet, 53(3):354–366.

Babayeva, N. D., Wilder, P., Shiina, M., Mino, K., Desler, M., Ogata, K., Rizzino, A., and Tahirov, T. H. (2010). Structural basis of Ets1 cooperative binding to palindromic sequences on stromelysin-1 promoter DNA. Cell Cycle, 9(15):3054–62.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C. F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., and Bulyk, M. L. (2009). Diversity and complexity in DNA recognition by transcription factors. Science, 324(5935):1720–1723.

Bailey, T. L. and Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. Nucleic Acids Res, 40(17):e128.

Baillat, D., Begue, A., Stehelin, D., and Aumercier, M. (2002). ETS-1 transcription factor binds cooperatively to the palindromic head to head ETS-binding sites of the stromelysin-1 promoter by counteracting autoinhibition. J Biol Chem, 277(33):29386–98.

Bell, R. J., Rube, H. T., Xavier-Magalhães, A., Costa, B. M., Mancini, A., Song, J. S., and Costello, J. F. (2016). Understanding TERT Promoter Mutations: A Common Path to Immortality. Mol Cancer Res, 14(4):315–323.

Berger, M. F. and Bulyk, M. L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc, 4(3):393–411.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol., 24(11):1429–1435.

Bianchi, M. et al. (2012). Distribution of metastatic sites in renal cell carcinoma: a population-based analysis. Ann Oncol, 23(4):973–980.

Bondurand, N., Girard, M., Pingault, V., Lemort, N., Dubourg, O., and Goossens, M. (2001). Human Connexin 32, a gap junction protein altered in the X-linked form of Charcot-Marie-Tooth disease, is directly regulated by the transcription factor SOX10. Hum Mol Genet, 10(24):2783–2795.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C. Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res., 22(9):1790–1797.

Bradford, A. P., Wasylyk, C., Wasylyk, B., and Gutierrez-Hartmann, A. (1997). Interaction of Ets-1 and the POU-homeodomain protein GHF-1/Pit-1 reconstitutes pituitary-specific gene expression. Mol Cell Biol, 17(3):1065–1074.

Buisson, R., Langenbucher, A., Bowen, D., Kwan, E. E., Benes, C. H., Zou, L., and Lawrence, M. S. (2019). Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science, 364(6447).

Bulyk, M. L., Johnson, P. L., and Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res, 30(5):1255–1261.

Choy, W. W., Datta, D., Geiger, C. A., Birrane, G., and Grant, M. A. (2014). Crystallization and preliminary X-ray analysis of a complex of the FOXO1 and Ets1 DNA-binding domains and DNA. Acta Crystallogr F Struct Biol Commun, 70(Pt 1):44–48.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., Mann, R. S., and Stern, D. L. (2015). Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell, 160(1-2):191–203.

Datta, V., Siddharthan, R., and Krishna, S. (2018). Detection of cooperatively bound transcription factor pairs using ChIP-seq peak intensities and expectation maximization. PLoS ONE, 13(7):e0199771.

Davidson, A. E., Liskova, P., Evans, C. J., Dudakova, L., Nosková, L., Pontikos, N., Hartmannová, H., Hodaňová, K., Stránecký, V., Kozmík, Z., Levis, H. J., Idigo, N., Sasai, N., Maher, G. J., Bellingham, J., Veli, N., Ebenezer, N. D., Cheetham, M. E., Daniels, J. T., Thaung, C. M., Jirsova, K., Plagnol, V., Filipec, M., Kmoch, S., Tuft, S. J., and Hardcastle, A. J. (2016). Autosomal-Dominant Corneal Endothelial Dystrophies CHED1 and PPCD1 Are Allelic Disorders Caused by Non-coding Mutations in the Promoter of OVOL2. Am J Hum Genet, 98(1):75–89.

Driever, W. and Nüsslein-Volhard, C. (1988). The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. Cell, 54(1):95–104.

Elliott, K. and Larsson, E. (2021). Non-coding driver mutations in human cancer. Nat Rev Cancer, 21(8):500–509.

Ezer, D., Zabet, N. R., and Adryan, B. (2014). Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. Comput Struct Biotechnol J, 10(17):63–69.

Fisher, R. (1934). Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd, 4 edition.

Fitzsimmons, D., Hodsdon, W., Wheat, W., Maira, S. M., Wasylyk, B., and Hagman, J. (1996). Pax-5 (BSAP) recruits Ets proto-oncogene family proteins to form functional ternary complexes on a B-cell-specific promoter. Genes Dev, 10(17):2198–2211.

French, J. D. and Edwards, S. L. (2020). The Role of Noncoding Variants in Heritable Disease. Trends Genet, 36(11):880–891.

Garvie, C. W., Pufall, M. A., Graves, B. J., and Wolberger, C. (2002). Structural analysis of the autoinhibition of Ets-1 and its role in protein partnerships. J Biol Chem, 277(47):45529–45536.

Goetz, T. L., Gu, T. L., Speck, N. A., and Graves, B. J. (2000). Auto-inhibition of Ets-1 is counteracted by DNA binding cooperativity with core-binding factor alpha2. Mol Cell Biol, 20(1):81–90.

Gordan, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M. L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep, 3(4):1093–1104.

Gordân, R., Hartemink, A. J., and Bulyk, M. L. (2009). Distinguishing direct versus indirect transcription factor-DNA interactions. Genome Res, 19(11):2090–2100.

Gotea, V., Visel, A., Westlund, J. M., Nobrega, M. A., Pennacchio, L. A., and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res., 20(5):565–577.

Gu, T. L., Goetz, T. L., Graves, B. J., and Speck, N. A. (2000). Auto-inhibition and partner proteins, core-binding factor beta (CBFbeta) and Ets-1, modulate DNA binding by CBFalpha2 (AML1). Mol Cell Biol, 20(1):91–103.

Guertin, M. J. and Lis, J. T. (2013). Mechanisms by which transcription factors gain access to target sequence elements in chromatin. Curr Opin Genet Dev, 23(2):116–123.

Guo, L., Du, Y., Qu, S., and Wang, J. (2016). rVarBase: an updated database for regulatory features of human variants. Nucleic Acids Res., 44(D1):D888–893.

Gégonne, A., Bosselut, R., Bailly, R. A., and Ghysdael, J. (1993). Synergistic activation of the HTLV1 LTR Ets-responsive region by transcription factors Ets1 and Sp1. EMBO J, 12(3):1169–1178.

Haberle, V. and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol, 19(10):621–637.

He, B., Li, T., Guan, L., Liu, F. E., Chen, X. M., Zhao, J., Lin, S., Liu, Z. Z., and Zhang, H. Q. (2016). CTNNA3 is a tumor suppressor in hepatocellular carcinomas and is inhibited by miR-425. Oncotarget, 7(7):8078–8089.

He, Q., Johnston, J., and Zeitlinger, J. (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat Biotechnol, 33(4):395–401.

Heard, N. A. and Rubin-Delanchy, P. (2018). Choosing between methods of combining $p$-values. Biometrika, 105(1):239–246.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 75(4):800–802.

Hollenhorst, P. C., Chandler, K. J., Poulsen, R. L., Johnson, W. E., Speck, N. A., and Graves, B. J. (2009). DNA specificity determinants associate with distinct transcription factor functions. PLoS Genet, 5(12):e1000778.

Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., Schadendorf, D., and Kumar, R. (2013). TERT promoter mutations in familial and sporadic melanoma. Science, 339(6122):959–961.

Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., and Garraway, L. A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science, 339(6122):957–959.

Hudson, T. J. et al. (2010). International network of cancer genome projects. Nature, 464(7291):993–998.

Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., and Bulyk, M. L. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res, 43(3):117–122.

Héberlé, and Bardet, A. F. (2019). Sensitivity of transcription factors to DNA methylation. Essays Biochem, 63(6):727–741.

Ibarra, I. L., Hollmann, N. M., Klaus, B., Augsten, S., Velten, B., Hennig, J., and Zaugg, J. B. (2020). Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. Nat Commun, 11(1):124.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature, 578(7793):82–93.

Inukai, S., Kock, K. H., and Bulyk, M. L. (2017). Transcription factor-DNA binding: beyond binding site motifs. Curr Opin Genet Dev, 43:110–119.

Iwafuchi-Doi, M. (2019). The mechanistic basis for chromatin regulation by pioneer transcription factors. Wiley Interdiscip Rev Syst Biol Med, 11(1):e1427.

Iwafuchi-Doi, M., Donahue, G., Kakumanu, A., Watts, J. A., Mahony, S., Pugh, B. F., Lee, D., Kaestner, K. H., and Zaret, K. S. (2016). The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. Mol Cell, 62(1):79–91.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E., and Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res, 20(6):861–873.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013). DNA-binding specificities of human transcription factors. Cell, 152(1-2):327–339.

Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature, 527(7578):384–388.

Jonsen, M. D., Petersen, J. M., Xu, Q. P., and Graves, B. J. (1996). Characterization of the cooperative function of inhibitory sequences in Ets-1. Mol Cell Biol, 16(5):2065–2073.

Jusakul, A. et al. (2017). Whole-Genome and Epigenomic Landscapes of Etiologically Distinct Subtypes of Cholangiocarcinoma. Cancer Discov, 7(10):1116–1135.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res., 46:D260–266.

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. Nat Rev Genet, 17(2):93–108.

Kim, E. et al. (2016). Systematic Functional Interrogation of Rare Cancer Variants Identifies Oncogenic Alleles. Cancer Discov, 6(7):714–726.

Kim, S., Brostromer, E., Xing, D., Jin, J., Chong, S., Ge, H., Wang, S., Gu, C., Yang, L., Gao, Y. Q., Su, X. D., Sun, Y., and Xie, X. S. (2013). Probing allostery through DNA. Science, 339(6121):816–819.

Kim, W. Y., Sieweke, M., Ogawa, E., Wee, H. J., Englmeier, U., Graf, T., and Ito, Y. (1999). Mutual activation of Ets-1 and AML1 DNA binding by direct interaction of their autoinhibitory domains. EMBO J, 18(6):1609–1620.

Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J., and Mann, R. S. (2019). Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. Annu Rev Cell Dev Biol, 35:357–379.

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., Medvedeva, Y. A., Magana-Mora, A., Bajic, V. B., Papatsenko, D. A., Kolpakov, F. A., and Makeev, V. J. (2017). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Research, 46(D1):D252–D259.

Kähärä, J. and Lähdesmäki, H. (2013). Evaluating a linear k-mer model for protein-DNA interactions using high-throughput SELEX data. BMC Bioinformatics, 14 Suppl 10:S2.

Lamber, E. P., Vanhille, L., Textor, L. C., Kachalova, G. S., Sieweke, M. H., and Wilmanns, M. (2008). Regulation of the transcription factor Ets-1 by DNA-mediated homo-dimerization. EMBO J, 27(14):2006–2017.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and T., W. M. (2018). The Human Transcription Factors. Cell, 172(4):650–665. [PubMed:29425488] [doi:10.1016/j.cell.2018.01.029].

Lawrence, M. S. et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature, 499(7457):214–218.

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature, 505(7484):495–501.

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., and Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nat. Genet., 47(8):955–961.

Lemon, B. and Tjian, R. (2000). Orchestrated response: a symphony of transcription factors for gene control. Genes Dev, 14(20):2551–2569.

Levitsky, V., Zemlyanskaya, E., Oshchepkov, D., Podkolodnaya, O., Ignatieva, E., Grosse, I., Mironova, V., and Merkulova, T. (2019). A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. Nucleic Acids Res, 47(21):e139.

Li, M., Xia, S., and Shi, P. (2020). DPM1 expression as a potential prognostic tumor marker in hepatocellular carcinoma. PeerJ, 8:e10307.

Li, Y., Zhu, Z., Zhang, S., Yu, D., Yu, H., Liu, L., Cao, X., Wang, L., Gao, H., and Zhu, M. (2011). ShRNA-targeted centromere protein A inhibits hepatocellular carcinoma growth. PLoS One, 6(3):e17794.

Lie-A-Ling, M., Mevel, R., Patel, R., Blyth, K., Baena, E., Kouskoff, V., and Lacaud, G. (2020). RUNX1 Dosage in Development and Cancer. Mol Cells, 43(2):126–138.

Lifanov, A. P., Makeev, V. J., Nazina, A. G., and Papatsenko, D. A. (2003). Homotypic regulatory clusters in Drosophila. Genome Res., 13(4):579–588.

Link, V. M., Romanoski, C. E., Metzler, D., and Glass, C. K. (2018). MMARGE: Motif Mutation Analysis for Regulatory Genomic Elements. Nucleic Acids Res, 46(14):7006–7021.

Lipták, T. (1958). On the combination of independent tests. Magyar Tud Akad Mat Kutato Int Kozl, 3:171–197.

Liu, J., Shively, C. A., and Mitra, R. D. (2020). Quantitative analysis of transcription factor binding and expression using calling cards reporter arrays. Nucleic Acids Res, 48(9):e50.

Liu, T., Ortiz, J. A., Taing, L., Meyer, C. A., Lee, B., Zhang, Y., Shin, H., Wong, S. S., Ma, J., Lei, Y., Pape, U. J., Poidinger, M., Chen, Y., Yeung, K., Brown, M., Turpaz, Y., and Liu, X. S. (2011). Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol., 12(8):R83.

Lizio, M. et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol, 16:22.

Lochovsky, L., Zhang, J., Fu, Y., Khurana, E., and Gerstein, M. (2015). LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. Nucleic Acids Res, 43(17):8123–8134.

Lochovsky, L., Zhang, J., and Gerstein, M. (2018). MOAT: efficient detection of highly mutated regions with the Mutations Overburdening Annotations Tool. Bioinformatics, 34(6):1031–1033.

Lu, R. and Rogan, P. K. (2018). Transcription factor binding site clusters identify target genes with similar tissue-wide expression and buffer against mutations. F1000Res, 7:1933.

Mansour, M. R., Abraham, B. J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A. D., Etchin, J., Lawton, L., Sallan, S. E., Silverman, L. B., Loh, M. L., Hunger, S. P., Sanda, T., Young, R. A., and Look, A. T. (2014). Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a non-coding intergenic element. Science, 346(6215):1373–1377.

Martin, V., Zhao, J., Afek, A., Mielko, Z., and Gordân, R. (2019). QBiC-Pred: quantitative predictions of transcription factor binding changes due to sequence variants. Nucleic Acids Res, 47(W1):W127–W135.

Mas-Ponte, D. and Supek, F. (2020). DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. Nat Genet, 52(9):958–968.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). "transfac and its module transcompel: transcriptional gene regulation in eukaryotes". Nucleic Acids Research, 34(suppl1):D108–D110.

Maurano, M. T. and others. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science, 337(6099):1190–1195.

Mirny, L. A. (2010). Nucleosome-mediated cooperativity between transcription factors. Proc Natl Acad Sci U S A, 107(52):22534–22539.

Moore, J. E. et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 583(7818):699–710.

Morgunova, E. and Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. Curr. Opin. Struct. Biol., 47:1–8.

O'Leary, N. A. et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res, 44(D1):D733–745.

Otálora-Otálora, B. A., Henríquez, B., López-Kleine, L., and Rojas, A. (2019). RUNX family: Oncogenes or tumor suppressors (Review). Oncol Rep, 42(1):3–19.

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet, 10(10):669–680.

Penvose, A., Keenan, J. L., Bray, D., Ramlall, V., and Siggers, T. (2019). Comprehensive study of nuclear receptor DNA binding provides a revised framework for understanding receptor specificity. Nat Commun, 10(1):2514.

Perera, D., Poulos, R. C., Shah, A., Beck, D., Pimanda, J. E., and Wong, J. W. (2016). Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature, 532(7598):259–263.

Petersen, J. M., Skalicky, J. J., Donaldson, L. W., McIntosh, L. P., Alber, T., and Graves, B. J. (1995). Modulation of transcription factor Ets-1 DNA binding: DNA-induced unfolding of an alpha helix. Science, 269(5232):1866–1869.

Poulos, R. C., Wong, Y. T., Ryan, R., Pang, H., and Wong, J. W. H. (2018). Analysis of 7,815 cancer exomes reveals associations between mutational processes and somatic driver mutations. PLoS Genet, 14(11):e1007779.

Rhee, H. S. and Pugh, B. F. (2012). ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. Curr Protoc Mol Biol, Chapter 21:Unit 21.24.

Rheinbay, E. et al. (2017). Recurrent and functional regulatory mutations in breast cancer. Nature, 547(7661):55–60.

Rheinbay, E. et al. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature, 578(7793):102–111.

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., and Mann, R. S. (2010). Origins of specificity in protein-DNA recognition. Annu Rev Biochem, 79:233–269.

Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and Lopez-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature, 532(7598):264–267.

Sacco, J. J., Yau, T. Y., Darling, S., Patel, V., Liu, H., Urbé, S., Clague, M. J., and Coulson, J. M. (2014). The deubiquitylase Ataxin-3 restricts PTEN transcription in lung cancer cells. Oncogene, 33(33):4265–4272.

Schwessinger, R., Suciu, M. C., McGowan, S. J., Telenius, J., Taylor, S., Higgs, D. R., and Hughes, J. R. (2017). Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. Genome Res., 27(10):1730–1742.

Shen, N., Zhao, J., Schipper, J. L., Zhang, Y., Bepler, T., Leehr, D., Bradley, J., Horton, J., Lapp, H., and Gordan, R. (2018). Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential In Vivo Binding. Cell Syst, 6(4):470–483.

Shen, Z., Hoeksema, M. A., Ouyang, Z., Benner, C., and Glass, C. K. (2020). MAGGIE: leveraging genetic variation to identify DNA sequence motifs mediating transcription factor binding and function. Bioinformatics, 36(Suppl_1):i84–i92.

Shi, W., Fornes, O., Mathelier, A., and Wasserman, Wyeth, W. (2016). Evaluating the impact of single nucleotide variants on transcription factor binding. Nucleic Acids Res, 44(21):10106–10116.

Shiina, M., Hamada, K., Inoue-Bungo, T., Shimamura, M., Uchiyama, A., Baba, S., Sato, K., Yamamoto, M., and Ogata, K. (2015). A novel allosteric mechanism on protein-DNA interactions underlying the phosphorylation-dependent regulation of Ets1 target gene expressions. J Mol Biol, 427(8):1655–1669.

Shrivastava, T., Mino, K., Babayeva, N. D., Baranovskaya, O. I., Rizzino, A., and Tahirov, T. H. (2014). Structural basis of Ets1 activation by Runx1. Leukemia, 28(10):2040–2048.

Siddharthan, R. (2010). Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. PLoS One, 5(3):e9722.

Sieweke, M. H., Tekotte, H., Frampton, J., and Graf, T. (1997). MafB represses erythroid genes and differentiation through direct interaction with c-Ets-1. Leukemia, 11 Suppl 3:486–488.

Sieweke, M. H., Tekotte, H., Jarosch, U., and Graf, T. (1998). Cooperative interaction of ets-1 with USF-1 required for HIV-1 enhancer activity in T cells. EMBO J, 17(6):1728–1739.

Siggers, T. and Gordân, R. (2014). Protein-DNA binding: complexities and multi-protein codes. Nucleic Acids Res, 42(4):2099–2111.

Skalicky, J. J., Donaldson, L. W., Petersen, J. M., Graves, B. J., and McIntosh, L. P. (1996). Structural coupling of the inhibitory regions flanking the ETS domain of murine Ets-1. Protein Sci, 5(2):296–309.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends Biochem Sci, 39(9):381–399.

Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat Genet, 45(9):1021–1028.

Soufi, A., Donahue, G., and Zaret, K. S. (2012). Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell, 151(5):994–1004.

Sun, W., Graves, B. J., and Speck, N. A. (1995). Transactivation of the Moloney murine leukemia virus and T-cell receptor beta-chain enhancers by cbf and ets requires intact binding sites for both proteins. J Virol, 69(8):4941–4949.

Taniuchi, I., Osato, M., and Ito, Y. (2012). Runx1: no longer just for leukemia. EMBO J, 31(21):4098–4099.

Taylor, I. A., McIntosh, P. B., Pala, P., Treiber, M. K., Howell, S., Lane, A. N., and Smerdon, S. J. (2000). Characterization of the DNA-binding domains from the yeast cell-cycle transcription factors Mbp1 and Swi4. Biochemistry, 39(14):3943–3954.

Thomas-Chollier, M., Hufton, A., Heinig, M., O'Keeffe, S., Masri, N. E., Roider, H. G., Manke, T., and Vingron, M. (2011). Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. Nat Protoc, 6(12):1860–1869.

Tomovic, A. and Oakeley, E. J. (2007). Position dependencies in transcription factor binding sites. Bioinformatics, 23(8):933–941.

Tseng, C. C., Wong, M. C., Liao, W. T., Chen, C. J., Lee, S. C., Yen, J. H., and Chang, S. J. (2021). Genetic Variants in Transcription Factor Binding Sites in Humans: Triggered by Natural Selection and Triggers of Diseases. Int J Mol Sci, 22(8).

Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science, 249(4968):505–510.

Tugrul, M., Paixao, T., Barton, N. H., and Tkacik, G. (2015). Dynamics of Transcription Factor Binding Site Evolution. PLoS Genet., 11(11):e1005639.

Tweedie, S., Braschi, B., Gray, K., Jones, T. E. M., Seal, R. L., Yates, B., and Bruford, E. A. (2021). Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res, 49(D1):D939–D946.

Uhlen, M. et al. (2017). A pathology atlas of the human cancer transcriptome. Science, 357(6352).

van Zwet, W. R. and Oosterhoff, J. (1967). On the combination of independent test statistics. The Annals of Mathematical Statistics, 38(3):659–680.

Wagih, O., Merico, D., Delong, A., and Frey, B. J. (2018). Allele-specific transcription factor binding as a benchmark for assessing variant impact predictors. bioRxiv, preprint: not peer reviewed.

Ward, L. D. and Kellis, M. (2015). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Research, 44(D1):D877–D881.

Wasson, T. and Hartemink, A. J. (2009). An ensemble model of competitive multifactor binding of the genome. Genome Res., 19(11):2101–2112.

Wasylyk, B., Wasylyk, C., Flores, P., Begue, A., Leprince, D., and Stehelin, D. (1990). The c-ets proto-oncogenes encode transcription factors that cooperate with c-Fos and c-Jun for transcriptional activation. Nature, 346(6280):191–193.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet, 46(11):1160–1165.

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell, 158:1431–1443.

Whitlock, M. C. (2005). Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol, 18(5):1368–1373.

Wotton, D., Ghysdael, J., Wang, S., Speck, N. A., and Owen, M. J. (1994). Cooperative binding of Ets-1 and core binding factor to DNA. Mol Cell Biol, 14(1):840–850.

Xin, B. and Rohs, R. (2018). Relationship between histone modifications and transcription factor binding is protein family specific. Genome Res.

Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W. W., Gordân, R., and Rohs, R. (2014). TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res, 42(Database issue):D148–155.

Yates, B., Gray, K. A., and Wright, M. W. (2014). Genenames.org: the HGNC resources in 2015. Nucleic Acids Research, 43(D1):D1079–D1085.

Zaykin, D. V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J Evol Biol, 24(8):1836–1841.

Zhang, Y., Ho, T. D., Buchler, N. E., and Gordân, R. (2021). Competition for DNA binding between paralogous transcription factors determines their genomic occupancy and regulatory functions. Genome Res.

Zhao, J., Li, D., Seo, J., Allen, A. S., and Gordân, R. (2017). Quantifying the Impact of Non-coding Variants on Transcription Factor-DNA Binding. Res Comput Mol Biol, 10229:336–352.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods, 12(10):931–934.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., Bussemaker, H. J., Gordân, R., and Rohs, R. (2015). Role of non-coding sequence variants in cancer. Proc Natl Acad Sci USA, 112:4654–4659.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A. C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res., 41(Web Server issue):56–62.

Zou, H., Chen, H., Zhou, Z., Wan, Y., and Liu, Z. (2019). ATXN3 promotes breast cancer metastasis by deubiquitinating KLF4. Cancer Lett, 467:19–28.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B, 67:301–320.

# Biography

Vincentius Martin received his Bachelor of Engineering degree in Informatics Engineering from Bandung Institute of Technology, Indonesia in 2014. He joined GIK Lab–a computer science (CS) research group consists of collaborations between Indonesian CS students and some CS faculties from the United States–where he performed research in cloud computing and operating systems. In 2016, he joined PhD program in the CS department at Duke University where he did research in computer network security. Later, he joined Dr. Raluca Gordân's lab in 2017, to perform research in computational genomics. After graduation in 2022, he will join GlaxoSmithKline as a Genomics Data Scientist.