Transcription Factor-Centric Approaches to Identify
Regulatory Driver Mutations in Cancer

by

Jingkang Zhao

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date:_____
Approved:

_____
Raluca Gordân, Advisor

_____
Andrew Allen

_____
Jen-Tsan Chi

_____
Sandeep Dave

Dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy in the Graduate Program in
Computational Biology and Bioinformatics
of Duke University

2020

ABSTRACT

Transcription Factor-Centric Approaches to Identify
Regulatory Driver Mutations in Cancer

by

Jingkang Zhao

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date:_____
Approved:

_____
Raluca Gordân, Advisor

_____
Andrew Allen

_____
Jen-Tsan Chi

_____
Sandeep Dave

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree of
Doctor of Philosophy in the Graduate Program in
Computational Biology and Bioinformatics
of Duke University

2020

Abstract

Most previous efforts to identify cancer driver mutations have focused on protein-coding genes. In recent years, the decreasing costs of DNA sequencing have enabled whole-genome sequencing (WGS) studies of thousands of tumor samples, making it possible to systematically survey non-coding regions for potential driver events. From these studies, millions of somatic mutations in cancer have been identified, the majority of which are non-coding. However, driver identification remains a far greater challenge in non-coding regions than in coding genes, primarily due to the incomplete annotation of the non-coding genome and the unknown functional impact of non-coding mutations.

In this work, we present new approaches to identify putative regulatory driver mutations in cancer, based on new methodology for predicting the quantitative effects of single nucleotide variants on transcription factor (TF) binding. Unlike most of the previous work on driver identification, our method does not require the driver mutations to be highly recurrent; instead, we assess the mutations' significance by testing if they cause larger TF binding changes than expected in the case of completely random mutations. Since gene regulation relies on the cooperation of multiple regulatory elements, we have devised a way to combine the effects of all regulatory mutations of a gene in order to identify genes whose regulation is likely to be

significantly perturbed by the mutations observed in their regulatory elements, through changes in TF binding.

We have applied our TF-centric approaches to analyze single nucleotide variants identified in a liver cancer data set from the International Cancer Genome Consortium (ICGC), and identified potentially dysregulated genes whose regulatory mutations could trigger significant TF binding changes. Notably, the genes identified by us are different from the ones prioritized by recurrence-based approaches. However, most of the potentially dysregulated genes we have identified have large changes in gene expression and/or are cancer prognostic genes. Our results suggest that regulatory mutations should be investigated further, not just by their recurrence, but also by their functional effects such as TF binding changes, to uncover dysregulated genes that may drive tumorigenesis.

Dedication

To my family

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Most cancer mutation analyses have focused on mutations that alter the amino acid sequences of protein coding genes. However, whole-genome sequencing (WGS) of tumor genomes have revealed that the vast majority of somatic mutations in cancer are non-coding mutations, suggesting that they could also play a role in cancer initiation and development. In fact, several recent studies have demonstrated that many variants that are associated with cancer susceptibility are non-coding variants (Maurano et al., 2012; Weinhold et al., 2014). The current belief is that cancer arises because of the accumulation of multiple driver mutations that confer growth advantage to the tumor cells, some of which are non-coding (Khurana et al., 2016). As only a small proportion of the mutations in cancer are driver mutations, it is important for us to identify them and distinguish them from the rest of the mutations (passenger mutations).

This chapter provides necessary background to understand the functional role of non-coding mutations in cancer, explains the difficulties to identify non-coding driver mutations, and reviews current approaches to overcome these difficulties. We start by introducing the functional elements in the non-coding genome.

1.1 Non-coding functional elements in the human genome

Only about 1 percent of human genome is made up of protein-coding genes, and the other 99 percent is non-coding. Although no discernible function has been identified

for most of the non-coding genome, it is becoming clear that at least some of it is integral to the function of cells, particularly the regulation of protein-coding genes.

The expression of eukaryotic protein-coding genes can be regulated at several steps, including transcription initiation and elongation, and mRNA processing, transport, translation, and stability. Most regulation, however, is believed to occur at the level of transcription initiation (Matson et al., 2006). Transcription initiation typically involves two distinct families of cis-acting regulatory elements: (a) a promoter, which is found near the transcription start site (TSS), providing binding sites for the protein machinery that carries out transcription, and (b) distal regulatory elements, including enhancers, silencers, and insulators, which can be found before or after the gene they control, sometimes far away, regulating gene expression by interacting with promoters in the three-dimensional (3D) structure of the genome (Figure 1). Such regulatory elements provide recognition sites for specialized proteins - transcription factors (TFs) to bind and either activate or repress transcription.

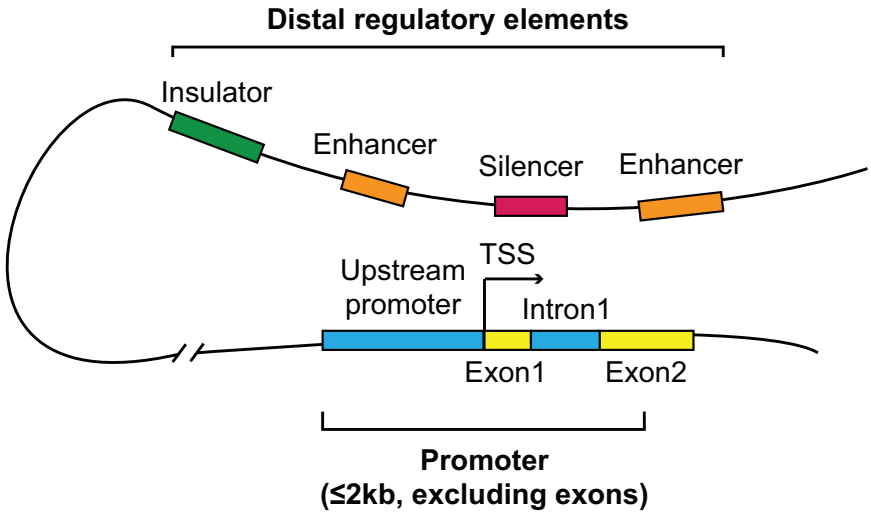

Figure 1: Cis-regulatory elements in the human genome

Besides cis-regulatory elements, non-coding RNAs (ncRNAs) also participate in the regulation of protein-coding genes. ncRNAs can be divided into several categories, such as tRNAs, rRNAs, small nucleolar RNAs (snoRNAs), small nuclear (snRNAs), miRNAs and long ncRNAs (lncRNAs; which are >200 nucleotides). All these RNAs act through different mechanisms to modulate gene expression, and many are known to have an important role in cancer biology, in particular miRNAs and lncRNAs (Khurana et al., 2016). miRNAs inhibit target gene expression by binding to the 3' UTRs of their mRNAs and causing mRNA degradation or repression of translation. lncRNAs have been shown to act as molecular scaffolds that bind proteins, DNA or other RNA molecules, and are able to tune gene expression.

It is worth pointing out that the annotation of non-coding functional elements in the human genome is far from complete. Researchers are working to identify more regulatory elements and understand the mechanism of their functions.

1.2 Role of non-coding mutations in cancer

As discussed in the previous section, non-coding regulatory elements have diverse roles in the regulation of protein-coding genes. Thus, mutations in these elements can lead to dysregulation of gene expression, potentially driving tumorigenesis. Here, we introduce some known examples of non-coding driver mutations.

One of the earliest findings of non-coding driver mutations are TERT promoter mutations (Heidenreich et al., 2014). These mutations contribute to cancer development by creating binding sites for activator transcription factors. The TERT (telomerase reverse transcriptase) gene encodes the catalytic subunit of the enzyme telomerase. Telomerase lengthens telomeres, allowing cells to escape apoptosis and become cancerous. TERT expression is generally repressed in normal somatic cells but can be overexpressed in cancer. TERT promoter mutations have been observed to create binding sites for ETS transcription factors, thus up-regulating TERT expression, and they are highly recurrent across multiple tumor types, suggesting that they likely act as drivers (Weinhold et al., 2014).

Mutations that disrupt TF-binding sites have also been observed in enhancers. 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 MYB transcription factor and forming a super-enhancer upstream of TAL1, which results in its overexpression (Mansour et al., 2014). Another study has discovered recurrent non-coding somatic mutations in chronic lymphocytic leukemia (CLL) patients, including mutations in a potential enhancer region close to the PAX5 gene, a transcription factor involved in B-cell differentiation (Puente et al., 2015).

Other important classes of non-coding mutations that are likely to contain drivers include mutations in splicing sites, mutations in UTRs, mutations in ncRNAs, and mutations in ncRNA binding sites. Large genomic rearrangements that lead to fusion events of active regulatory elements with oncogenes can also promote tumorigenesis.

So far, most of the findings of non-coding driver mutations have come from focused studies of cancer genes and their regulatory regions, as most cancer genomics studies have used exome sequencing rather than whole-genome sequencing. However, thanks to the international collaboration of cancer projects in ICGC and TCGA PCAWG, cancer genomes from over 2,000 patients across various types of cancer have been collected, enabling us to systematically survey non-coding regions for potential driver

events. As more WGS data become available, we expect to see new types of mutational effects and extend our understanding of the role of non-coding mutations in cancer.

1.3 Difficulties in the identification of non-coding driver mutations

Despite the increasing availability of WGS data, driver identification remains a far greater challenge in non-coding regions than in coding genes, owing to sequencing and mapping artifacts, poorly understood localized hypermutation processes, incomplete annotation of regulatory regions, inaccurate estimation of the background mutation rate and the unknown functional effect of non-coding mutations (Rheinbay et al., 2020). Here, we explain these difficulties in more details.

1.3.1 Sequencing and mapping artifacts

The sequencing and mapping artifacts are mostly due to the variant calling process. Unlike protein-coding genes, a large proportion of noncoding genome consists of repetitive DNA, making it difficult to align the sequences to the reference genome and call the variants using short reads, especially for calling the structural variants. There is also a difference between calling somatic and germline variants. In germline variant calling, the reference genome is the same for all samples. However, in somatic variant calling, which is used to detect cancer mutations, the reference is a related tissue from the same individual. Here, we expect to see difference in sequences both between tumor cells and between normal cells, making it harder to call variants correctly.

1.3.2 Poorly understood localized hypermutation processes

Hypermutation can be caused by environmental factors such as UV light and carcinogens. In recent years, several intrinsic sources of hypermutation have also been described. For example, dysregulation of apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) family members, has been shown to result in increased levels of C to T transitions in a wide range of cancers. Defective DNA replication repair or DNA mismatch repair is associated with hypermutation in colorectal, endometrial, and other cancers (Campbell et al., 2017). More conservative approaches need to be applied to identify non-coding driver mutations in dense hypermutated regions, as the mutations can be originated from APOBEC activity, defective repair, or other mechanisms. Overall, the driving forces and dynamics of hypermutagenesis are mostly unknown, making it difficult to distinguish driver mutations from passenger mutations.

1.3.3 Incomplete annotation of regulatory regions

The massive size of the non-coding genome implies that the chance of detecting a significant mutation after correcting for multiple hypothesis testing is almost zero and an analysis with sufficient power usually needs to focus on regulatory regions of interest instead of probing the entire genome. However, the incomplete annotation of the functional elements implies that we may not be able to consider all potential driver

mutations because some regulatory elements have not been discovered yet and associations between the elements and the genes they regulate can be missing.

1.3.4 Inaccurate estimation of the background mutation rate

A common computational approach to prioritize potential driver mutations is to identify genomic regions (hotspots) with high mutation frequency across different cancer samples, as driver mutations are expected to be under positive selection. The identification of these mutational hotspots involves comparing the mutation rate within a DNA window to a background distribution of mutation rate. To estimate the background mutation rate, the heterogeneity across different patients and across different genomic regions need to be taken into account. It is very challenging to precisely estimate the background mutation rate and we will review some of the current approaches in Section 1.4.1.

1.3.5 The unknown functional effect of non-coding mutations

Hotspot analyses usually generate a large set of candidate driver mutations, most of which are false positives. To further narrow down the candidates, functional impacts of these mutations need to be predicted. Unlike mutations in protein-coding genes, for which there are established metrics (SIFT (Ng et al., 2003), PolyPhen (Adzhubei et al., 2010), etc.) to quantify their effect on gene function, the functional impact of non-coding mutations is more difficult to predict due to the lack of knowledge about the non-coding genome. A common approach is to integrate various functional annotations of each non-

coding mutation into a functional impact score, which we will discuss in more details in Section 1.4.2.

1.4 Recent approaches to identify non-coding driver mutations

Broadly speaking, driver identification uses two lines of evidence: detection of signals of positive selection, or predictions of mutations with high functional impact (Khurana et al., 2016). As these two lines of evidence apply to both coding and non-coding driver mutations, early computational approaches to identify non-coding driver mutations learn from successful works for coding mutations. For example, they build background mutation rate models by accounting for genomic mutation rate covariates such as transcriptional activity and DNA replication timing.

However, as discussed in Section 1.3, the identification of non-coding mutations is more challenging due to our poor understanding of the non-coding genome and the multiple types of effects they can have on regulatory functions. Therefore, recent methods try to achieve better results by including new covariates about mutation process in the non-coding regions and integrating various sources of non-coding functional annotations.

1.4.1 Improvements on the background mutation rate models

LARVA (Large-scale Analysis of Variants in noncoding Annotations) (Lochovsky et al., 2015) integrates a comprehensive set of noncoding functional elements, modeling their mutation count with a beta-binomial distribution to handle overdispersion. It uses

regional genomic features such as replication timing to better estimate local mutation rates and mutational enrichments.

MutSigNC (Non-coding significance analysis) (Rheinbay et al., 2017) takes into consideration patient-specific mutation rate, patient-specific sequencing coverage, as well as information about regional mutation clustering to correct for variation in mutation rate. To account for the effect of APOBEC mutations, they exclude mutations with ≥ 0.8 probability of originating from any of the APOBEC mutation signatures.

MOAT (Mutations Overburdening Annotations Tool) (Lochovsky et al., 2017) is a computational system for identifying significant mutation burdens in genomic elements with an empirical, nonparametric method. Taking a set of variant calls and a set of annotations, MOAT calculates which annotations have observed variant counts that are substantially elevated with respect to a distribution of expected variant counts determined by permutation of the input data.

1.4.2 Integration with non-coding functional annotations

The functional impact analyses usually come after hotspot analyses to narrow down the set of candidate driver mutations. 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. Difference in TF binding between alleles can be predicted using position weight matrices (PWM) and motif prediction algorithms.

However, these methods are limited by the availability of high-quality PWMs and by the high false positive and false negative predictions rates of motif finding algorithms.

More recent methods (FunSeq2 (Fu et al., 2014), ActiveDriverWGS (Zhu et al., 2020), and DriverPower (Shuai et al., 2020)) use an overall functional impact score considering multiple features of the sequence context of the mutation (DHSs, histone marks, evolutionary conservation, etc.) to weight and prioritize candidate driver mutations. They share a lot of similarities in integrating the functional impact score into the driver identification process despite their different definitions of the score.

Table 1: Some of the recent methods to identify non-coding driver mutations

	Method	Reference
Improvements on the background mutation rate models	LARVA	Lochovsky et al., 2015
	MutSigNC	Rheinbay et al., 2017
	MOAT	Lochovsky et al., 2017
Integration with non-coding functional annotations	FunSeq2	Fu et al., 2014
	ActiveDriverWGS	Zhu et al., 2020
	DriverPower	Shuai et al., 2020

1.4.3 Comprehensive evaluation of non-coding somatic driver mutations raises doubts about drivers identified in previous studies

In a very recent work, Rheinbay et al. (Rheinbay et al., 2020) presented analyses of driver point mutations and structural variants in non-coding regions across 2,658 genomes. This is the most comprehensive evaluation of non-coding somatic mutations in terms of the number of methods employed, the number of samples analyzed, and the

number of cancer types studied. For single nucleotide variants (SNVs), they developed a strategy for combining significance levels from 13 methods (including the 6 methods in Table 1) of driver discovery to overcome the limitations of individual methods.

The team identified novel driver candidates, including SNVs in the 5′ region of TP53 gene, and in the 3′ untranslated regions of NFKBIZ and TOB1. However, the results indicated that non-coding regulatory driver mutations in known cancer genes besides TERT are much less frequent than protein-coding drivers. Moreover, some non-coding drivers identified in previous studies were found to be the result of less accurate methodologies or the result of previously uncharacterized hypermutation processes. Therefore, in order to correctly identify new regulatory driver mutations, either more cancer genomes need to be sequenced, or new methodologies that depend less on the frequency of driver mutations need to be developed.

1.5 The novelty and scope of our approach

1.5.1 Transcription factor-centric

Although the functional impact score has gained popularity in recent years, it is often hard to interpret or validate since it represents a mixed effect of multiple regulatory functions. In addition, the prediction of each regulatory function usually has its own training data, assumptions and algorithms, making it difficult to constrain the analyses to the tissue of interest.

Instead of building a complicated scoring scheme to balance each regulatory function, we want to focus on TF binding. Our aim is to develop more accurate prediction tools to characterize TF binding change due to mutations, expand the repertoire of TFs that can be precisely characterized, and speed up the prediction process so that all the regulatory mutations can be studied.

1.5.2 Identification of driver mutations outside of mutational hotspots

Recent WGS studies have identified a handful of somatic mutations in regulatory regions that affect TF binding and target gene expression. However, the number of functional non-coding mutations associated with cancer is expected to be much higher given the low overlap between those reported in different studies, and the hundreds of non-coding mutations identified in genome-wide association studies (Gan et al., 2017).

As discussed in Section 1.4.3, one possible reason is that most of the current driver identification pipelines assume that the regions that harbor driver mutations should be mutational hotspots. However, given the large number of regulatory elements in the human genome and the heterogeneity across patients, we may not be able to detect driver mutations that have moderate mutation frequency or only affect a small subset of the patients. Therefore, we aim at developing a new pipeline that does not require drivers to be highly recurrent.

1.5.3 Targeting SNVs in enhancers and promoters

In our work, we are going to focus on single nucleotide variants (SNVs) in cis-regulatory elements, in particular promoters and enhancers, because the functional impact of mutations in these regions is closely related to altered TF binding activity, while the functional role of mutations in other elements are more difficult to determine. We are more interested in SNVs than small insertions and deletions (indels) or large structural variants because they consist of the majority of somatic mutation calls from whole-genome sequencing data and their functional impact is usually within the regulatory element. Our approach can also be applied to small indels after some modification. For further information on identifying functional structural variants, the readers can refer to (Rheinbay et al., 2020).

Chapter 2. An OLS method to quantify the impact of non-coding variants on transcription factor-DNA binding

In this chapter, we present a method for assessing the impact of non-coding mutations on TF-DNA interactions, based on regression models of DNA-binding specificity trained on high-throughput *in vitro* data. We use ordinary least squares (OLS) to estimate the parameters of the binding model for each TF, and we show that our predictions of TF-binding change due to DNA mutations correlate well with measured changes in gene expression. In addition, by leveraging distributional results associated with OLS estimation, for each predicted change in TF binding we also compute a normalized score (z-score) and a significance value (p-value) reflecting our confidence that the mutation affects TF binding. We use this approach to analyze a large set of pathogenic non-coding variants, and we show that these variants lead to significant differences in TF binding between alleles, compared to a control set of common variants.

This work has been published in (Zhao et al., 2017).

2.1 Background

Single nucleotide variants (SNVs) play important roles in the pathogenesis of many complex diseases (Fredriksson et al., 2017). For mutations that occur within protein-coding genes, there are established metrics that attempt to quantify the effect of a variant on gene function. However, coding variants are only a small fraction of all genetic variants: recent studies estimate that 93% of disease- and trait-associated human

genetic variants fall within non-coding genomic regions (Maurano et al., 2012), and their functional impact is difficult to assess and quantify.

Non-coding variants can play a functional role in the cell by disrupting interactions between transcription factors (TFs) and their genomic target sites (Khurana et al., 2016). TFs are regulatory proteins that bind short DNA sites, typically in the neighborhood of the regulated genes, and promote or repress gene expression. Predicting the effect of SNVs on TF binding is an important area of research still lacking good solutions. Binding models for many human TFs are currently available (Robasky et al., 2011; Jolma et al., 2013; Mathelier et al, 2014; Weirauch et al, 2014) in the form of position weight matrices (PWMs). A PWM is a matrix of scores (or weights) for each nucleotide at each position in a TF binding site. Although they are easy to use and visualize, PWMs make the assumption that individual base pairs in a TF binding site (TFBS) contribute independently to the binding affinity. This assumption does not always hold (Bulyk et al., 2002; Udalova et al., 2002; Zhao et al., 2012; Tomovic et al., 2007). Nevertheless, although it is now recognized that PWMs cannot accurately capture TF-DNA binding affinity (Maerkl et al., 2007; Stormo, 2013; Siggers and Gordân, 2014), current methods for determining whether a SNV is likely to affect TF-DNA binding are based on differences in PWM scores (Andersen, et al., 2008; Thomas-Chollier et al., 2011; Ward and Kellis, 2012; McVicker, 2013). Such methods are generally able to detect large changes in TF binding affinity (from high affinity to non-specific binding), but they

ignore less drastic changes, which can have important phenotypic effects (e.g. (Rowan et al., 2010)).

Another drawback of using PWM models to predict the effect of SNVs is the fact that many mammalian TFs have several PWMs available in the literature, oftentimes from databases such as Transfac (Matys et al., 2006), Jaspar (Mathelier et al., 2016), UniPROBE (Newburger et al., 2009), or Cis-BP (Weirauch et al., 2014). Different PWMs can result in different predictions on whether or not a SNV will affect binding of the TF of interest, and there is no objective method to choose the best PWM to use, as quality metrics are not reported for these models. Ideally, a method for characterizing the effect of non-coding SNVs 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.

Here, we present a new method for assessing the impact of non-coding variants on TF-DNA binding. Based on high-throughput data from protein-binding microarray (PBM) experiments (Berger et al., 2006; Berger and Bulyk, 2009), we build k-mer-based models of TF binding specificity, estimating the model parameters with ordinary least squares (OLS). We use the estimated regression coefficients, as well as the variance-covariance matrix, to compute for any given mutation: 1) a quantitative prediction of the change in TF binding due to the mutation, and 2) a z-score and a p-value indicating the significance of the predicted change, given the model properties. Our approach is novel

compared to previous regression models trained on PBM data because, by using OLS, we obtain not only estimates of the regression coefficient for each k-mer, but also the variance of the coefficient estimates. Thus, our predictions of the effects of mutations on TF-DNA binding implicitly take into account the quality of the training data and model, such that in the case of poor predictive models we require a larger change in binding for a mutation to be called significant. In addition, the computed variance in the estimates of the model parameters allows us to make objective choices between different models corresponding to the same TF.

2.2 Data

2.2.1 Universal protein-binding microarray (PBM) data

Accurate methods for predicting the effect of SNVs on TF binding require accurate models of TF-DNA binding specificity. Here, to train such models we use high-throughput *in vitro* data from universal PBM assays. Each universal PBM data set is specific to one TF, and it contains quantitative measurements of the binding specificity of that TF for ~40,000 DNA sequences. The PBM protocol is described in detail in (Berger and Bulyk, 2009). Briefly, double-stranded DNA molecules attached to a glass slide (microarray) are incubated with an epitope-tagged TF. To detect the amount of TF bound to each DNA spot, the microarray is labeled with a fluorophore-conjugated antibody specific to the epitope tag, and scanned using a microarray scanner. The

fluorescence intensity of each DNA spot provides a quantitative measurement of the TF specificity for the DNA sequence in that spot.

PBM experiments are typically performed using Agilent microarrays printed with custom 60-bp DNA sequences (Berger and Bulyk, 2009). For a universal PBM array design, the DNA sequences printed on the array are computationally designed according to a deBruijn sequence of order 10 over the {A,C,G,T} alphabet, which, by definition, is guaranteed to contain all possible 10-bp DNA sequences, with each 10-mer occurring once and only once. To computationally generate the DNA library, the deBruijn sequence is split into sequences of 35 or 36 bases, depending on the design, and the remaining 25 or 24 bases, respectively, are set to the complement of a primer used to double-strand the DNA molecules. Table 2 shows as example one of the 973 PBM data sets used in our analysis of pathogenic non-coding variants.

Table 2: Example of universal PBM data set for transcription factor Arid5a

DNA sequences of length L = 60	TF binding intensity
TTGAATCAAT.....GTCCGTGCTG	74573.8653
CCAAGACAGT.....GTCCGTGCTG	45399.3011
CCAAGACAGT.....GTCCGTGCTG	40440.2397
……	……
ACTTCCGATA.....GTCCGTGCTG	39895.925

2.2.2 Massively parallel reporter assay (MPRA) data

To validate that the quantitative predictions of TF-binding changes made using our OLS k-mer models are biologically relevant, we leveraged high-throughput gene expression data from massively parallel reporter assays (MPRA) (Melnikov et al., 2012). Briefly, in an MPRA experiment one first synthesizes tens of thousands of oligonucleotides that contain a library of regulatory elements (enhancers), each coupled to a short DNA tag. The oligonucleotides are used to generate a pool of plasmids, where each plasmid contains one of the regulatory elements of interest upstream of an open reading frame followed by the sequence tag corresponding to that regulatory element. The pool of plasmids is co-transfected into cells, where the regulatory elements drive transcription of mRNA molecules containing the tags. The tags in the reporter mRNAs, as well as the original plasmid pool, are sequenced and counted. The ratio of these counts, or the logarithm of the ratio, is taken as a measurement of the gene expression driven by each regulatory element.

Here, we use MPRA data from two recent studies. Melnikov et al. (Melnikov et al., 2012) reported the expression levels of a reporter gene (an inert open reading frame) downstream of variants of a synthetic, 87-bp cAMP-regulated enhancer. The mutants were either generated by single nucleotide substitutions (for a total of 261 variants) or by random multiple 1-bp nucleotide substitutions, introduced at a rate of 10% per position (~27,000 variants). The expression level of each variant was reported as the median of

the mRNA-based counts normalized by the DNA-based counts, taken over multiple tags. In our analyses, we used the natural logarithm of the ratios of the expression levels of mutants to the expression level of the wild-type sequence.

2.3 Methods

2.3.1 Training k-mer regression models of TF binding specificity using ordinary least squares (OLS)

In a universal PBM experiment, TF binding to each of the ~40,000 pre-designed L-bp DNA sequence is measured as fluorescent signal (Table 2). We apply a logarithmic transformation to the fluorescent signal, which makes the experimental noise uncorrelated with the signal, and we use the natural log-transformed fluorescence intensities as the dependent variable Y. As for independent variables X, we use the counts of each k-mer within the L-bp DNA sequences, with the value of k decided based on validation experiments (Section 2.4.1). Since the DNA is double -stranded, and binding of TFs is not strand-specific, we regard each sequence and its reverse complement as the same feature. Thus, the number of features n_k for a k-mer model is $4^k/2$ when k is an odd number, and $(4^k - 2^k)/2 + 2^k$ when k is even.

Suppose there are a total of N L-bp sequences. We convert each sequence into the counts of all n_k k-mers in an overlapping fashion, generating a $N \times n_k$ covariate matrix X. There is an inherent restriction for the rows of the matrix. For any row i, the sum of the counts is $L - k + 1$, which is due to the fact that every L-bp sequence contains

$L - k + 1$ overlapping k-mers. The linear dependency of the $n_k$features renders the intercept term redundant, and we therefore train our models without it:

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_{n_k} x_{in_k} + \varepsilon_i$$

We compute the ordinary least square (OLS) estimates for the coefficients $\hat{\beta}$'s, as well as the covariance matrix $\hat{\Sigma}$:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\Sigma} = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{N - n_k}(X'X)^{-1}$$

2.3.2 Statistical testing using OLS k-mer models of TF binding specificity

By assuming independence and normality on the error terms $\varepsilon \sim N(0, \sigma^2 I)$, we can perform statistical tests on linear combination of β's. Given a vector c of the same length as β, we can test:

$$H_0: c'\beta = 0$$

$$H_1: c'\beta \neq 0$$

A t-statistic can be derived from the data

$$t = \frac{c'\hat{\beta}}{\sqrt{c'\hat{\Sigma}c}} \sim t_{N-n_k}$$

In practice, since we have a large number of observations, the distribution of the test statistic is approximately normal, and we can thus compute a z-score for $c'\hat{\beta}$.

2.3.3 Using OLS k-mer models to predict the effect of SNVs on TF-DNA binding

The method in Section 2.3.2 can be directly applied to predict the effect of single base-pair variants on TF binding. To illustrate this, we provide an example for a mutation (A to C) that affects a binding site for TF Creb1 (Table 3). The wild-type and mutated binding sites are shown in bold. The mutated position is underlined. The 6-mers in parentheses are the reverse complements of 6-mers in the original sequence. In these cases, the reverse complement 6-mers were used as features because they are alphabetically ranked lower that the corresponding 6-mers on the forward strand.

We used 6-mer features to train a regression model from universal PBM data for Creb1, available from (Weirauch et al., 2014). In a 6-mer model, there are a total of 2080 features, and we can derive the estimates of coefficients $\hat{\beta}$ for all 6-mers, as well as the covariance matrix estimate $\hat{\Sigma}$.

Table 3: Single base-pair mutation overlapping a binding site for TF Creb1

Wild-Type CCCAT **TGACGTCA**ATGGG	**Mutant** CCCAT **TGCCGTCA**ATGGG
CATTGA	CATTGC
ATTGAC	ATTGCC
TTGACG (CGTCAA)	TTGCCG (CGGCAA)
TGACGT (ACGTCA)	TGCCGT (ACGGCA)
GACGTC	GCCGTC (GACGGC)
ACGTCA	CCGTCA

Given a k-mer model, a single base-pair mutation leads to a change in every k-mer in a $2k-1$ bp region centered at the mutated base. Thus, the total change is:

$$\sum_{p=1}^{k}(\hat{\beta}_{j_p}-\hat{\beta}_{i_p})$$

Here, j_p is the index of the pth k-mer in the mutated sequence, and i_p is the index of the corresponding k-mer in the original sequence.

For the example in Table 3, the mutation causes a change in 6 consecutive 6-mers, and the total effect of the mutation is:

$$\hat{\beta}_{CATTGC}+\hat{\beta}_{ATTGCC}+\hat{\beta}_{CGGCAA}+\hat{\beta}_{ACGGCA}+\hat{\beta}_{GACGGC}+\hat{\beta}_{CCGTCA}$$
$$-\hat{\beta}_{CATTGA}-\hat{\beta}_{ATTGAC}-\hat{\beta}_{CGTCAA}-\hat{\beta}_{ACGTCA}-\hat{\beta}_{GACGTC}-\hat{\beta}_{ACGTCA}$$

In vector notation, this effect can be written as $c'\hat{\beta}$ where the elements of c' are all zero except for the ones corresponding to elements of $\hat{\beta}$ that appeared in the expression above. Next, we compute the test statistic for the difference in predicted binding affinity $c'\hat{\beta}$ between the mutant and the wild-type sequences, and test its significance.

2.4 Results

2.4.1 OLS 6-mer models can accurately predict TF binding intensity

To check the accuracy of the k-mer models and determine the best value for k, we used 115 TFs from the Cis-BP database (Weirauch et al., 2014) for which universal PBM data is available from two distinct array designs. We learned OLS k-mer models from one array design and tested them on the independent data obtained using the second array design. The Pearson correlation coefficient (R) between our predicted TF

binding intensities and the measured intensities from both the training and the test data sets are summarized in Figure 2. We note that PBM experiments performed on different arrays are not replicate experiments, as the array designs contain different DNA sequences. In addition, data quality is highly variable across the PBM data sets, so we expect the performance of any model trained on these data sets to also vary. Importantly though, our 6-mer OLS models are designed to implicitly take data quality into account.

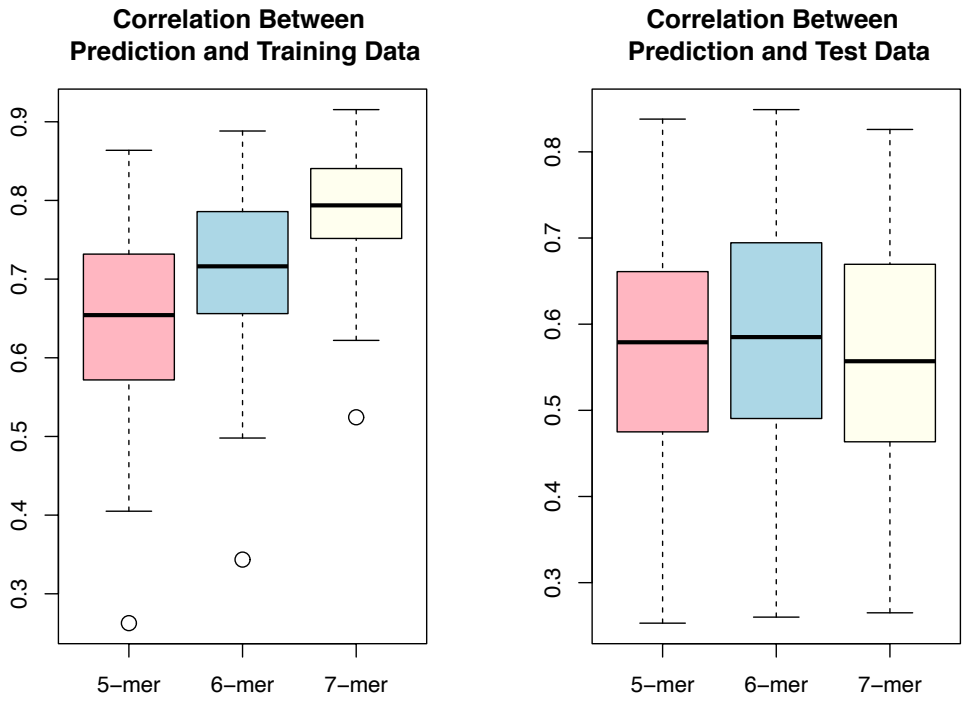

Figure 2: Performance of OLS k-mer models for k = 5, 6, 7. Boxplots show Pearson correlation coefficients between predicted and measured TF binding intensities, for 115 TFs with data available from two universal PBM designs.

Figure 2 shows the performance of 5-mer, 6-mer, and 7-mer OLS models, which have 512, 2080, and 8192 features, respectively. For $k = 8$, the models have a total of

32,896 features. Since we only have ~40,000 observations, models with $k > 8$ run into dimensionality problems and we cannot get OLS estimates for the parameters. Among 5-mer, 6-mer, and 7-mer models, we found that 7-mers models perform best on the training data (Figure 2, left panel). However, on independent test data from a different array design, 7-mer models perform worse than 6-mers models (Figure 2, right panel), indicating that they are likely over-fitting the data. Thus, our results indicate that k-mer models with $k = 6$ have the best accuracy in predicting TF binding intensity for new DNA sequences. All results presented below use 6-mer OLS models.

The main goal of our method is to predict changes in TF binding, not absolute TF binding levels. Nevertheless, to ensure that our 6-mer OLS models are accurate in predicting TF binding levels, we compared them to previous models trained and tested on PBM data from different array designs. For this comparison, we used the PBM data from the DREAM5 TF-DNA Motif Recognition Challenge (Weirauch et al., 2013), which includes independent data sets obtained using two different array designs, for 66 mouse TFs. In the challenge, PBM intensity data were provided only for one array design, and the performance of each algorithm was evaluated by assessing the prediction accuracy on the other array design, using the Pearson correlation coefficient (R). Weirauch et al. (Weirauch et al., 2013) used several normalization techniques to transform the PBM data before using it for training and testing, and for each algorithm they selected the combination of normalization steps that resulted in the best prediction accuracy on the

test data. In contrast to their approach, we use the PBM data directly in our algorithm, applying only a logarithmic transformation to all PBM intensities, and thus keeping the test PBM data completely independent from the training step. The performance of our 6-mer OLS method was above average compared to the 15 methods tested in (Weirauch et al., 2013). Thus, we conclude that the accuracy of our method in predicting TF binding intensity is comparable to existing algorithms, with our method having the unique advantage that it implicitly incorporates data quality into the TF binding models.

2.4.2 TF binding change predictions based on OLS 6-mer models correlate well with gene expression changes

To validate that our OLS 6-mer models are able to quantitatively predict the effect of nucleotide mutations, we leveraged high-throughput reporter expression data generated using massively parallel reporter assays (MPRA). This part is a collaborative work with Dongshunyi Li.

First, we focused on MRPA data for an 87-bp synthetic enhancer that contains four binding sites for transcription factor Creb1. Melnikov et al. (Melnikov et al., 2012) reported expression measurements for the wild-type enhancer (Figure 3a), for all possible single base pair mutations, as well as a large number of enhancer variants with multiple mutations randomly distributed across the enhancer region. The expression values are reported as ratios of tag counts in the reporter mRNA versus tag counts in the plasmid pool (see Section 2.2.2). Based on the expression values reported in (Melnikov et al., 2012), we computed for each mutant enhancer the natural logarithm of the ratio

between the expression of the mutant and the expression of the wild-type enhancer. We asked whether these changes in gene expression can be explained, at least in part, by changes in Creb1-DNA binding predicted according to 1) our OLS 6-mer model for TF Creb1; and 2) the mouse Creb1 PWM reported in the Cis-BP database (motif identifier M0297_1.02). To score DNA sites according to the PWM we used the log-likelihood (LLR) score, i.e. we computed the base 2 logarithm of the ratio between the probability of the site according to the PWM model, and the probability of the site according to a uniform background model over the four nucleotides.

Before applying our approach to predict the effect of mutations on Creb1-DNA binding, we verified the accuracy of Creb1 OLS 6-mer models trained on PBM data, and we selected the most accurate model. There are two universal PBM datasets available for mammalian Creb1, from two distinct universal designs, denoted HK and ME (Weirauch et al., 2014) (Section 2.2.1). We trained OLS 6-mer models on each array design, and we compared the models according to their predicted variance for the parameter estimates, i.e. the diagonal of the estimated covariance matrix $\hat{\Sigma}$. The parameter estimates for the model trained on the ME data set showed lower variances (Mann-Whitney U test $p < 10^{-6}$), and thus it was selected as the final Creb1 OLS 6-mer model.

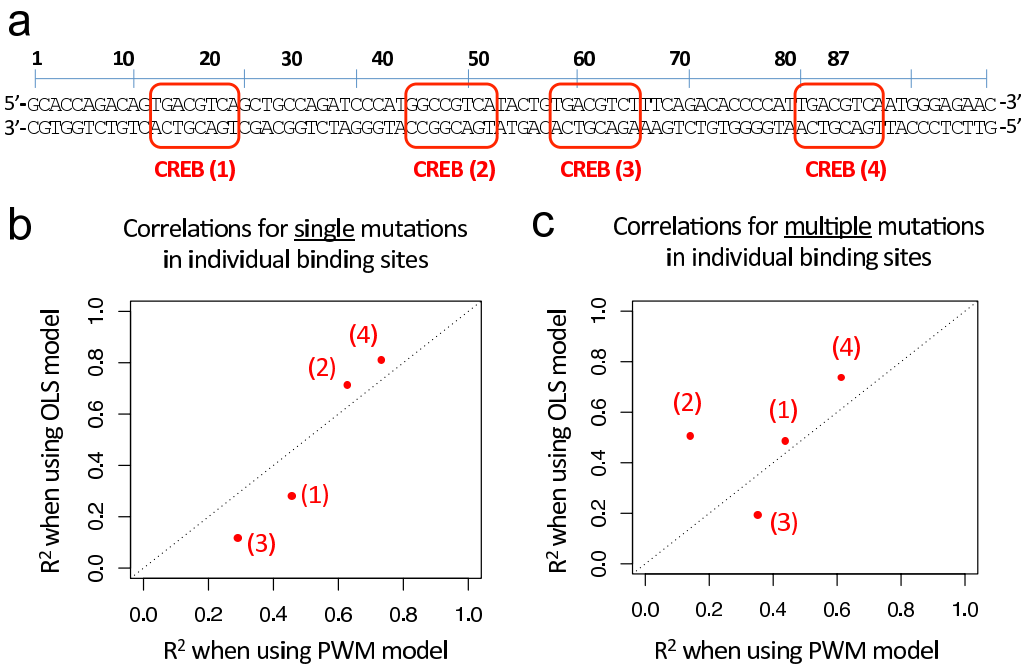

Figure 3: Correlations between measured gene expression changes and TF binding changes predicted by OLS and PWM models, for individual TF binding sites

We first compared the OLS and PWM models on variant enhancers with single bp mutations in each of the four Creb1 binding sites (defined as shown in Figure 3a, red rectangles mark the four annotated Creb1 binding sites). For each binding site, we asked how well the measured gene expression changes due to 1-bp mutations within the binding site correlate with the predicted changes in TF binding. The OLS model performed better than the PWM for mutations in sites 2 and 4 (where both models have good prediction) and worse than the PWM for mutations in sites 1 and 3 (where both models performed poorly) (Figure 3b).

Next, we compared the OLS and PWM models on enhancers with multiple 1-bp mutations in each of the four Creb1 binding sites. The OLS model outperformed the PWM on three of the four binding sites (Figure 3c). This result was as expected. Unlike our OLS k-mer models, PWM models cannot capture dependencies between positions within TF binding sites, and this shortcoming can lead to poor predictions when multiple mutations are introduced in a site.

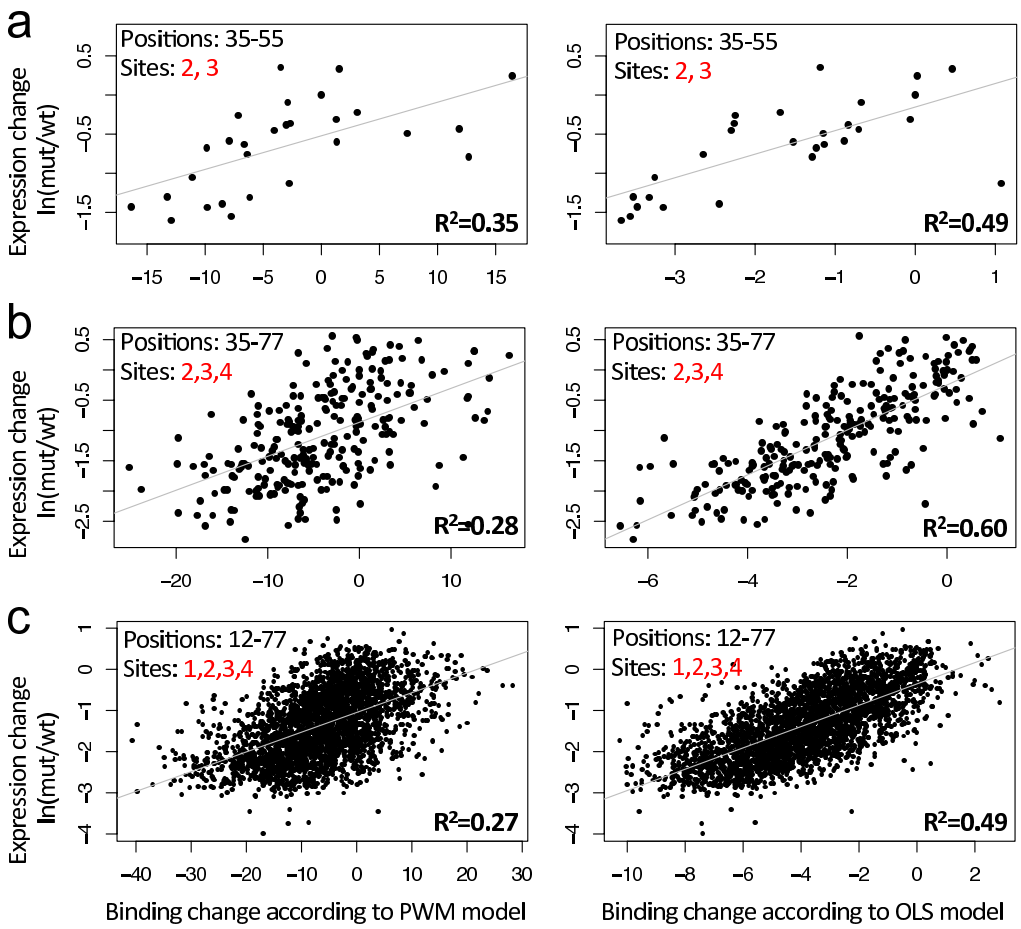

Figure 4: Correlations between measured gene expression changes and TF binding changes predicted by OLS and PWM models, for multiple TF binding sites and multiple mutations.

Finally, we compared the OLS and PWM models on enhancers with multiple 1-bp mutations in regions that cover several of the Creb1 binding sites (Figure 4). For such regions, using the OLS 6-mer model is straightforward, since the model can be applied to predict TF binding for sequences of any width. In contrast, PWM models have a fixed width. To apply the PWM model to longer regions, we used a sliding window of size 8 (the same size as the Creb1 PWM), we scored each window according to the PWM, and we summed up the scores above a certain cutoff, expressed in terms of the maximum LLR score that can be obtained using the PWM model (e.g. 20% the maximum score, 30%, 40%, 50%, 60%, etc.). We also tested other approaches to score long DNA regions using PWMs, such as the GOMER model (Granek and Clarke, 2005), but the thresholding approach described above worked best. We found that a cutoff of 60% leads to the best performance of the PWM model, so we used this cutoff in our comparisons. Figure 4 shows that as we include more binding sites in our analysis, the performance of the PWM decreases, reaching an $R^2$of 0.27 when all four binding sites are included (Figure 4c, left panel). In contrast, the OLS model continues to perform well regardless of the number of binding sites included in the analysis, and it constantly achieves correlations of 0.49 or higher (Figure 4, right panels).

We also tested additional Creb1 PWM models, including the curated human Creb1 motif from the HocoMoco database (Kulakovskiy et al., 2013) (downloaded from Cis-BP, motif identifier M6180_1.02), which achieved correlations < 0.1 in all analyses of

mutations in multiple binding sites. Overall, the Cis-BP motif M0297_1.02 resulted in the highest correlations with the gene expression data. Thus, we focused on this motif for all comparisons described in this section.

Our results show that changes in TF binding, predicted using the OLS 6-mer model, can explain ~50% of the change in gene expression due to DNA mutations. This fraction is remarkable, given the complexity of gene regulation. We do not expect TF binding changes to completely explain gene expression changes, nor to correlate linearly with expression changes observed in the cell. The large correlation between changes predicted by the OLS model and measured changes in gene expression demonstrates that our predictions are quantitative and biologically relevant.

2.4.3 Analysis of pathogenic non-coding variants

To further illustrate how our methods can be used to analyze non-coding variants, we performed a broad analysis of all non-coding pathogenic SNVs annotated in the Human Gene Mutation Database (HGMD) (Stenson et al, 2014) and ClinVar (Landrum et al., 2016). This part is a collaborative work with Jungkyun Seo.

Starting with 101,833 SNVs, we excluded any variants that overlapped with consensus coding sequences, leaving 4,655 unique variants. Next, we removed variants considered to reside within coding/canonical splice of any Ensembl coding transcript. We also excluded the variants on sex chromosomes or mitochondrial chromosomes, leaving a total of 3,422 unique non-coding pathogenic autosomal variants for analysis.

We also selected a set of control variants among the common variants annotated in the 1000 Genomes Project (1000 Genomes Project Consortium, 2015), following similar filtering steps. We first downloaded all SNVs from phase 3 of the 1000 Genomes Project and excluded all rare variants (i.e. variants with minor allele frequency <0.01). To obtain non-coding variants, we annotated the filtered variants using the Variant Effect Predictor (VEP) tool (McLaren et al., 2016), and we excluded variants annotated to reside in a coding region. After this process, a total of 11.9 million non-coding SNVs were retained from 84.4 million 1000 Genomes variants. Finally, we randomly selected 3,422 non-coding autosomal variants that followed a similar genomic distribution as the pathogenic variants.

We trained 6-mer regression models for all 973 PBM data sets available for human and mouse TFs, and applied our models to predict the binding changes due to SNVs in the pathogenic and control data sets. For each SNV, we took the maximum absolute value of the 973 predicted z-scores as the measure of the binding change due to the SNV. Figure 5 displays the empirical cumulative density functions of the predicted binding changes for the 3,422 variants in each data set. Our result shows that the binding changes caused by the pathogenic variants are significantly larger than the changes caused by the control variants ($p < 10^{-6}$, Mann-Whitney U test), indicating that there is a strong regulatory component for the annotated pathogenic variants

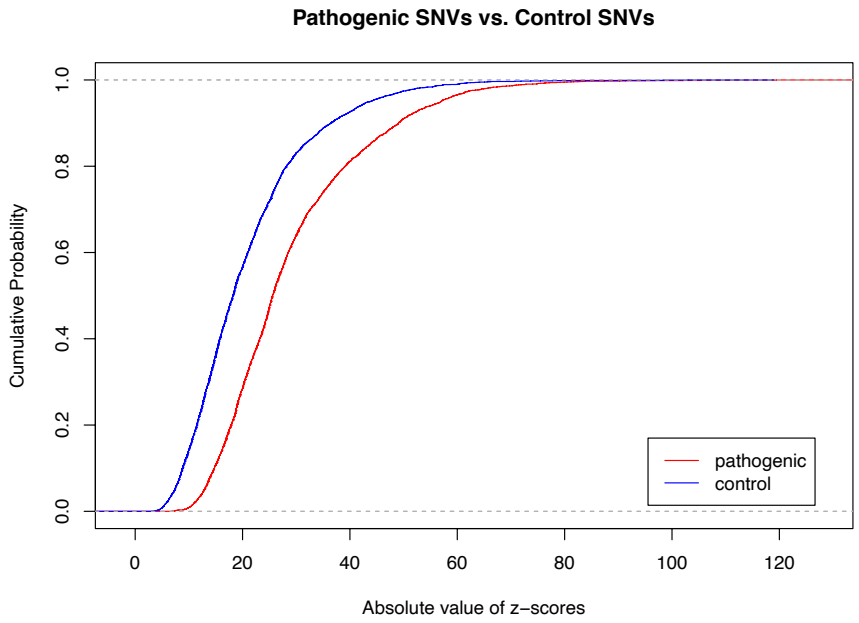

Figure 5: Comparison of predicted TF binding changes between pathogenic SNVs (red line) and control SNVs (blue line). Overall, pathogenic SNVs have a larger effect on TF binding, as predicted by our OLS 6-mer models.

2.5 Discussion

We developed a new method to assess the impact of non-coding mutations on TF-DNA binding, using high-throughput PBM data. Such data is currently available for almost 1,000 mammalian TFs covering a broad range of TF families. Each PBM data set contains binding measurements for ~40,000 short DNA sequences. We utilize the data to build k-mer linear regression models, estimating the model parameters with OLS. The novelty of our approach, compared to previous work, is that we can use the estimated regression coefficients together with the estimated covariance matrix to compute not

only the change in TF binding due to a mutation (or set of mutations), but also a z-score and a p-value indicating the significance of the change.

Importantly, for any given mutation, the z-scores and p-values obtained for different TFs are directly comparable and can be combined to assess the broad regulatory effects of mutations, as illustrated in Section 2.4.3. In contrast, given that PWM scores are not directly comparable across different models, combining differences in PWM scores for large sets of TFs is not straightforward. As another advantage of OLS k-mer models over PWMs, we note that our k-mer models can be used to assess the effect of mutations over long regions containing multiple mutations and binding sites, without the need to call binding sites according to some score cutoffs. As shown in Section 2.4.2 for mutations in an enhancer regulated by Creb1, our OLS model was able to quantitatively capture the effects of DNA mutations over long regions, explaining ~50% of the change in gene expression. Thus, we expect any method that uses PWM models to assess the functional effects of non-coding variants to benefit from using our OLS models instead of PWMs.

We note that individual k-mer features in our models are not independent, so their estimated coefficients should not be interpreted individually. Overall, though, the change in binding score and the corresponding z-scores and p-values, computed as described in Section 2.3.3, can be interpreted directly because they take into account all overlapping k-mers affected by the mutation of interest, and the z-scores and p-values

also take into consideration the correlation between features through the estimated variance-covariance matrix. One concern about the z-scores and p-values is their dependence on the normality of the random error, which can be approximately achieved by exploring the transformation of the raw intensity score. In our study we did not elaborate on finding the optimal transformation, since we have a sufficiently large sample size for the statistical tests to be applicable even in cases of non-normality. The main limitation of our OLS approach is that the number of features cannot exceed the number of observations. In future work we will focus on Bayesian methods that can be applied to higher dimensional data, while at the same time providing posterior distributions that allow us to make statistical inferences about the predictions.

Chapter 3. QBiC-Pred: a web service for fast and quantitative predictions of transcription factor binding changes based on the OLS method

This chapter introduces QBiC-Pred, a web server for predicting quantitative TF binding changes due to nucleotide variants. QBiC-Pred uses regression models of TF binding specificity trained on high-throughput *in vitro* data. The training is done using ordinary least squares (OLS), as described in Chapter 2, and we leverage distributional results associated with OLS estimation to compute, for each predicted change in TF binding, a p-value reflecting our confidence in the predicted effect. Here we present extensive validations and show that OLS models are accurate in predicting the effects of mutations on TF binding *in vitro* and *in vivo*, outperforming widely used PWM models as well as recently developed deep learning models of specificity. QBiC-Pred takes as input mutation data sets in several formats, and it allows post-processing of the results through a user-friendly web interface. QBiC-Pred is freely available at http://qbic.genome.duke.edu.

This is a collaborative work with Vincentius Martin, a Computer Science PhD student who also works in the Gordan Lab. Vincentius has accelerated the algorithm and implemented the web service and I have curated available PBM data sets and carried out comparisons with other available methods. This work is published in (Martin et al., 2019).

3.1 Introduction

In Chapter 2 we have introduced an ordinary least squares (OLS)-based method for assessing the impact of non-coding mutations on TF-DNA interactions. Briefly, we used high-throughput *in vitro* TF binding data from universal protein-binding microarray (uPBM) experiments (Berger et al., 2006) to train regression models of TF-DNA 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.

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

3.1.1 Input

QBiC takes as input mutation/variant data sets containing single nucleotide variants (SNVs), in several formats: 1) variant files in the standard variant call format (VCF); 2) "simple somatic mutations" files generated by the International Cancer Genome Consortium (ICGC) (International Cancer Genome Consortium, 2010); 3) tab- or comma-separated values files with the columns: chromosome, chromosome_ pos, mutated_from, and mutated_to; and 4) text files containing 17-bp DNA sequences with the "mutated from" 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 mutations 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) (Gray et al., 2015). The list of available TFs and models is available on the QBiC website in the Downloads section.

3.1.2 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. Similar 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 data sets used to obtain individual predictions and provides links to the HGNC database 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) 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, 2016; 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 Methods

3.2.1 Curation of universal PBM data for training

The OLS models used by QBiC were trained on curated uPBM data from literature and our laboratory, mapped to 582 human TF proteins. Each uPBM experiment measures the binding specificity of a TF for ~40,000 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). 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.

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, as shown in Chapter 2. 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 22 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 1,450 uPBM data sets to 633 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. (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 6 uPBM data sets, including the top 3 data sets with the highest cross-validation accuracy, as well as the top 3 data sets obtained for TFs with the highest amino-acid identify to the human TFs. The final mapping, which includes 666 uPBM data sets and 582 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 Results).

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).

3.2.4 QBiC-Pred web server implementation

QBiC-Pred was developed using the Flask web framework and it runs under Apache 2.4. Predictions of the effects of input variants on TF binding are made using pre-computed 12-mer tables encoding the predicted TF binding changes, z-scores and p-values for all possible mutations in all possible contexts. To further speed up the computations, QBiC uses asynchronous multiprocessing with the Celery framework, where 4 workers (i.e. processes) are spawned for each request. Each worker extracts predictions for a subset of the input TFs. The prediction results are saved in a Redis database for two days; during this time the user can access the results using a unique job identifier and can interactively process the results within QBiC (Figure 6). Users can also download the prediction results and re-upload them later, even after the job expired, for further processing within the QBiC framework.

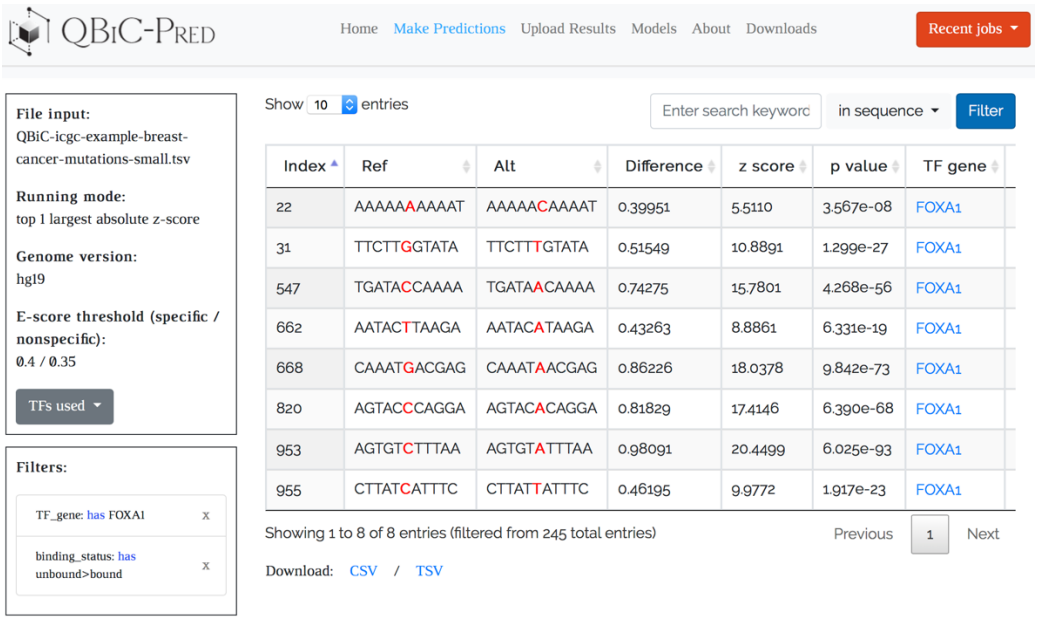

Figure 6: Web server results page for a sample mutation file containing ICGC breast cancer simple somatic mutation data

Users can leave the QBiC website while the predictions are being calculated 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 number of TFs selected as input, as QBiC 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. After all predictions are computed, they are displayed in a table format with filtering capabilities. Users can post-process the results and downloaded them as csv or tsv files.

3.3 Results

In Chapter 2 we showed that our OLS model-based predictions of TF binding changes due to DNA mutations correlate well with measured changes in gene

expression. 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 models of TF binding specificity outperform PWMs and DeepBind models in predicting *in vitro* TF binding changes

As described in Section 3.2.2, 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., 2018) 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, 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. (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 7. 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.

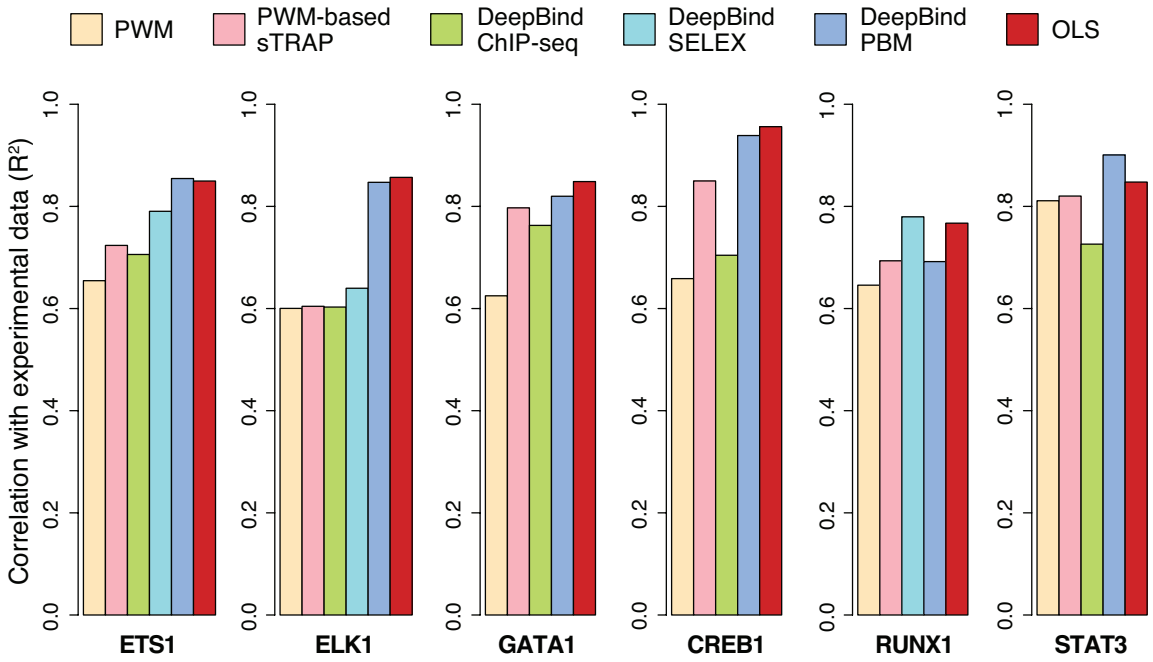

Figure 7: Performance of OLS models in predicting *in vitro* TF binding changes, compared to PWM and DeepBind models

Figure 8 shows a detailed comparison of five models (OLS, PWM, sTRAP, DeepBind SELEX, and DeepBind PBM) for a binding site of TF ELK1. The input mutation file used in QBiC to generate the ELK1 binding change predictions shown in Figure 8 can be downloaded from the QBiC website as the sample input file in sequence format. 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. There are also positions where the magnitude of the TF binding change seems to be overestimated by our OLS model but not so much by PWM-based and DeepBind models, such as the T to C mutation at the last position. We note, however, that in this case the correctness of the magnitude of the predicted increase is difficult to assess. For the PWM and the DeepBind SELEX models, the largest predicted increases are incorrect, so we cannot compare them directly to predicted increase at the last position. For the PWM-based sTRAP model and the DeepBind PBM model, the magnitude of the predicted increase at the last position is larger than for other correctly predicted increases, similarly to our OLS model. Thus, it is difficult to judge which model performed best at predicting this particular increase. Nevertheless, over all mutations tested, the OLS model has the best performance (see also Figure 7).

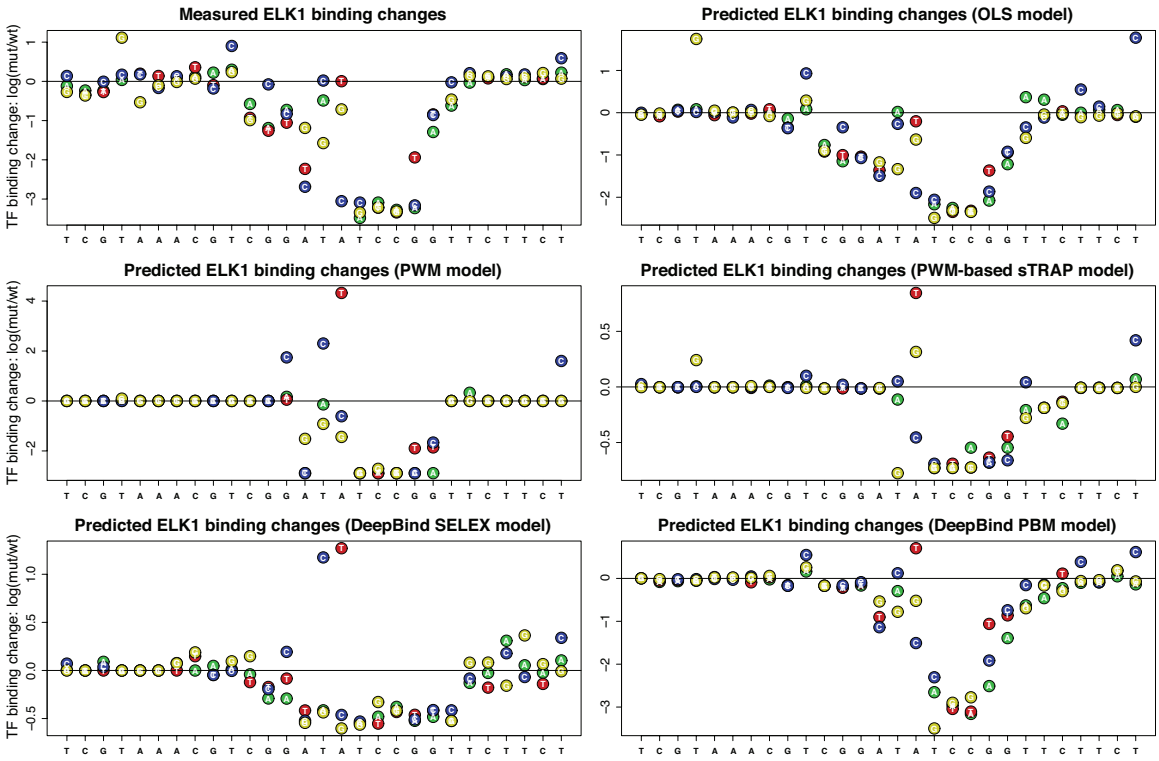

Figure 8: Measured and predicted effects of single nucleotide mutations in an ELK1 binding site and its flanking regions

3.3.2 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 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 9). 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.

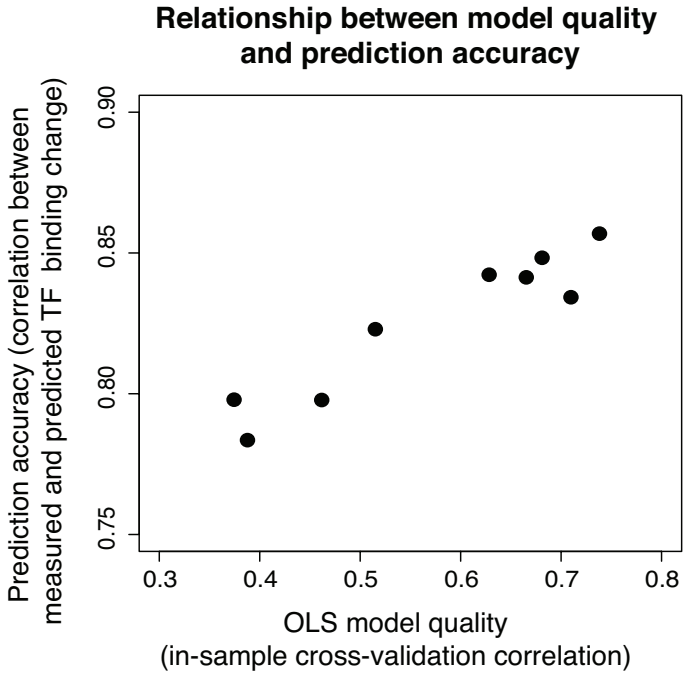

Figure 9: Relationship between OLS model quality and the prediction accuracy on independent *in vitro* mutation data. Figure shows the performance of OLS models trained on 9 different uPBM data sets

3.3.3 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. (Wagih et al., 2018). For DeepBind SELEX and PBM models we derived the binding change scores similarly to Wagih et al. (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.

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 10a), which was expected given the limitations of PWM models in capturing TF binding specificity (Shen et al., 2018). OLS models also outperformed DeepBind SELEX models trained on *in vitro* binding data from HT-SELEX experiments (Figure 10b) and DeepBind PBM models trained on *in vitro* data from PBM experiments (Figure 10c) demonstrating that, when using only DNA sequence information for training, OLS models perform best in predicting *in vivo* allele-specific binding variants.

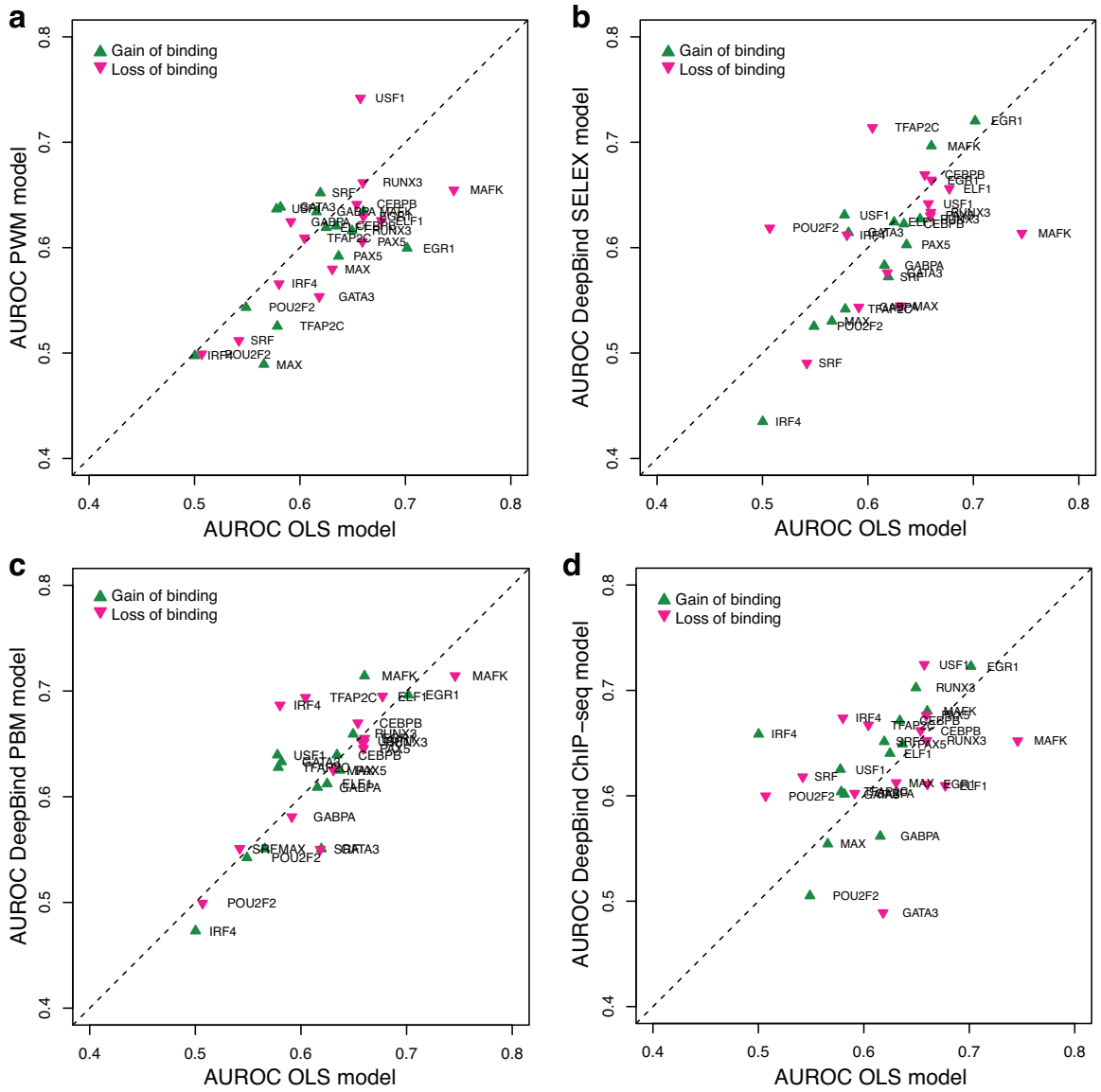

Figure 10: Performance of OLS, DeepBind, and PWM models in distinguishing between ASB and non-ASB variants identified from *in vivo* ChIP-seq data

We also compared the performance of OLS models to DeepBind models trained on *in vivo* ChIP-seq data (Figure 10d). 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.

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, 2016), 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 QBiC-Pred.

In summary, QBiC-Pred 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 ~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). 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. Utilizing accurate transcription factor binding change predictions to identify regulatory driver mutations in cancer

In this chapter, we make use of our fast and accurate TF binding change predictions to develop TF-centric approaches to prioritize non-coding mutations that can potentially act as driver mutations. Unlike most of the existing methods, our method does not require the prioritized mutations/regions to be highly recurrent, but detect their significance by testing if they give rise to more TF binding changes than expected from random mutations within the regulatory element of interest. Another unique feature of our method is that we link enhancers and promoters to the genes they regulate, combining evidence from all regulatory elements of each gene to infer whether a gene is potentially dysregulated due to regulatory mutations.

We applied our method to the LIRI-JP (virus associated hepatocellular carcinoma) data set from ICGC and identified potentially dysregulated genes whose regulatory mutations in their promoters and/or enhancers could trigger significant TF binding changes. The results demonstrated that the gene expression change is more drastic for the potentially dysregulated genes compared to a set of control genes that also have regulatory mutations. At last, we adapted our method to perform patient-level analysis and identify potentially dysregulated genes for each patient. Interestingly, the set of genes that are identified across multiple patients is enriched with cancer prognostic genes.

Overall, our results indicate that TF-centric approaches have the potential to identify dysregulated cancer genes that cannot be discovered by existing recurrence-based methods.

4.1 Data

4.1.1 ICGC simple somatic mutations (SSM) data

We chose the LIRI-JP project from ICGC for our study because it has a fairly large number of donors with WGS simple somatic mutations data (258 donors) as well as gene expression data (232 donors).

The SSM data set contains ~3.8 million mutations, most of which are SNVs (~3.5 million). Other mutation types in the data set are indels and multiple base substitutions (MSub). Since we want to focus on predicting the effect of TF binding change for SNVs, we filtered out other types of mutations. For each mutation, the information we want to keep is its position in the genome, its reference and alternative alleles, and which donor it is from. The ICGC mutation ID is also kept for easy lookup, resulting in the data format in Table 4.

Table 4: Format of ICGC simple somatic mutations (SSM) data after initial data processing

chromosome	start	end	icgc_mutation_id	icgc_donor_id	ref	mut
chr1	20252	20252	MU76919899	DO45117	A	G
chr1	54366	54366	MU1027595	DO23518	A	G
chr1	63684	63684	MU5724057	DO48730	G	A
chr1	66257	66257	MU76862748	DO45287	T	A
chr1	77904	77904	MU1277428	DO23543	A	G
…	…	…	…	…	…	…

4.1.2 ICGC sequence-based gene expression (EXP-S) data

The EXP-S data set contains the RNA-Seq results for samples corresponding to 232 donors. Most of the donors have gene expression data available for both cancer cells and normal liver cells, allowing for direct comparisons between the two types of cells. From the data available in ICGC, the fields of interest for our analyses are donor ID, gene, normalized read count, and cell type, resulting in the data format in Table 5.

Table 5: Format of ICGC gene expression data after initial data processing

icgc_donor_id	gene	norm_count	submitted_sample_id
DO227643	A1BG	822.797991	RK039_Cancer
DO227643	A1BG	999.247366	RK039_Liver
DO227643	A1BG-AS1	164.229661	RK039_Cancer
DO227643	A1BG-AS1	190.027983	RK039_Liver
DO227643	A1CF	24.5990974	RK039_Cancer
DO227643	A1CF	21.6770884	RK039_Liver
…	…	…	…

4.1.3 Promoters

TF binding sites are located in the promoter and/or enhancer regions of their target genes. Given our focus on mutations that affect TF binding, we need to specify where promoters and enhancers are located across the genome.

Promoters are defined as DNA sequences near transcription start sites (TSS), which provide the regulatory information necessary for transcription initiation. However, promoters do not have clear boundaries, and the lengths of promoters are quite arbitrary and vary from study to study. In our analysis, we consider promoters for all protein-coding genes and specify promoters as the DNA sequences from 1000 bp upstream to 1000 bp downstream of each RefSeq (O'Leary et al., 2016) TSS, excluding any RefSeq exon sequences. According to this definition, there are 21,543 promoters, taking up ~1.5% of the genome.

Our methods can be applied to promoters defined on other gene sets, such as genes annotated in GENCODE (Frankish et al., 2019).

4.1.4 Enhancers

Similar to promoters, our methods work with any enhancer data set. The quality of the data set, however, can significantly influence the results. In our analysis, we used the largest set of experimentally determined enhancers, from the FANTOM project (Andersson et al., 2014; Lizio et al., 2015). In addition, we note that: 1) this enhancer set has been frequently used in recent publications that analyze non-coding mutations (e.g.

(Weinhold et al., 2014; Khurana et al., 2016)), 2) these enhancers do not overlap with each other, which is a convenient feature when we combine results across them, and 3) FANTOM provides linkage between enhancers and associated TSSs, which is critical for being able to connect our enhancer results to gene expression data. Most of the FANTOM enhancers are between 100 bp and 500 bp in length. After removing the ones that overlap with the promoters defined in Section 4.1.3, there are a total of 41,254 enhancers, taking up ~0.4% of the genome.

The FANTOM project uses Cap Analysis of Gene Expression (CAGE) to map the sets of transcripts, transcription factors, promoters and enhancers in the majority of mammalian primary cell types and a series of cancer cell lines. To build the linkage between enhancers and TSSs, they use expression correlation to identify all intra-chromosomal enhancer-promoter pairs within 500kb. Overall, 64% of enhancers have at least one associated TSS. On average, a TSS is associated with 4.9 enhancers and an enhancer with 2.4 TSSs.

We are aware that other approaches to determine the linkages between enhancers and the genes they regulate exist, such as the Roadmap epigenomics enhancer-gene linking (Bernstein et al., 2010). In our analysis, we use the links provided by the FANTOM project since they correspond directly to the FANTOM enhancers, but our methods can work with any enhancer sets and corresponding enhancer-gene linkages.

4.1.5 The human protein atlas data

To validate the functional effect of the genes we prioritized in cancer, we downloaded pathology data from The Human Protein Atlas (Uhlen et al., 2017) (http://www.proteinatlas.org). The information we used in the pathology data is whether a gene is a prognostic marker in cancer, which is inferred by Kaplan-Meier analysis of correlation between mRNA expression level and patient survival.

4.1.6 An overview of mutation burden in promoters and enhancers

We calculated the mutation rate for all enhancers together and all promoters together. The average mutation rate is 3.6×10^{-6} in enhancers and 3.1×10^{-6} in promoters. Since most enhancers are hundreds of bp long and promoters thousands of bp, the majority of enhancers and a fair number of promoters do not contain any mutation (Table 6). In the current stage, our method can only make statistical inference on regulatory elements that contain mutations (a potential way to make inference on all elements will be discussed in Chapter 5). After removing these elements without mutations, we are left with 9,018 enhancers and 12,612 promoters.

Table 6: An overview of mutation burden in regulatory elements

Mutation count	Number of enhancers	Number of promoters
0	32236	8931
1	7455	4776
2	1275	3937
3	222	1741
4	55	1174
>4	11	984
Total	**41254**	**21543**

4.2 Methods

4.2.1 Definition of TF binding change for each regulatory element

Most mutated regulatory elements (enhancers and promoters) only harbor one mutation (see Section 4.1.6). In this case, for a given TF, the binding change caused by the mutation can be predicted using QBiC. The result is in the form of a signed number, with positive meaning increase of binding and negative meaning decrease. When there is more than one mutation in the element, we take the prediction with the greatest absolute value to represent the TF binding change for the element, so the result is still in the form of a signed number.

More formally, suppose there are m regulatory elements with mutations, and a total of n TFs considered (n can be the number of all available TFs or the number of all TFs that are expressed in the cell type of interest). If there are k mutations in the ith element, to define the binding change of the jth TF on this element, we use QBiC to

predict the binding change due to each mutation $y_1, \ldots, y_k$, and take the prediction with the greatest absolute value.

$$t_{ij} = y_{argmax(|y_l|)} \quad l \in \{1, \ldots, k\}$$

The results can be put in an m by n matrix (Table 7).

Table 7: Format of TF binding change predictions for regulatory elements

	TF 1	**…**	**TF n**
element 1	t_{11}	…	t_{1n}
…	…		…
element m	t_{m1}	…	t_{mn}

4.2.2 Background mutation model

Suppose that there are k mutations in a regulatory element i, the question we want to ask is given the number of mutations in this element, do they cause more binding change for TF j than expected if the mutations happen randomly? Here, the null hypothesis is that all k mutations in this element are passenger mutations. Under the null hypothesis, the k mutations can happen at any position in the element and can be mutated from the reference nucleotide to any of the other three nucleotides, with probabilities specified by a background mutation model.

The background mutation model here needs to account for far less parameters than the models used by methods discussed in Section 1.4.1 for two reasons. First, as we control for the total number of mutations in the element, we do not need to estimate the

absolute rate for each mutation to happen, but only the relative rate for each mutation compared to all other possible mutations in the element. Second, each regulatory element is short, so the global parameters associated with the chromosomal location of the mutation such as DNA replication timing can be considered the same for each base in the element. The local parameter that matter most to the mutation rate is the trinucleotide context of the mutation and the nucleotide being mutated from and to. Since reverse complements represent the same trinucleotide, for example, 5′-ACG-3′ is the same as 5′-CGT-3′, there are $4^3/2 = 32$ possible trinucleotides, leading to $32 \times 3 = 96$ possible trinucleotide to trinucleotide mutation types.

To build the background mutation model, we first estimated the mutation rate of the 32 trinucleotides. This is simply done by counting their occurrences in all enhancers/promoters and the number of times they got mutated. We estimated the trinucleotide mutation rates separately for enhancers and promoters since they looked different (Figure 11). Then for each trinucleotide, we counted the proportion of the three trinucleotides they mutated to, which means given the trinucleotide is mutated, the probability that it is mutated to each of the three alternatives. Finally, we multiplied the mutation rate of each trinucleotide by the probability that it got mutated to another trinucleotide to represent the trinucleotide to trinucleotide mutation rate (illustrated in Table 8).

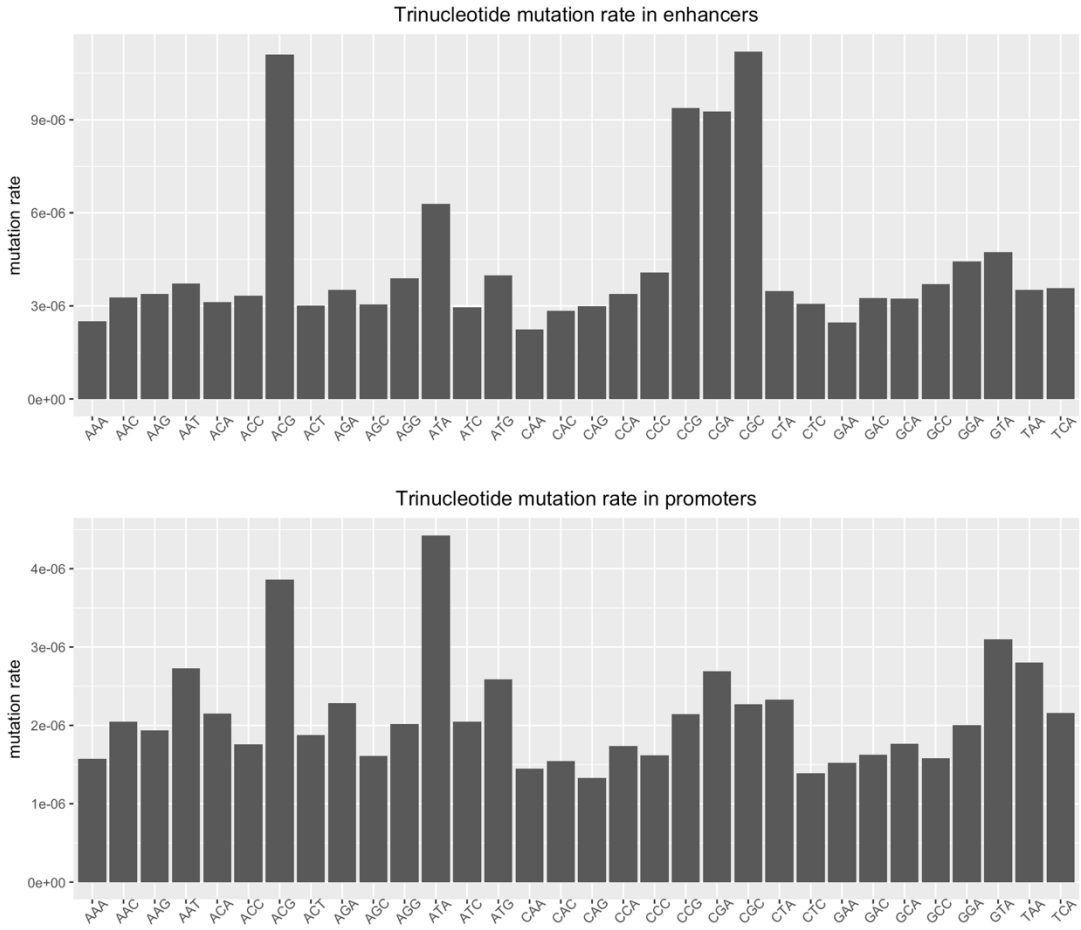

Figure 11: Trinucleotide mutation rates in enhancers and promoters

Table 8: Sample trinucleotide to trinucleotide mutation rates in promoters

ref	mut	mut_rate	proportion	tri_mut_rate
AAA	ACA	1.58E-06	0.33309456	5.25E-07
AAA	AGA	1.58E-06	0.45773639	7.21E-07
AAA	ATA	1.58E-06	0.20916905	3.30E-07
AAC	ACC	2.05E-06	0.16234498	3.33E-07
AAC	AGC	2.05E-06	0.69334837	1.42E-06
AAC	ATC	2.05E-06	0.14430665	2.96E-07
AAG	ACG	1.93E-06	0.29359286	5.68E-07
AAG	AGG	1.93E-06	0.47364152	9.16E-07
AAG	ATG	1.93E-06	0.23276561	4.50E-07
...	...	...	...	...

Now coming back to the null hypothesis, suppose that the length of the regulatory element is L bp, then there are $3L$ possible mutations. The trinucleotide to trinucleotide mutation rate tells us how likely each one of the $3L$ mutations to happen relative to other mutations. Given that there are k mutations in the element, the distribution of the number of times each possible mutation happens follows a multinomial distribution:

$$(X_1, \dots, X_{3L}) \sim Multinomial(k, (\pi_1, \dots, \pi_{3L}))$$

$(\pi_1, \dots, \pi_{3L})$ can be directly taken from Table 8 according to the reference trinucleotide and alternative trinucleotide of each possible mutation, and normalized so that they add up to 1.

4.2.3 Resampling-based background distribution of TF binding change

With the help of the background mutation model, we can resample the k mutations within the regulatory element i. This process is an unordered sampling of k mutations from a pool of $3L$ mutations with replacement, so the total number of possible outcomes is

$$N_i = \binom{3L + k - 1}{k}$$

Here, it is sampling with replacement because the mutations in this element come from different patients, which implies the exact same mutation can appear more than once. The probability for each outcome to appear is given by the multinomial distribution in Section 4.2.2.

For each resampling outcome $(x_1, \dots, x_{3L})$, we can use QBiC to predict the TF binding changes $(y_1, \dots, y_k)$ for the k mutations, and the binding change on regulatory element i can be calculated according our definition in Section 4.2.1. Therefore, we can generate a resampling-based background distribution of TF binding change for every TF (Figure 12). In the example shown in Figure 12, the length of the enhancer chr1:1124600-1124942 is $L = 343$, and there is $k = 1$ mutation in this enhancer, so there are 1029 resampling outcomes. The TF binding change predicted here is for transcription factor MYC, which ranges from -0.28 to 1.25.

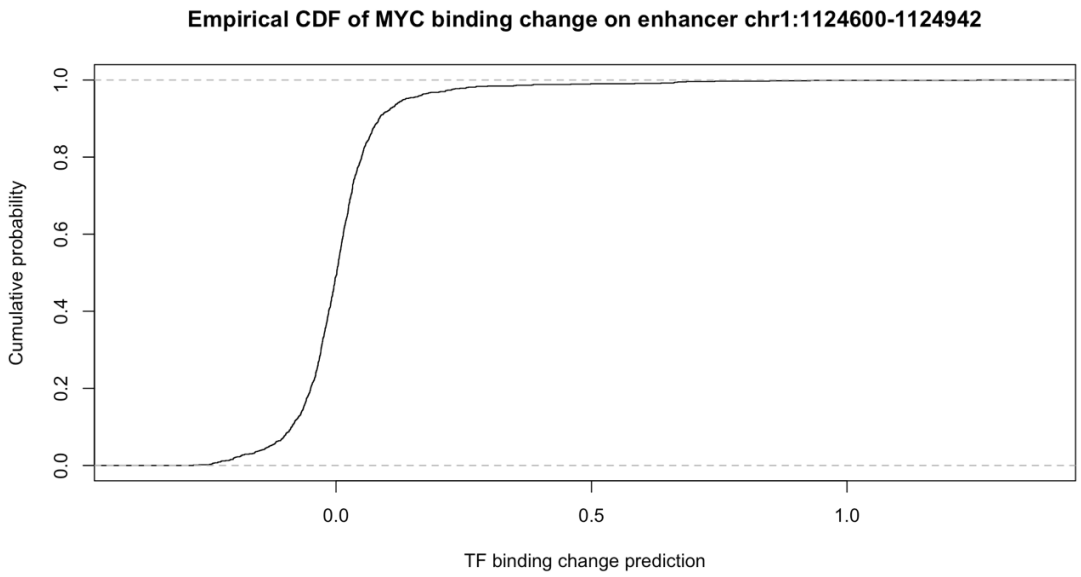

Figure 12: An example of the resampling-based background distribution. The regulatory element is enhancer chr1:1124600-1124942 (L=343), TF is MYC, and k=1

4.2.4 Testing the significance of the observed TF binding change

Using the resampling-based background distribution, we can test whether the observed TF binding change t_{ij} (Section 4.2.1) is larger than expected from random mutations and derive an empirical p-value p_{ij} by calculating the cumulative probability of seeing an equally or more extreme binding change than the observed one. Carrying out hypothesis testing for every entry in Table 7, the result is still an m by n matrix (Table 9).

Table 9: Testing the significance of observed binding change of each TF on each regulatory element

	TF 1	**…**	**TF n**
element 1	p_{11}	…	p_{1n}
…	…		…
element m	p_{m1}	…	p_{mn}

We tested increase and decrease of TF binding separately for the convenience of subsequent integration of evidence over multiple regulatory elements. In other words, we did two one-sided hypothesis testing for each observed binding change, resulting in two copies of Table 9, one for increase of binding and one for decrease. However, each pair of one-sided p-values add up to one, so we only need to store one copy of the table.

4.2.5 Integrating results across multiple regulatory elements that regulate the same gene

Regulatory driver mutations contribute to cancer initiation and development by affecting the expression of the protein-coding genes they regulate. Therefore, determining the links between cis-regulatory elements (in our work, enhancers) and their target genes is very important. We mentioned two available linkage data sets in Section 4.1.4. However, even when the links between regulatory regions and target genes are known, it still remains a challenge to study the effects of mutations in all elements controlling gene expression in a comprehensive manner.

So far, we have answered the following question: given the number of mutations in a regulatory element, do these mutations cause a significantly larger binding change, for a given TF, than expected if the mutations happened randomly? Now we want to integrate the results for multiple regulatory elements that regulate the same gene.

Suppose that a gene has k mutated regulatory elements, then for a given TF, we can compute one-sided p-values $p_1, \dots, p_k$ for the observed binding change in each element. Intuitively, if all the mutated elements show weak evidence of increased binding, we want to reinforce this evidence; if some mutated elements show strong evidence of increased binding, while the other mutated elements show strong evidence of decreased binding, they contradict each other and we do not want to assert that there is a significant TF binding change in the regulatory region of the gene.

The way we define enhancers and promoters ensures that there is no overlap between them. Therefore, we can assume that the p-values $p_1, \dots, p_k$ are independent and apply Stouffer's method (Stouffer et al., 1949) to combine the p-values. If we let $Z_i = \Phi^{-1}(1 - p_i)$ $(i = 1, \dots, k)$, where Φ is the standard normal cumulative distribution function, then

$$Z = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}}$$

is a z-score for the regulatory elements combined. Then we use this z-score to derive a two-sided p-value to show the significance level of TF binding change in all mutated regulatory regions of the gene.

It is worth to mention here that we can also introduce weights in Stouffer's method, if we let

$$Z = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$

then it is z-score that follows standard normal distribution under the null hypothesis. The weights can be a function of the length of the regulatory element and the number of mutations to account for the fact that the hypothesis testing in each element is not equally sized.

After integrating evidence across multiple regulatory elements of the same gene, the results are two-sided p-values for each gene (Table 10). Based on these p-values, we can prioritize genes whose regulation is significantly affected by the mutations in their regulatory elements through a change in TF binding.

Table 10: Testing the significance of observed TF binding changes in all regulatory elements combined for each gene

	TF 1	**…**	**TF n**
gene 1	p_{11}	…	p_{1n}
…	…		…
gene *l*	p_{l1}	…	p_{ln}

Lastly, we use Hochberg (Hochberg, 1988) correction to adjust p-values for multiple testing.

4.2.6 Adaptation of our method to perform patient-level analysis

The method we've introduced so far uses mutations from all patients combined. However, with minor modifications, we can perform the same analysis for each patient. On patient level, the number of mutated regulatory elements and the number of mutations they contain are much less than all patients combined. Thus, most of the existing methods are not able to detect any driver gene since the mutation burden in each patient is not large enough. However, our way of driver identification does not require the regulatory elements to be recurrently mutated. Even if there is only one mutation in the regulatory elements of a gene, it can still turn out to be significant provided that the mutation causes a large enough TF binding change. Therefore, our method can better adapt to small data sets or patient-level analyses.

The modifications we need to make are in the resampling process. In Section 4.2.3, we sampled with replacement from all possible mutations in the regulatory region. For patient-level analysis, we should sample without replacement because a patient cannot have the same mutation.

4.3 Results

4.3.1 Analytical and simulation-based implementations of our method generate similar result

There are two ways to implement the resampling process (Section 4.2.3) in our method. Here, the context is that there are k mutations in a regulatory element i, whose length is L, and we want to obtain the background TF binding change distribution in this

element by resampling the k mutations from $3L$ possible mutations and predict the TF binding change in each sample.

One solution is to simulate the k mutations according to the background mutation model (Section 4.2.2), and predict the TF binding change in each simulation to generate the background distribution. The simulation process is a random sampling of k mutations from a pool of $3L$ possible mutations, with replacement for the aggregated analysis and without replacement for the patient-level analysis. The probability weights for the $3L$ mutations to be drafted are $(\pi_1, \dots, \pi_{3L})$ determined by the trinucleotide context of each mutation. The simulation-based implementation is simple to understand and can work for both resampling with and without replacement. However, simulations are time consuming, and it is currently impossible to generate the background distributions for every entry in Table 7 (tens of thousands of rows, about 600 columns, which means we would need to generate millions of background distributions).

An alternative solution is to directly calculate the probability of each possible resampling outcome. For resampling with replacement, we can calculate it for the

$$N_i = \binom{3L + k - 1}{k}$$

possible outcomes using the probability mass function (pmf) of the multinomial distribution $Multinomial(k, (\pi_1, \dots, \pi_{3L}))$. In other words, we pre-compute the N_i probabilities and predict the TF binding change for each possible outcome in order to derive the background distribution, which takes a much shorter amount of time than

simulations. The drawback of this analytical approach is that it requires us to know the distribution of resampling outcomes, so that we can directly calculate their probability. For resampling without replacement, there is no closed form for the pmf, so we need to find approximations for the probability. For example, in our patient-level analysis, we first use the previous multinomial pmf to generate $\binom{3L+k-1}{k}$ probabilities. However, only $\binom{3L}{k}$ of them correspond to possible outcomes, and the rest are invalid outcomes because they contain replicates of the same mutation. After discarding the invalid ones, we rebalance the $\binom{3L}{k}$ probabilities so that they add up to one and use them as approximations for the actual probabilities.

We confirmed that the results from analytical and simulation-based approaches agree with each other (Figure 13). In Figure 13, we plotted the p-values of MYC binding change in all 9018 mutated enhancers (the column for TF MYC in Table 9). The x coordinate is the p-value of the MYC binding change calculated by the analytical approach, and the y coordinate is the p-value calculated by the simulation-based approach. The number of runs in the simulation is 1,000,000. The results in the upcoming sections all use the analytical implementation of our method.

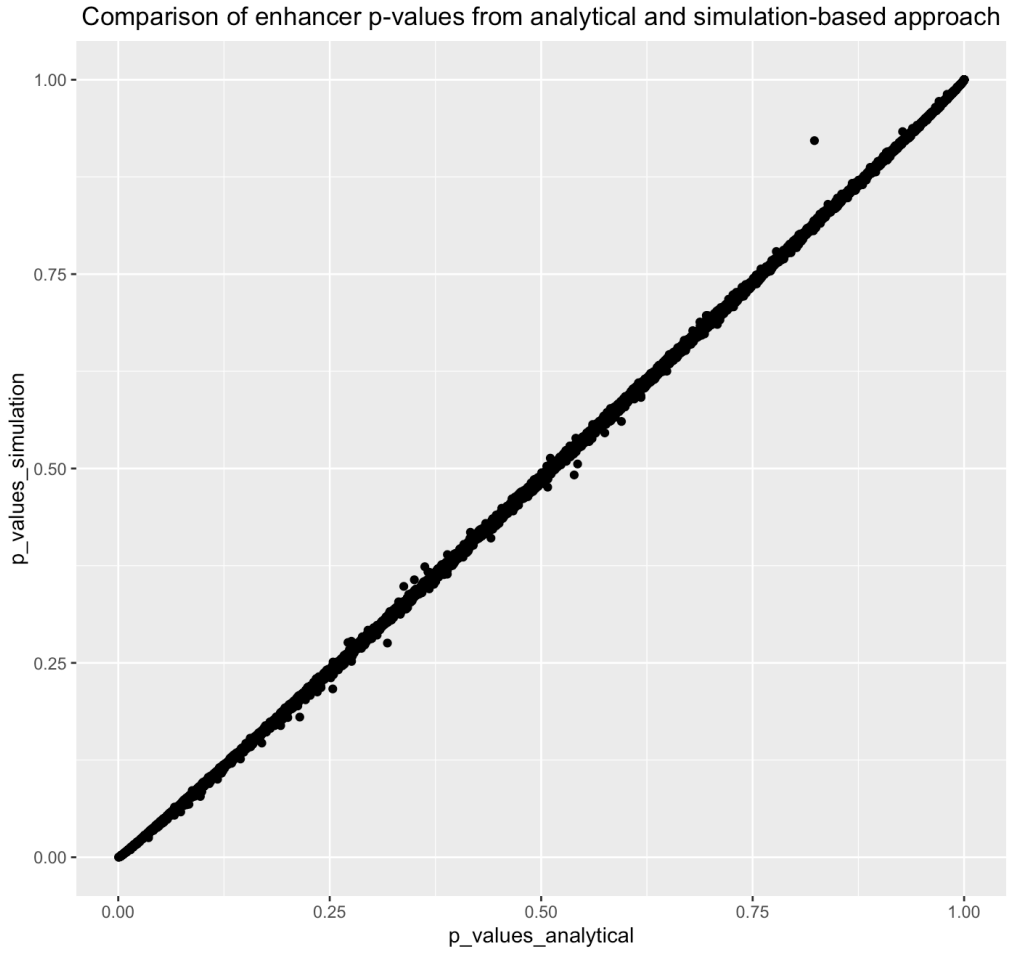

Figure 13: Comparison of p-values of MYC binding change in 9018 enhancers calculated from analytical and simulation-based approach

4.3.2 Integrated analysis across multiple regulatory elements of each gene identifies 82 genes whose regulatory mutations can lead to significant TF binding changes

We implemented our method for all 582 TFs whose binding change can be predicted by QBiC (Section 3.2.1). As discussed in Section 4.1.6, the regulatory elements we considered were the mutated ones, a total of 9,018 enhancers and 12,612 promoters. We mapped these enhancers to the 5,336 genes they regulate, using the enhancer-gene

links from the FANTOM project (Section 4.1.4), and mapped these promoters to 11,721 genes by their TSSs. The number of genes is smaller than the number of promoters because some genes have multiple TSSs.

We first analyzed enhancers and promoters separately (Section 4.3.2.1 and Section 4.3.2.2), since some TFs only work on enhancers and some only on promoters (Andersson et al., 2019). Then we performed a third analysis that considered all regulatory elements of the same gene (Section 4.3.2.3) by using Stouffer's method again to combine the results from the two separated analysis.

The direct output of our analysis is an implementation of Table 10 in Section 4.2.5, which is shown in Table 11 for integrating evidence across all enhancers of the same gene. We applied the same procedure to all promoters.

Table 11: Testing the significance of observed TF binding changes in all enhancers combined for each gene

	TF 1 (ALX1)	**…**	**TF 582 (ZNF8)**
gene 1 (AADACL4)	0.181	…	0.843
…	…		…
gene 5336 (ZZEF1)	0.915	…	0.002

The entries in Table 11 are p-values before correction for multiple testing. We checked their distribution and found that the distribution of p-values in each column is approximately a uniform distribution (an example for p-values of ALX1 binding change in Figure 14). This is as expected because the majority of mutations are passenger

mutations and the majority of genes are not contributing to tumorigenesis. Therefore, the null hypothesis should hold for most of the 5,336 tests, leading to uniformly distributed p-values.

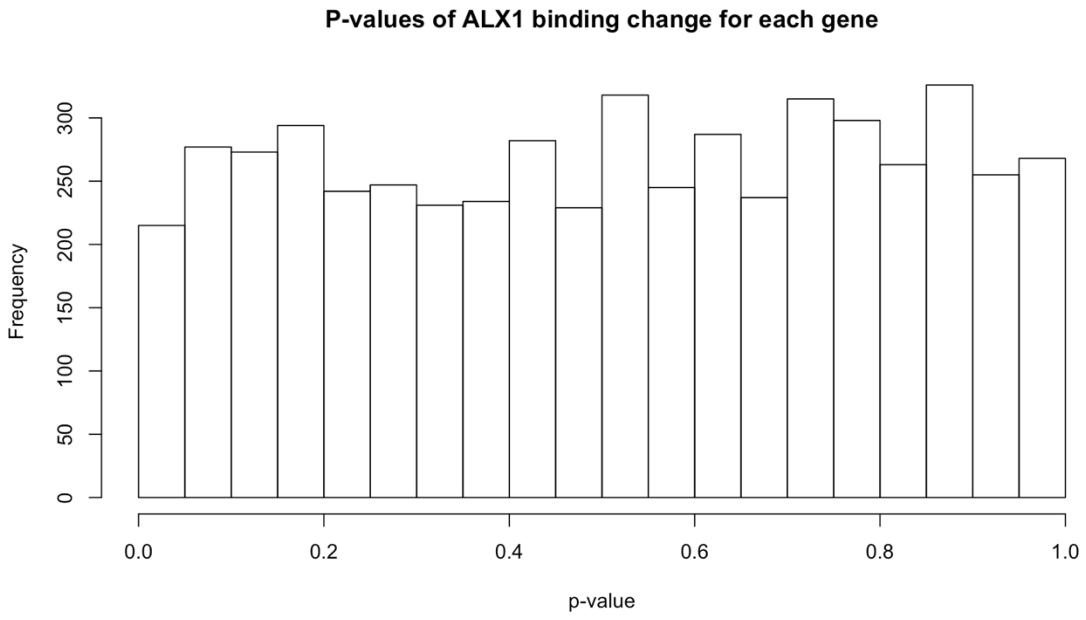

Figure 14: A histogram of the p-values of ALX1 binding change for each of the 5336 genes

The direct output, which is in the form of a large matrix, is not straightforward for interpretation or prioritization of candidate driver genes. Therefore, we want to integrate the information in the matrix and present it in a more compact form. Initially we aimed to do a meta-analysis by combining results from all 582 TFs. However, it is hard to combine p-values across columns because they are not independent. For example, paralogous factors have similar binding preferences, so their binding changes due to mutations are correlated.

One possible way to prioritize potentially dysregulated genes is by their smallest adjusted p-values across all TFs (i.e. row minimum). If the smallest adjusted p-value is less than 0.05, we know that the binding change of at least one TF is significant in the regulatory elements of the gene. According to this criterion, we have identified 82 potentially dysregulated genes whose regulatory mutations can lead to significant TF binding changes.

4.3.2.1 Analysis on all enhancers of each gene identifies 5 significant genes

After adjusting for multiple testing, 5 out of the 5,336 genes that have enhancer mutations are considered significant (Table 12). The table shows the gene name, the number of mutations in all of its enhancers, the minimum p-value before and after adjusting for multiple testing, and the TF whose binding changes the most and results in this minimum p-value.

Table 12: Prioritized genes from all enhancers analysis

gene	count	min_p	min_adjusted_p	TF
ATXN3	6	4.81E-07	0.0026	RUNX1
POLM	3	6.44E-07	0.0034	KMT2A
CEBPD	3	1.04E-06	0.0056	TFAP4
AP1S2	3	3.08E-06	0.0164	GCM1
PLEKHM2	4	8.73E-06	0.0466	DLX1

ATXN3 (Ataxin-3) is a member of the deubiquitinating enzymes (DUBs). These enzymes catalyze the removal of ubiquitin from protein substrates and regulate several aspects of protein fate. ATXN3 has been shown to restrict transcription of PTEN, a

tumor suppressor gene and major antagonist of the phosphatidylinositol-3-kinase (PI3K) pathway commonly hyperactivated in cancer (Sacco et al., 2014). Interestingly, in our study, cancer patients with mutations in the regulatory elements of ATXN3 gene show a significantly higher ATXN3 expression than patients without mutations ($p = 0.04314$, Wilcoxon rank-sum test, details in Section 4.3.3).

CEBPD (CCAAT/enhancer binding protein delta) is an important transcription factor regulating the expression of genes involved in immune and inflammatory responses. It is a prognostic marker in renal cancer and breast cancer according to The Human Protein Atlas (Uhlen et al., 2017). CEBPD amplification and overexpression has been shown to be a driver of tumor metastasis in urothelial carcinoma (Wang et al., 2015). In our study, cancer patients with mutations in the regulatory elements of CEBPD gene exhibit an almost significantly higher CEBPD expression than patients without mutations ($p = 0.05821$, Wilcoxon rank-sum test, Appendix A).

4.3.2.2 Analysis on all promoters of each gene identifies 74 significant genes

After adjusting for multiple testing, 74 out of the 11,721 genes that have promoter mutations are significant. In Table 13, we list the top 10 genes according to their minimum adjusted p-values. Notably, we do not see any overlap between the genes prioritized by enhancers and by promoters. One reason is that some genes only have either enhancer or promoter mutations, but not both, so they only appear in one

analysis. Another reason is that some TFs work mostly on enhancers or promoters, as discussed in Section 4.3.2.

Table 13: Prioritized genes from all promoters analysis

gene	count	min_p	min_adjusted_p	TF
FDPS	4	2.01E-08	0.0002	HBP1
ZNF248	5	3.57E-08	0.0004	ATF1
C1R	6	3.86E-08	0.0005	CXXC1
WFIKKN2	4	7.90E-08	0.0009	LBX1
CCNH	4	1.05E-07	0.0012	ELK1
RNF32	6	1.15E-07	0.0013	HNF1A
TECR	8	1.38E-07	0.0016	HOXA13
ANGPTL6	8	1.42E-07	0.0017	EGR1
EHBP1	8	1.87E-07	0.0022	TCF12
A4GALT	4	2.04E-07	0.0024	RORA

FDPS (Farnesyl diphosphate synthase) is a mevalonate pathway enzyme that is highly expressed in several cancers. Although the mechanistic, functional, and clinical significance of FDPS in cancer remains mostly unexplored, a recent study shows that FDPS plays an oncogenic role in PTEN-deficient prostate cancer through GTPase/AKT axis (Seshacharyulu et al., 2019). Unfortunately, there is no gene expression data available for patients in the LIRI-JP project who have mutations in the regulatory elements of the FDPS gene, but it is a prognostic marker in liver cancer according to The Human Protein Atlas (Uhlen et al., 2017), indicating that there is a strong correlation between its expression level and the survival time of liver cancer patients.

4.3.2.3 Analysis on all regulatory elements of each gene identifies 53 significant genes

After adjusting for multiple testing, 53 out of the 13,982 genes that have either enhancer or promoter mutations were considered significant. In Table 14, we list the top 10 genes according to their minimum adjusted p-values. A full list of the 53 significant genes can be found in Appendix A.

Of these 53 prioritized genes, 4 of them are also prioritized by the enhancer analysis, and 46 of them also prioritized by the promoter analysis. Three genes, ETS1, CELF6, and PALT1 are new findings in the combined analysis.

Table 14: Prioritized genes from all regulatory elements analysis

gene	count	min_p	min_adjusted_p	TF
FDPS	4	2.01E-08	0.0003	HBP1
C1R	6	3.86E-08	0.0005	CXXC1
KANK4	17	5.63E-08	0.0008	FOXM1
WFIKKN2	4	7.90E-08	0.0011	LBX1
CCNH	4	1.05E-07	0.0015	ELK1
CENPA	9	1.38E-07	0.0019	DMRTA2
TECR	8	1.38E-07	0.0019	HOXA13
ANGPTL6	8	1.42E-07	0.0020	EGR1
A4GALT	4	2.04E-07	0.0029	RORA
C10orf67	4	2.53E-07	0.0035	RORA

The ETS1 gene encodes a member of the ETS family of transcription factors, which are defined by the presence of a conserved ETS DNA-binding domain that

recognizes the core consensus DNA sequence GGAA/T in target genes. These proteins function either as transcriptional activators or repressors of numerous genes, and are involved in stem cell development, cell senescence and death, and tumorigenesis. Dysregulation of these transcription factors facilitates cell proliferation in cancers, and several members participate in invasion and metastasis by activating gene transcription (Fry et al., 2018). For example, ETS1 overexpression is found in human breast cancer (BC) associated with invasiveness and poor prognosis (Furlan et al., 2014). By overexpressing ETS1 or a dominant-negative mutant in BC cells, they showed that ETS1 plays a key role in coordinating multiple invasive features of cancer cells. Our analysis suggests that ETS1 is potentially dysregulated in some liver cancer patients due to regulatory mutations, and similar experiments can be designed to further explore whether ETS1 plays an oncogenic role in liver tumorigenesis.

4.3.3 Genes with significant mutations in their regulatory regions show larger expression changes

To test if the genes we prioritized based on mutations in their regulatory elements are different from randomly chosen genes, we asked whether the selected mutations lead to larger expression changes than mutations in the regulatory elements of control genes. For each gene, we compared the gene expression (normalized read counts, Section 4.1.2) of patients with regulatory mutations versus the gene expression of patients without regulatory mutations, and we used the negative log of the p-value from a Wilcoxon rank-sum test of the two populations to represent the expression change.

For example, ATXN3 is prioritized in both all enhancers and all regulatory elements analyses. There are 9 patients with ATXN3 regulatory mutations and 8 of them have expression data. The other 249 patients don't have ATXN3 regulatory mutations and 229 of them have expression data. We compared the ATXN3 expression between these two groups of patients (Figure 15). In general, patients with ATXN3 regulatory mutations have higher ATXN3 expression than patients without mutations ($p = 0.04314$, Wilcoxon rank-sum test). We took the negative log of the p-value ($-\log(p) = 3.143$) to represent the expression change.

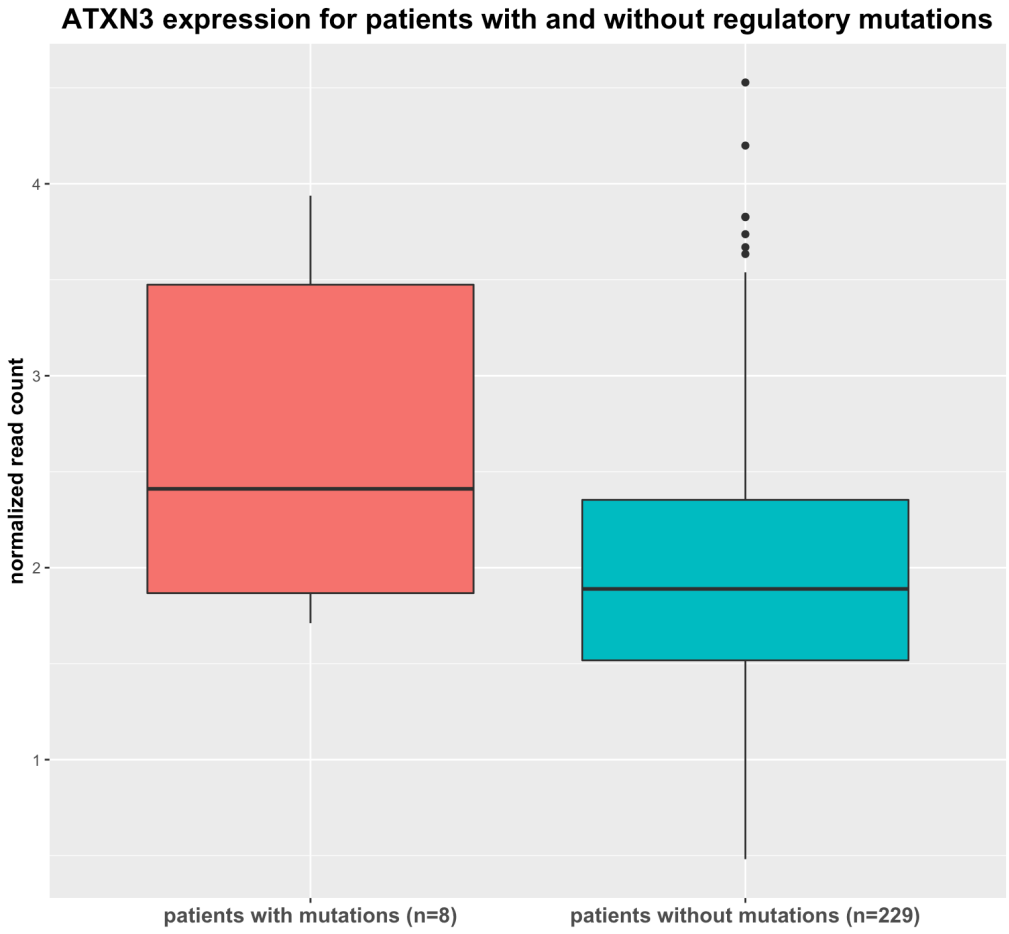

Figure 15: Patients with ATXN3 regulatory mutations (red) have higher AXTN3 expression than patients without regulatory mutations (green)

We calculated the gene expression change for all 13,982 genes that have regulatory mutations. For prioritized genes, we chose the ones identified from the all regulatory elements analysis (Section 4.3.2.3), i.e. a total of 53 genes (details in Appendix A). For controls, we used the rest of the 13,982 genes, i.e. a total of 13,919 genes. After filtering out the genes without available expression data, we obtained 41 prioritized genes and 10,897 control genes.

Figure 16 shows a direct comparison between the expression changes due to regulatory mutations for our prioritized genes versus the control genes. The expression changes are significantly larger in the prioritized group ($p = 0.00148$, Wilcoxon rank-sum test). A non-parametric test is used because we want to ensure that the transformation applied to calculate the expression change (in our case, negative log) does not affect the p-value in this test, provided that it is a monotonic transformation.

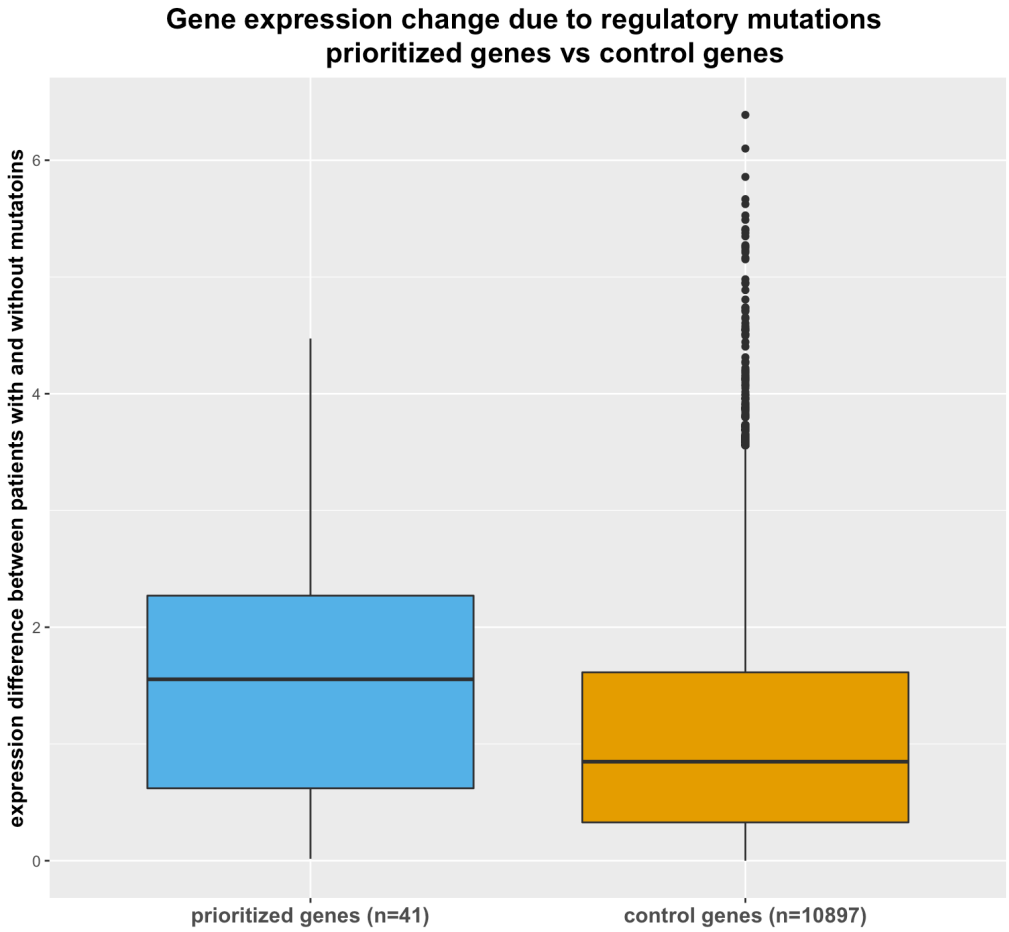

Figure 16: Genes prioritized by our analysis (blue) have larger expression changes due to regulatory mutations compared to control genes (yellow)

4.3.4 The gene set prioritized by the patient-level analysis is enriched with cancer prognostic genes

The aggregated analyses presented above combine mutations from different patients to gain more statistical power in order to identify significant regulatory mutations. However, a potential disadvantage of this analysis is that it may not work well on a population of patients that are highly heterogeneous. Therefore, we developed the patient-level analysis as a complementary method to address this situation.

Basically, we can think of the patient-level analysis as applying the aggregated analysis procedure to each individual patient, with minor modifications, as described in Section 4.2.6. The direct output for each patient is a much shorter version of Table 11 (because the number of mutated regulatory elements in each patient is small), but with the same number of columns. In the LIRI-JP data set, individual patients have 42 mutated enhancers and 158 mutated promoters on average.

We applied the patient-level analysis on all-enhancers of each gene for all 258 patients in the LIRI-JP data set. For each patient, we identified a set of significant genes in the same format as Table 12 in Section 4.3.2.1. Therefore, we had 258 sets of genes (some are empty sets). Then, we ranked the genes by the number of sets they appeared in and prioritized top genes in the list for functional validation (Table 15).

Table 15: Top 5 genes identified in the patient-level analysis

gene	occurrence	function
KLF6	8	Prognostic marker in renal cancer (favourable)
NR3C1	7	Prognostic marker in renal cancer (unfavourable) and endometrial cancer (unfavourable)
BRD2	5	Gene product is not prognostic
CXCR4	5	Prognostic marker in renal cancer (unfavourable), ovarian cancer (favourable) and stomach cancer (unfavourable)
FPR1	5	Prognostic marker in renal cancer (unfavourable)

Interestingly, these top genes are enriched in prognostic markers annotated by The Human Protein Atlas (introduced in Section 4.1.5), indicating most of them are cancer-related genes. We also notice that none of the 5 genes is close to being prioritized in the aggregated analysis, but they all have high regulatory mutation counts. In fact, for a gene to be prioritized in the patient-level analysis, it has to have regulatory mutations in several patients, so the patient-level analysis depends more on the recurrence of mutations than the aggregated analysis.

Table 16: Results from population-level analysis for the top 5 genes identified in the patient-level analysis

gene	count	min_p	min_adjusted_p	TF
KLF6	30	0.0001	0.5884	FOXC1
NR3C1	19	0.0063	1	TGIF1
BRD2	15	0.0017	1	NKX2-2
CXCR4	35	0.0028	1	GMEB1
FPR1	21	0.0093	1	ATF1

4.4 Discussion

In summary, we have developed new approaches to analyze regulatory mutations and identify potentially dysregulated genes in cancer patients. Notably, the set of potentially dysregulated genes we have identified have larger gene expression effects than random gene sets, and are enriched for prognostic genes of cancer.

Here, we used the LIRI-JP data set to illustrate the workflow of our method. But the method is applicable to any cancer mutation study, including any simple somatic mutation data set in ICGC. In addition, our method can work both on a large data set that aggregates mutations from hundreds of patients or on a small set of mutations from an individual patient. Overall, it is a versatile tool that can be applied to a wide range of cancer projects.

While our method provides a new perspective to identify potential driver genes and regulatory driver mutations, and is supported by gene expression data and

previously identified cancer prognostic marker data, experimental follow-up will be needed to test whether the regulatory mutations identified by our method affect the expression of target genes and subsequently promote the survival and/or proliferation of cancer cells. We will discuss possible improvements and validations in Chapter 5.

Chapter 5. Conclusions

In this work, we have developed new approaches to identify putative regulatory driver mutations in cancer, based on new methodology for predicting the quantitative effects of single nucleotide variants on TF binding. Compared to previous work, our method is innovative in that it does not require the driver mutations to be highly recurrent; instead, we assess the mutations' significance by testing if they cause larger TF binding changes than expected in the case of completely random mutations.

In order to implement our method, we first established an OLS regression model to quantitatively predict TF binding changes due to single nucleotide mutations, taking into account the mutations' sequence context (Chapter 2). Then, we validated our OLS model on multiple *in vitro* and *in vivo* data sets (Chapter 2 and 3), and demonstrated that it outperforms state-of-the-art sequence-based prediction tools (Chapter 3). Collaborating with Computer Science PhD candidate Vincentius Martin, we optimized the OLS model predictions so that it could be deployed on large cancer mutation data sets for hundreds of TFs.

Utilizing this fast and accurate TF binding change prediction tool, which we call QBiC-Pred, we generated predictions for all single nucleotide variants in the LIRI-JP liver cancer mutation data set from ICGC. To test whether the predicted binding changes were significant, we devised a resampling-based approach to derive the background distribution of TF binding changes in each enhancer and promoter.

Since transcription initiation relies on the cooperation of multiple enhancers and the promoters of the gene, we mapped enhancers to the genes they regulate using enhancer-TSS links provided by the FANTOM project (Lizio et al., 2015) and combined the effects across multiple regulatory elements using Stouffer's method (Stouffer et al., 1949). Therefore, we could perform hypothesis testing at the gene level and identify genes whose regulation is likely to be significantly perturbed by the mutations observed in their regulatory elements, through changes in TF binding.

For the potentially dysregulated genes identified from our analysis, we found that their expression changes were significantly larger than the changes of genes not identified as dysregulated. Among these potentially dysregulated genes, the subset of genes whose expression changes were abnormally large could be of particular interest for follow-up experimental validations, because 1) these genes contain regulatory mutations that could significantly change the binding of specific TFs and 2) the gene expression of patients with regulatory mutations is significantly different from patients without regulatory mutations.

It is worth noting that most driver identification approaches, including ours, combine mutations from different patients to gain more statistical power to identify significant regulatory mutations. However, a potential disadvantage is that it may not work well on a population of patients that are highly heterogeneous. For approaches that identify drivers by their mutational burden, it is impossible to run the analysis on

the patient-level because the number of mutations is too small. Nevertheless, since our method does not require the driver mutations to be highly recurrent, with minor modifications it can be applied to identify potentially dysregulated genes for each patient. The patient-level analysis needs to be further refined so that it has more coherent evidence to prioritize genes for follow-up experimental validations, but it provides a very different perspective than the population-level approaches.

In the future, we can improve our method from several aspects. First, now we generate the background TF binding change distribution in a regulatory element based on a fixed number of mutation counts in the element, so the background distribution is a conditional distribution $P(binding\ change|count = k)$. If we know the probability of having k mutations in the element $P(count = k)$, we can calculate the marginal distribution $P(binidng\ change) = \sum_k P(count = k)\, P(binding\ change|count = k)$. This is especially useful if we want to include the regulatory elements without mutations into the analysis. To get $P(count = k)$, we may borrow the results from existing recurrence-based driver identification approaches (e.g. (Lochovsky et al., 2015)), since the key in those analyses is calculating the background mutation rate in each non-coding functional element. Other ways to improve our method include considering only TFs that are expressed in the tissue of interest, introducing weights when combining evidence from multiple regulatory elements (discussed in Section 4.2.5), integrating results across different TFs to more coherently identify potentially dysregulated genes

(discussed in Section 4.3.2), and running gene set enrichment analysis (GSEA) to get a functional profile of the genes we prioritize.

Overall, our TF-centric approaches use a distinctive pipeline to identify regulatory driver mutations in cancer. From the results we have seen so far, most of the potentially dysregulated genes prioritized by us either have corresponding large expression changes or are cancer prognostic genes (or both). Although experimental validations are needed to determine whether these genes actually contribute to cancer development, our results suggest that regulatory mutations should be investigated further, not just by their recurrence, but also by their functional effects such as TF binding changes, to uncover dysregulated genes that may drive tumorigenesis.

Appendix A

In appendix A, we list a full table of the 53 genes identified from our all regulatory elements analysis, which means they have potentially significant TF binding change for at least one TF in their regulatory elements. The table shows the gene name, the number of mutations in all of its enhancers and promoters, the minimum p-value (after correction for multiple comparison) for testing whether the TF binding change is larger than expected from random mutations, and the p-value for comparing the gene expression of patients with regulatory mutations versus the gene expression of patients without regulatory mutations. Twelve of these genes do not have expression data available, which results in NA in the p_value_expression column.

gene	count	min_adjusted_p_binding	p_value_expression
FDPS	4	**0.0003**	**NA**
C1R	6	**0.0005**	**0.0114**
KANK4	17	**0.0008**	**0.1106**
WFIKKN2	4	**0.0011**	**NA**
CCNH	4	**0.0015**	**0.1499**
CENPA	9	**0.0019**	**0.0196**
TECR	8	**0.0019**	**0.5937**
ANGPTL6	8	**0.0020**	**0.2276**
A4GALT	4	**0.0029**	**NA**
C10orf67	4	**0.0035**	**NA**
C10orf82	12	**0.0036**	**NA**
USP45	4	**0.0036**	**0.2698**

CHRDL1	4	**0.0037**	**0.9837**
GPN3	7	**0.0038**	**0.9628**
DPM3	5	**0.0051**	**0.0859**
ATXN3	9	**0.0054**	**0.0431**
PPP1R18	12	**0.0059**	**NA**
MYRIP	4	**0.0080**	**0.7625**
POLM	3	**0.0090**	**0.4331**
CCDC191	8	**0.0094**	**NA**
SLC35F5	4	**0.0119**	**0.2635**
MYH2	9	**0.0120**	**0.1030**
EHBP1	9	**0.0131**	**0.1901**
CEBPD	3	**0.0146**	**0.0582**
ATL3	4	**0.0152**	**0.1811**
C18orf54	11	**0.0153**	**0.3661**
CERS3	6	**0.0156**	**NA**
TMED2	6	**0.0171**	**0.1584**
DTWD2	4	**0.0174**	**0.1418**
RUFY2	4	**0.0174**	**NA**
ARPIN	4	**0.0187**	**NA**
ZNF417	4	**0.0204**	**0.9592**
NPAS1	2	**0.0208**	**0.2177**
PRICKLE3	5	**0.0231**	**0.1161**
TRMT44	6	**0.0235**	**NA**
ZNF347	8	**0.0243**	**0.1001**
MTMR11	9	**0.0245**	**0.7061**
PEX26	4	**0.0255**	**0.1128**
ETS1	24	**0.0272**	**0.1607**

DMTF1	8	**0.0277**	**0.2114**
CSPG5	5	**0.0280**	**0.1032**
ZNF526	12	**0.0284**	**0.6449**
MRGPRX2	11	**0.0348**	**0.8333**
CELF6	8	**0.0357**	**0.7793**
ZNF561	6	**0.0357**	**0.0914**
PATL1	4	**0.0395**	**0.5223**
ZNF248	6	**0.0408**	**0.6625**
TM4SF18	8	**0.0410**	**0.0497**
AP1S2	3	**0.0430**	**0.2473**
TXNL4B	8	**0.0438**	**0.3384**
CTNNA3	3	**0.0461**	**0.0862**
HIGD1A	8	**0.0464**	**0.5377**
SPACA3	5	**0.0498**	**NA**

References

Maurano, Matthew T., et al. "Systematic localization of common disease-associated variation in regulatory DNA." *Science* 337.6099 (2012): 1190-1195.

Weinhold, Nils, et al. "Genome-wide analysis of noncoding regulatory mutations in cancer." *Nature genetics* 46.11 (2014): 1160-1165.

Khurana, Ekta, et al. "Role of non-coding sequence variants in cancer." *Nature Reviews Genetics* 17.2 (2016): 93.

Maston, Glenn A., Sara K. Evans, and Michael R. Green. "Transcriptional regulatory elements in the human genome." *Annu. Rev. Genomics Hum. Genet.* 7 (2006): 29-59.

Heidenreich, Barbara, et al. "TERT promoter mutations in cancer development." *Current opinion in genetics & development* 24 (2014): 30-37.

Mansour, Marc R., et al. "An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element." *Science* 346.6215 (2014): 1373-1377.

Puente, Xose S., et al. "Non-coding recurrent mutations in chronic lymphocytic leukaemia." *Nature* 526.7574 (2015): 519-524.

Rheinbay, Esther, et al. "Analyses of non-coding somatic drivers in 2,658 cancer whole genomes." *Nature* 578.7793 (2020): 102-111.

Campbell, Brittany B., et al. "Comprehensive analysis of hypermutation in human cancer." *Cell* 171.5 (2017): 1042-1056.

Ng, Pauline C., and Steven Henikoff. "SIFT: Predicting amino acid changes that affect protein function." *Nucleic acids research* 31.13 (2003): 3812-3814.

Adzhubei, Ivan A., et al. "A method and server for predicting damaging missense mutations." *Nature methods* 7.4 (2010): 248-249.

Lochovsky, Lucas, et al. "LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations." *Nucleic acids research* 43.17 (2015): 8123-8134.

Rheinbay, Esther, et al. "Recurrent and functional regulatory mutations in breast cancer." *Nature* 547.7661 (2017): 55-60.

Lochovsky, Lucas, Jing Zhang, and Mark Gerstein. "MOAT: efficient detection of highly mutated regions with the Mutations Overburdening Annotations Tool." *Bioinformatics* 34.6 (2018): 1031-1033.

Fu, Yao, et al. "FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer." Genome biology 15.10 (2014): 1-15.

Zhu, Helen, et al. "Candidate cancer driver mutations in distal regulatory elements and long-range chromatin interaction networks." *Molecular Cell* 77.6 (2020): 1307-1321.

Shuai, Shimin, Steven Gallinger, and Lincoln Stein. "Combined burden and functional impact tests for cancer driver discovery using DriverPower." *Nature communications* 11.1 (2020): 1-12.

Gan, Kok A., et al. "Identification of single nucleotide non-coding driver mutations in cancer." *Frontiers in genetics* 9 (2018): 16.

Zhao, Jingkang, et al. "Quantifying the impact of non-coding variants on transcription factor-DNA binding." *International Conference on Research in Computational Molecular Biology*. Springer, Cham, 2017.

Fredriksson, Nils Johan, et al. "Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature." *PLoS genetics* 13.5 (2017): e1006773.

Robasky, Kimberly, and Martha L Bulyk. "UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions." *Nucleic acids research* vol. 39,Database issue (2011): D124-8. doi:10.1093/nar/gkq992

Jolma, Arttu et al. "DNA-binding specificities of human transcription factors." *Cell* vol. 152,1-2 (2013): 327-39. doi:10.1016/j.cell.2012.12.009

Mathelier, Anthony et al. "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles." *Nucleic acids research* vol. 42,Database issue (2014): D142-7. doi:10.1093/nar/gkt997

Weirauch, Matthew T et al. "Determination and inference of eukaryotic transcription factor sequence specificity." *Cell* vol. 158,6 (2014): 1431-1443. doi:10.1016/j.cell.2014.08.009

Bulyk, Martha L et al. "Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors." *Nucleic acids research* vol. 30,5 (2002): 1255-61. doi:10.1093/nar/30.5.1255

Udalova, Irina A et al. "Quantitative prediction of NF-kappa B DNA-protein interactions." *Proceedings of the National Academy of Sciences of the United States of America* vol. 99,12 (2002): 8167-72. doi:10.1073/pnas.102674699

Zhao, Yue et al. "Improved models for transcription factor binding site identification using nonindependent interactions." *Genetics* vol. 191,3 (2012): 781-90. doi:10.1534/genetics.112.138685

Tomovic, Andrija, and Edward J Oakeley. "Position dependencies in transcription factor binding sites." *Bioinformatics (Oxford, England)* vol. 23,8 (2007): 933-41. doi:10.1093/bioinformatics/btm055

Maerkl, Sebastian J, and Stephen R Quake. "A systems approach to measuring the binding energy landscapes of transcription factors." *Science (New York, N.Y.)* vol. 315,5809 (2007): 233-7. doi:10.1126/science.1131007

Stormo, Gary D. "Modeling the specificity of protein-DNA interactions." *Quantitative biology (Beijing, China)* vol. 1,2 (2013): 115-130. doi:10.1007/s40484-013-0012-4

Siggers, Trevor, and Raluca Gordân. "Protein-DNA binding: complexities and multi-protein codes." *Nucleic acids research* vol. 42,4 (2014): 2099-111. doi:10.1093/nar/gkt1112

Andersen, Malin C et al. "In silico detection of sequence variations modifying transcriptional regulation." *PLoS computational biology* vol. 4,1 (2008): e5. doi:10.1371/journal.pcbi.0040005

Thomas-Chollier, Morgane et al. "RSAT 2011: regulatory sequence analysis tools." *Nucleic acids research* vol. 39,Web Server issue (2011): W86-91. doi:10.1093/nar/gkr377

Ward, Lucas D, and Manolis Kellis. "Interpreting noncoding genetic variation in complex traits and human disease." *Nature biotechnology* vol. 30,11 (2012): 1095-106. doi:10.1038/nbt.2422

McVicker, Graham et al. "Identification of genetic variants that affect histone modifications in human cells." *Science (New York, N.Y.)* vol. 342,6159 (2013): 747-9. doi:10.1126/science.1242429

Rowan, Sheldon et al. "Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity." *Genes & development* vol. 24,10 (2010): 980-5. doi:10.1101/gad.1890410

Matys, V et al. "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes." *Nucleic acids research* vol. 34,Database issue (2006): D108-10. doi:10.1093/nar/gkj143

Mathelier, Anthony et al. "JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles." *Nucleic acids research* vol. 44,D1 (2016): D110-5. doi:10.1093/nar/gkv1176

Newburger, Daniel E, and Martha L Bulyk. "UniPROBE: an online database of protein binding microarray data on protein-DNA interactions." *Nucleic acids research* vol. 37,Database issue (2009): D77-82. doi:10.1093/nar/gkn660

Berger, Michael F et al. "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities." *Nature biotechnology* vol. 24,11 (2006): 1429-35. doi:10.1038/nbt1246

Berger, Michael F, and Martha L Bulyk. "Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors." *Nature protocols* vol. 4,3 (2009): 393-411. doi:10.1038/nprot.2008.195

Melnikov, Alexandre et al. "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay." *Nature biotechnology* vol. 30,3 271-7. 26 Feb. 2012, doi:10.1038/nbt.2137

Kheradpour, Pouya et al. "Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay." *Genome research* vol. 23,5 (2013): 800-11. doi:10.1101/gr.144899.112

Stenson, Peter D et al. "The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine." *Human genetics* vol. 133,1 (2014): 1-9. doi:10.1007/s00439-013-1358-4

Weirauch, Matthew T et al. "Evaluation of methods for modeling transcription factor sequence specificity." *Nature biotechnology* vol. 31,2 (2013): 126-34. doi:10.1038/nbt.2486

Granek, Joshua A, and Neil D Clarke. “Explicit equilibrium modeling of transcription-factor binding and gene regulation.” *Genome biology* vol. 6,10 (2005): R87. doi:10.1186/gb-2005-6-10-r87

Kulakovskiy, Ivan V et al. “HOCOMOCO: a comprehensive collection of human transcription factor binding sites models.” *Nucleic acids research* vol. 41,Database issue (2013): D195-202. doi:10.1093/nar/gks1089

Landrum, Melissa J et al. “ClinVar: public archive of interpretations of clinically relevant variants.” *Nucleic acids research* vol. 44,D1 (2016): D862-8. doi:10.1093/nar/gkv1222

1000 Genomes Project Consortium et al. “A global reference for human genetic variation.” *Nature* vol. 526,7571 (2015): 68-74. doi:10.1038/nature15393

McLaren, William et al. “The Ensembl Variant Effect Predictor.” *Genome biology* vol. 17,1 122. 6 Jun. 2016, doi:10.1186/s13059-016-0974-4

Martin, Vincentius, et al. "QBiC-Pred: quantitative predictions of transcription factor binding changes due to sequence variants." *Nucleic acids research* 47.W1 (2019): W127-W135.

International Cancer Genome Consortium et al. “International network of cancer genome projects.” *Nature* vol. 464,7291 (2010): 993-8. doi:10.1038/nature08987

Gray, Kristian A et al. “Genenames.org: the HGNC resources in 2015.” *Nucleic acids research* vol. 43,Database issue (2015): D1079-85. doi:10.1093/nar/gku1071

Shen, Ning et al. “Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential *In Vivo* Binding.” *Cell systems* vol. 6,4 (2018): 470-483.e8. doi:10.1016/j.cels.2018.02.009

Alipanahi, Babak et al. “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.” *Nature biotechnology* vol. 33,8 (2015): 831-8. doi:10.1038/nbt.3300

Ward, Lucas D, and Manolis Kellis. “HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease.” *Nucleic acids research* vol. 44,D1 (2016): D877-81. doi:10.1093/nar/gkv1340

Boyle, Alan P et al. “Annotation of functional variation in personal genomes using RegulomeDB.” *Genome research* vol. 22,9 (2012): 1790-7. doi:10.1101/gr.137323.112

Hume, Maxwell A et al. “UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions.” *Nucleic acids research* vol. 43,Database issue (2015): D117-22. doi:10.1093/nar/gku1045

Lambert, Samuel A et al. “The Human Transcription Factors.” *Cell* vol. 172,4 (2018): 650-665. doi:10.1016/j.cell.2018.01.029

Wagih, Omar, et al. "Allele-specific transcription factor binding as a benchmark for assessing variant impact predictors." *BioRxiv* (2018): 253427.

Shi, Wenqiang et al. “Evaluating the impact of single nucleotide variants on transcription factor binding.” *Nucleic acids research* vol. 44,21 (2016): 10106-10116. doi:10.1093/nar/gkw691

Thomas-Chollier, Morgane et al. “Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs.” *Nature protocols* vol. 6,12 1860-9. 3 Nov. 2011, doi:10.1038/nprot.2011.409

Khan, Aziz et al. “JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework.” *Nucleic acids research* vol. 46,D1 (2018): D260-D266. doi:10.1093/nar/gkx1126

Schwessinger, Ron et al. “Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints.” *Genome research* vol. 27,10 (2017): 1730-1742. doi:10.1101/gr.220202.117

Zhou, Jian, and Olga G Troyanskaya. “Predicting effects of noncoding variants with deep learning-based sequence model.” *Nature methods* vol. 12,10 (2015): 931-4. doi:10.1038/nmeth.3547

Lee, Dongwon et al. “A method to predict the impact of regulatory variants from DNA sequence.” *Nature genetics* vol. 47,8 (2015): 955-61. doi:10.1038/ng.3331

Guo, Liyuan et al. “rVarBase: an updated database for regulatory features of human variants.” *Nucleic acids research* vol. 44,D1 (2016): D888-93. doi:10.1093/nar/gkv1107

Amlie-Wolf, Alexandre et al. “INFERNO: inferring the molecular mechanisms of noncoding genetic variants.” *Nucleic acids research* vol. 46,17 (2018): 8740-8753. doi:10.1093/nar/gky686

Jolma, Arttu et al. “Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities.” *Genome research* vol. 20,6 (2010): 861-73. doi:10.1101/gr.100552.109

O'Leary, Nuala A., et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." *Nucleic acids research* 44.D1 (2016): D733-D745.

Frankish, Adam, et al. "GENCODE reference annotation for the human and mouse genomes." *Nucleic acids research* 47.D1 (2019): D766-D773.

Andersson, Robin, et al. "An atlas of active enhancers across human cell types and tissues." *Nature* 507.7493 (2014): 455-461.

Lizio, Marina, et al. "Gateways to the FANTOM5 promoter level mammalian expression atlas." *Genome biology* 16.1 (2015): 1-14.

Bernstein, Bradley E., et al. "The NIH roadmap epigenomics mapping consortium." *Nature biotechnology* 28.10 (2010): 1045-1048.

Uhlen, Mathias, et al. "A pathology atlas of the human cancer transcriptome." *Science* 357.6352 (2017).

Andersson, Robin, and Albin Sandelin. "Determinants of enhancer and promoter activities of regulatory elements." *Nature Reviews Genetics* (2019): 1-17.

Stouffer, Samuel A., et al. "The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1." (1949).

Sacco, J. J., et al. "The deubiquitylase Ataxin-3 restricts PTEN transcription in lung cancer cells." *Oncogene* 33.33 (2014): 4265-4272.

Wang, Yu-Hui, et al. "CEBPD amplification and overexpression in urothelial carcinoma: a driver of tumor metastasis indicating adverse prognosis." Oncotarget 6.31 (2015): 31069

Seshacharyulu, Parthasarathy, et al. "FDPS cooperates with PTEN loss to promote prostate cancer progression through modulation of small GTPases/AKT axis." Oncogene 38.26 (2019): 5265-5280.

Fry, Elizabeth A., and Kazushi Inoue. "Aberrant expression of ETS1 and ETS2 proteins in cancer." Cancer reports and reviews 2.3 (2018).

Furlan, Alessandro, et al. "Ets-1 controls breast cancer cell balance between invasion and growth." International journal of cancer 135.10 (2014): 2317-2328.

Biography

Jingkang Zhao received his BS degree in Statistics from Peking University in 2013 and his MS degree in Biostatistics from Duke University in 2015. Then he was admitted into the PhD program in Computational Biology and Bioinformatics at Duke University. He met Professor Raluca Gordân in 2014, during his master study, and has been working in the Gordân Lab since then.

Publications:

1. Martin V*, Zhao J*, Afek A, Mielko Z, Gordan R (2019) QBiC-Pred: quantitative predictions of transcription factor binding changes due to sequence variants. Nucleic Acids Research 47(W1):W127-W135 (* co-first authors)
2. Shen N, Zhao J, Schipper J, Zhang Y, Bepler T, Leehr D, Bradley J, Horton J, Lapp H, Gordan R (2018) Divergence in DNA specificity among paralogous transcription factors contributes to their differential *in vivo* binding. Cell Systems 6(4):470-483
3. Zhao J*, Li D*, Seo J, Allen AS, Gordân R (2017) Quantifying the impact of non-coding variants on transcription factor-DNA binding. Research in Computational Molecular Biology 2017 (RECOMB17). Lecture Notes in Computer Science 10229:336-352. (* co-first authors)