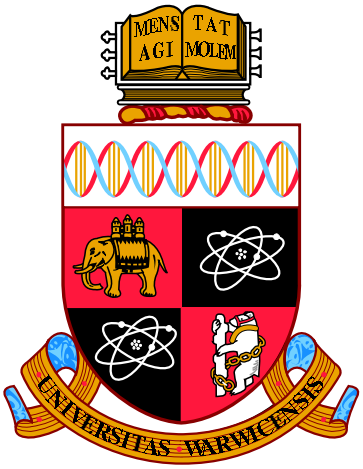

Predictive models of enhancer-gene regulation

by

Joseph Nasser

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy (by Published Work)

Warwick Medical School

December 2021

Contents

List of Figures

Acknowledgments

I would like to thank Sascha Ott for his guidance through the PhD process.
I would like to thank all those who have positively influenced me including my family, friends, colleagues and mentors.

Declarations

I hereby declare that this thesis has been composed by myself. I have not submitted any material in this thesis towards a degree at another university. This thesis is based on peer reviewed publications. See Appendix A for a list of publications included in this thesis and my contributions to them.

Joseph Nasser
December 2021

Abstract

Enhancers are DNA elements which play crucial roles in the spatio-temporal regulation of gene expression. An open question in enhancer biology is to identify which enhancers regulate which genes in which cell types. Recent experimental advances have enabled the design of high throughput screens to functionally interrogate putative enhancer-gene connections. These datasets have facilitated the development of predictive models of enhancer-gene regulation.

The main contribution of this thesis is an in depth analysis of the predictive ability of a specific enhancer-gene prediction method known as the Activity-by-Contact (ABC) model. We show that ABC is an effective predictor and outperforms other previously published prediction methods. We consider variations of the score which help to illustrate why the model performs so well. We consider the implications of success of the model on the role of genome architecture in gene regulation.

The ABC model is practically implementable and we describe how it was used to generate a database of enhancer-gene predictions across 131 cell types. We illustrate case studies which describe how this database can be used to interpret non-coding human genetic variation.

We discuss recent advances in single-cell sequencing which may form the basis for future larger-scale enhancer screens. We highlight a power-calculation that must be conducted to design such an experiment.

We also consider a formal mathematical representation of the ABC model. We show that the mathematical model is tractable and compute the mean of the mRNA distribution under this model. We show how the ABC Score formula can be derived from the mathematical model and generally describe the relationship between conceptual modeling and formal modeling.

We conclude by discussing the importance of these results within the field of gene regulation as a whole.

Chapter 1

Introduction

1.1 Discovery of enhancers

This thesis investigates the role of enhancers in eukaryotic gene regulation. Enhancers are non-coding DNA sequences which act as cis-regulators of gene expression. Enhancers were originally discovered in the early 1980s through experiments conducted by the laboratories of Schaffner and Chambon [1, 26]. In these experiments a non-coding sequence from the SV40 genome was shown to increase (enhance) the expression of genes on a plasmid. These early experiments also showed that the enhancer could activate expression regardless of its orientation and precise location relative to the gene promoter.

In the subsequent decades enhancers were discovered in their native genomic contexts (for the remainder of this thesis the term enhancer will refer to a DNA sequence which is a cis-regulator of a gene in its native context in the genome - not the ability of the sequence to drive expression in an episomal reporter assay). Enhancers were discovered in a wide range of model systems including drosophila, mouse, human etc. Such enhancers were found to play critical roles in development and to mediate disease risk [21, 7]. Through these studies it also became clear that some enhancers act in a cell-type specific manner - that is an enhancer sequence may control gene expression in one specific cell type or tissue - but may not be relevant in another cell type.

1.2 Enhancer mechanisms

The biochemical mechanisms by which enhancers function to activate genes has been an active area of study. Early experiments noted that enhancers tend to be accessible by the enzyme DNase suggesting that enhancer sequences are devoid of nucleosomes [7]. This has been confirmed with genome-wide assays of chromatin accessibility such as DNase-Seq and ATAC-Seq. ChIP experiments have begun to illuminate the set of factors present at enhancers [35]. It is now appreciated that transcription factors (TF) bind to enhancer regions and the role of many TFs have been elucidated. ChIP studies have also shown that the histones flanking active enhancers seem to be marked by a certain set of post-translational modifications [4]. These studies have suggested a guiding model of enhancer function - enhancers serve as landing pads for transcription factors which then recruit other activators and components of the transcriptional machinery.

A central open question is how the set of factors at the enhancer locus are able to affect the gene promoter which may be located distal (over tens to hundreds of kilobases or more) to the enhancer. Early experiments suggested that even though enhancers and their target gene promoters may be far away in the linear genome, they are actually in much closer proximity in the 3D nature of the nucleus [37]. With the advent of high throughput chromosome conformation mapping and live imaging of DNA the relationship between enhancer-promoter distances has been more fully investigated [34, 3]. Such studies have hinted at the role of 3D genome architecture in regulating enhancer-promoter communication although the precise relationship between genomic distance, 3D distance and gene regulation is still unclear.

1.3 Identifying enhancers in the genome

The progress made in understanding enhancer biology has formed the basis for the enhancer identification problem: given a gene in a particular cell type, or cell state, can we find all of its regulatory elements? The gold standard experiment to identify an enhancer is to perturb the DNA element through a genetic deletion and to read out the effect on expression of its putative target gene. However, such experiments are low throughput and prior to 2016 only a few dozen such examples have been found. Instead numerous attempts at computationally predicting enhancers have been proposed. Such attempts are typically based on identifying enhancers through epigenetic marks or through 3D genome conformation [10, 5, 38, 22]. Yet, the lack of gold standard functional experimental data has made it impossible to test these predictive models.

1.4 Relation of enhancers to human genetics studies

The ability to identify which enhancers regulate which genes in which cell types may impact our understanding of complex human traits and diseases. The 2000s have seen hundreds of genome-wide association studies (GWAS) conducted for a wide range of complex diseases and traits [6]. Such studies have implicated tens of thousands of variants (more precisely haplotypes) which are statistically associated with the traits/diseases. Crucially, however, such studies only identify the risk haplotype, they do not assign a mechanism by which the risk haplotype mediates disease risk.

The vast majority of risk variants are in non-coding regions of the genome. A seminal observation is that many of the risk variants appear to be in regions of the genome which are predicted to be enhancers [25]. A hypothesis has emerged wherein many variants mediate disease risk by altering enhancer function which in turn modulates gene expression. Accurate maps of enhancer-gene regulation are thus a potential stepping stone to investigate disease-associated non-coding genetic variation.

1.5 Technologies to experimentally identify enhancers at scale

The advances in this thesis are based on experimental work to test whether a DNA element acts as an enhancer for a given gene in a particular cellular context. Such an experiment typically consists of a strategy to perturb the enhancer element and read out its effect on gene expression.

The predominant perturbation strategy that has emerged is known as CRISPR interference (CRISPRi) [20]. This strategy consists of fusing the protein domain KRAB to catalytically inactive cas9 (dcas9). The dCas9-KRAB fusion is directed to a DNA element using a guide RNA complementary to the sequence of the DNA element. Upon localization to an enhancer element, KRAB recruits HP1 to heterochromatinize the chromatin and abrogate enhancer function. This strategy is favored over catalytically active cas9 strategies because it is amenable to high throughput experiment and it is unknown how the enhancer will be affected due to the short indels that occur due to homology directed repair. In general CRISPRi has been shown to be both robust and specific [19]. A limitation of CRISPRi is that it the dCas9-KRAB fusion may interfere with transcription making CRISPRi unsuitable for identifying enhancers-gene connections where the enhancer lies within the transcribed region of the gene.

CRISPRi has recently been combined with readout technologies to enable high throughput enhancer screens. In one of the first such screens [13] identified all regulatory sequences for the genes MYC and GATA1 in the K562 cell line. This initial functional datasets of enhancer-promoter connections facilitated the testing of predictive models of which enhancers regulate which genes. Based upon this

preliminary data, an enhancer prediction framework based on chromatin state and 3D genome architecture was hypothesized. This predictor was able to accurately distinguish DNA elements which acted as enhancers for MYC in K562 from those which did not. This predictor was later called the Activity by Contact Model

More experimental data was needed in order to rigorously test this predictive model. This led to the development of the CRISPRi-FlowFISH screening strategy (Fig 1) [14]. This strategy, in which the readout is based on RNA FISH and fluorescence activated cell sorting, identifies all regulatory elements for a given gene in a given cell type. This strategy is sensitive enough to detect approximately 20% effects on gene expression. Limitations of this strategy include that it cannot distinguish between cis and trans effects. This technique was used to screen all DNase peaks located within 500kb of 30 genes in the K562 cell line. In addition, we curated a set of other CRISPR experiments from the published literature. In total we have a dataset of 3863 experimentally interrogated element-gene pairs - of which 141 are found to be statistically significant. We will now use this data set to assess various predictive models of enhancer-gene regulation.

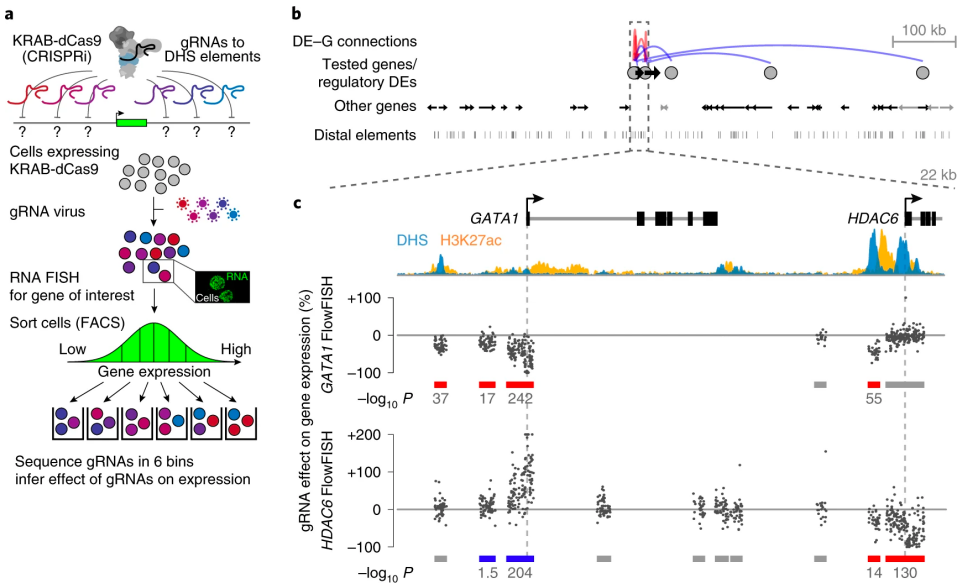

Figure 1: Figure reproduced from [[14], Main fig 1] **(a)** CRISPRi-FlowFISH Screening strategy. Cells expressing KRAB-dCas9 are infected with a pool of gRNAs targeting DHS elements near a gene of interest, labeled using RNA FISH against that gene and sorted into bins of fluorescence signal by FACS. The quantitative effect of each gRNA on the expression of the gene is determined by sequencing the gRNAs within each bin. Inset: example of K562 cells labeled for RPL13A. **(b)** Distal elements (DE) affecting GATA1 and HDAC6 expression in K562 cells. Genes expressed in K562 cells are shown in black; those not expressed are shown in gray. Red/blue arcs: perturbation of a DE resulted in a significant decrease/increase in the expression of the tested gene. Gray circles are distal elements where perturbation with CRISPRi affects the expression of at least one tested gene as measured by CRISPRi-FlowFISH. Distal elements are DHS peaks. **(c)** Close-up of region containing GATA1 and HDAC6. Points represent the effect on gene expression of a single gRNA. HDAC6 vertical axis capped at 200%. Gray, red and blue bars: DHS elements in which CRISPRi leads to either no detectable change (gray) or a significant decrease (red) or increase (blue) in expression. Elements overlapping the assayed gene are excluded from analyses because recruitment of KRAB-dCas9 in a gene body directly interferes with transcription. Such elements are included in analyses for other genes, as shown for the elements overlapping GATA1

Chapter 2

The predictive ability of the ABC Score

2.1 Precision-recall framework

Our first task is to compare the performance of various models in predicting the experimental data. In order to do so we require a quantitative framework for making such a comparison. Assuming that the prediction method is quantitative (as opposed to binary), we can make a scatter plot whose x axis is the predictor and y axis is the experiment. An example is given in Figure 2 using genomic distance as a continuous predictor. In Figure 2 it's clear that red points are shifted to the left relative to gray, indicating that a predictor solely based on enhancer-gene distance has better than random performance. In order to formalize and quantify this relationship, as well as support binary predictors, we use a precision-recall (PR) framework.

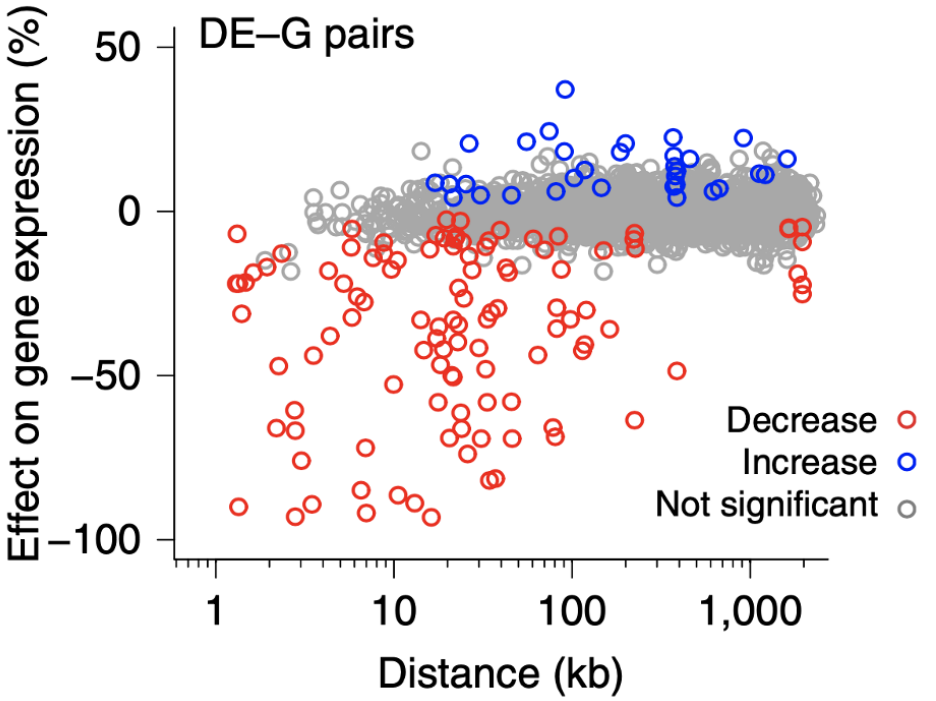

Figure 2: Figure reproduced from [14] Main fig 2e. Relationship between element-gene distance and effect size observed in CRISPRi Flow-FISH screen. Each dot represents an element-gene pair. The x-axis is the genomic distance between the element and the transcription start site of the gene. The y-axis represents the change in gene expression upon CRISPRi perturbation of the element as measured by FlowFISH. Red and blue dots represent statistically significant pairs, gray dots are not statistically significant.

For the PR framework, the set of ground truth positives are taken to be the element-gene pairs in the CRISPR dataset which are statistically significant and for

which CRISPR perturbation of the element leads to a decrease in gene expression. We then classify each element-gene pair in the CRISPR dataset as a true positive (TP), false positive (FP), true negative (TN), false negative (FN) relative to each predictive model. The precision of a predictive model is TP/(TP + FP) and the recall is TP/(TP + FN). Ie, precision measures whether the predicted positives are experimental positives and recall measures how many of the experimental positives are predicted positives. For a continuous predictor the PR curve is formed by computing precision and recall using a binary cutoff throughout the predictors range. A perfect predictor has both a recall and precision of 1 and would be represented as a dot on the top right of the PR curve. A predictor that makes random predictions would be represented by a horizontal line whose y-intercept is equal to the percentage of positives in the dataset. We note an important limitation of the PR framework is that the quantitative effect size of the experimental data is ignored - only its statistical significance is considered.

With the PR framework in hand, we can evaluate the performance of different predictors (Fig 3a). We find that predictors based on proximity such as distance or closest gene have modest performance, predictors based on binary features of the 3D genome (such as TADs and Loops, see Chapter 3) or correlation techniques have poor performance. We find that the ABC Score (described in detail below) has the best performance.

2.2 ABC Score - conceptual explanation

We now describe the conceptual notion of the ABC model (Fig 3b), details of its computation are in the next section. The ABC model is based on the following conceptual notions:

- Enhancers activate their target genes upon contact (or close physical proximity) between the enhancer element and the target gene promoter
- Each enhancer element is assumed to have an intrinsic activity level. Activity may reflect the frequency with which the TFs and co-activators are recruited to the enhancer
- Each enhancer has its own level of activity which is the same for each gene.
- The expression of a gene is proportional to sum of enhancer activity weighted by its contact frequency to the gene promoter

Putting this together we have the linear formula

$$G \sim \sum_{e} A_e C_{e,G} \tag{2.1}$$

Where G is the expression of a gene, A_e is the activity of an enhancer element, $C_{e,G}$ is the contact frequency between the enhancer and the target gene promoter and the sum is over all regulatory elements.

We now use this linear formula to derive the ABC Score. In the CRISPR experiments we measure the fraction change in gene expression

$$\text{Fraction Change} := \frac{G - G^{\Delta}}{G} \tag{2.2}$$

where G is the expression of a gene in the control condition and G^{Δ} is the expression of the gene in the CRISPR perturbation condition. Making the assumption that perturbation of element e causes the Activity of e to go to zero, combining 2.1 and 2.2 we have the

$$\text{ABC Score(G,e)} := \frac{A_e C_{e,G}}{\sum_{e'} A_{e'} C_{e',G}} \tag{2.3}$$

where the sum is taken over all enhancers e' within 5Mb of the transcription start site of G. See chapter 6 for a more formal mathematical derivation of the ABC Score.

2.3 ABC Score - practical implementation

In order to make the ABC Score practically implementable, we need to assign values to Activity and Contact. The Activity is intended to represent the ability of the element to bind transcription factors and other components of the transcriptional machinery. We estimate this quantity by the amount chromatin accessibility of the element and the amount of histone acetylation at the element. The Contact of an element-gene pair is taken to be the Hi-C contact frequency between the element and the gene promoter. We now describe this process in more detail.

The first step in computing the ABC Score is defining the set of putative enhancer elements (elements). Elements are taken to be regions of DNA which display elevated levels of chromatin accessibility. These are determined by calling peaks in an ATAC-Seq or DNase-Seq experiment. We make the heuristic assumption that each element consists of roughly one nucleosomes worth of DNA (150bp) and

that activating histone marks are present on neighboring nucleosomes. As such the typical element is approximately 500bp in length. The Activity of the element is computed as the geometric mean of the number of read counts (normalized by read depth) of Dnase-seq (or ATAC-Seq) and H3K27ac ChIP-Seq over the element (Fig 3b)

The Contact of an element-gene pair is derived from Hi-C experiments. We begin with a KR normalized Hi-C matrix at 5kb resolution. This matrix is then minimally processed in the following ways:

- We replace each diagonal element of the matrix with the maximum of its neighboring bins. We do this because we noted that the Hi-C signal on the diagonal is not correlated with either of its neighboring bins. This suggests that the Hi-C signal on the diagonal bins is highly influenced by the ability of the sequence to self-circularize and is not reflective of DNA contacts occuring within 5kb.

- We add a small pseudocount to each bin. For bins less than 1Mb the pseudocount is equal to the expected contact frequency at 1Mb (based on the powerlaw distribution), for bins at distance greater than 1Mb the pseudocount is equal to the expected contact frequency at that distance. This pseudocount is a crude regularizer which is designed to hedge against counting noise in the Hi-C experiment and to ensure that each bin has a non-zero count.

The Contact of an element gene pair is then the entry in this modified matrix corresponding to the midpoint of the element and the transcription start site of the gene.

We show below that the performance of the ABC score is robust to many data processing decisions that go into computing the ABC Score (Fig 6). We emphasize that the ABC Score is a parameter-free predictor - there are no free parameters: each component of ABC is derived from epigenomic/Hi-C experiments orthogonal to the CRISPR dataset. There is no model fitting, or training/testing framework involved.

2.4 ABC Score performance

The ABC Score has high predictive ability on the CRISPR dataset. Considered as a binary predictor, the ABC Score achieves a precision of 59% at a recall of 70%. We use the area under the PR curve (AUPRC) as a summary metric of effectiveness of

a predictor. The ABC Score achieves an AUPRC of .65. As a continuous predictor, the ABC Score is correlated with the effect size observed in the CRISPR experiments (Spearman correlation of -.63). This is a crucial point! The ABC Score is not a statistical model, it is inherently a (crude) biophysical model. The fact that the score is correlated with effect size is a strong signal that the model has some biological relevance.

2.5 Performance of variations of ABC Score

We now dig deeper into ABC performance. We first note that the ABC Score performs better than each of its individual components (Fig 4a). This appears to be because there are some strong distal enhancers and some weak proximal enhancers (Fig 4b).

Next we investigate different ways to measure Activity. We compute the performance of the ABC Score using different histone marks, or measures of transcription at the enhancer in place of H3K27ac. In general we find that many of these variations of the ABC score have reasonably good performance which is roughly comparable to H3K27ac, but some marks do not (Fig 4c).

Next we investigate a variation of the ABC Score which uses Hi-ChIP data (Fig 4d). Hi-ChIP is an assay which performs chromatin immunoprecipitation on a Hi-C library. As such a H3K27ac Hi-ChIP dataset can be roughly viewed as an experimental convolution of Hi-C and H3K27ac ChIP-Seq. We computed a variation of the ABC Score using H3K27ac Hi-ChIP data in K562 cells from [27] . We find the performance of this variation is similar to using separate Hi-C and H3K27ac ChIP-Seq datasets. Importantly, the predictor using quantitative Hi-ChIP signal far outperformed the Hi-ChIP loop calls (see next chapter for discussion of loops). This result gives us further confidence that the ABC concept is robust to the exact means by which to experimentally access Activity and Contact.

In analyzing the CRISPRi-FlowFISH dataset we noticed that ABC Score performance seemed to vary from gene to gene. In particular there was a subset of genes for which ABC tended to make many false positive predictions. We investigated whether these genes had a common property and realized that many of them are classified as ubiquitously expressed genes (sometimes denoted housekeeping genes.) We denoted genes as 'ubiquitously expressed' based on previous annotations which enumerated genes that had detectable or uniform expression across many tissues [39, 11, 8]. We find that the ubiquitously expressed genes tested in the

CRISPRi-FlowFISH assay tend not to have any enhancers, even though there are elements with high ABC Scores (Fig 5).

We also examined how the performance of ABC varies with respect to various decisions that need to be made in terms of data processing (Fig 6). We recomputed the ABC score by varying the range over which the sum in the ABC Score is computed (default 5mb), the size of the candidate elements (default 500 bp), the distance used in the Hi-C pseudocount regularization (default 1Mb) and the replacement of the Hi-C diagonal bin (default equal to the maximum of neighboring bins). We find that ABC performance is robust to all of these choices.

2.6 ABC Performance in other cell types

We also compared ABC predictions against CRISPR data from other cell types besides K562. This includes

- element-gene connections tested using allele-specific deletions alongside allele-specific RNA-Seq in a hybrid mouse stem cell line [9]
- element-gene connections tested using CRISPRi through the CARGO system followed by RNA-Seq in NCCIT cells [12]

In Fulco 2019 we found ABC performance in the non-K562 data to have an AUPRC of .73 (Fig 7). In [28] we also applied CRISRPi-FlowFISH to cell lines derived from immune lineages and found ABC performance in these cell types to be similar to K562. Overall we find that the success of the ABC model is not limited to K562 and performance is similar across other cell types and experimental measurement modalities.

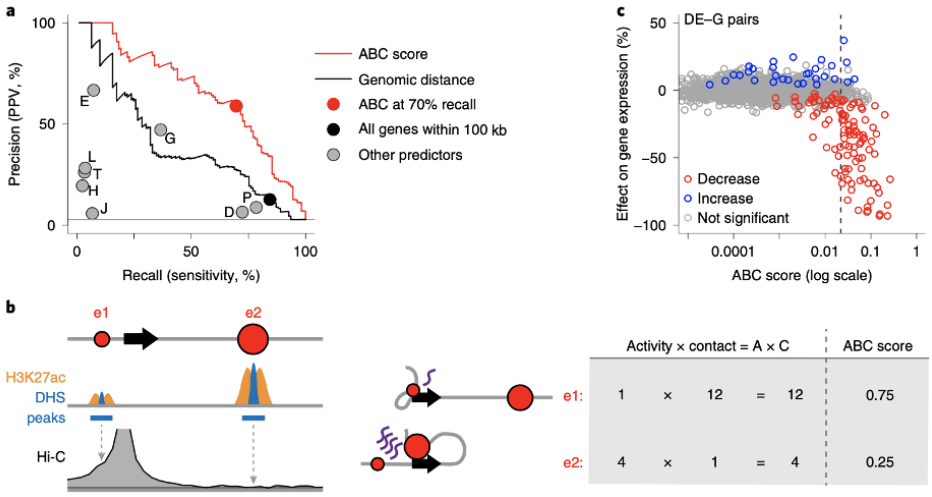

Figure 3: Figure reproduced from [14], Main fig 3] **(a)** Precision-recall plot of various predictive models evaluated on CRISPR dataset. Dots E (each element is predicted to only regulate the closest gene), and G (each gene is predicted to be regulated by its closest DHS element) represent proximity based predictors. Dots L (element-gene pairs are predicted to be regulatory if they are at opposite ends of Hi-C loops) and D (element-gene pairs are predicted to be regulatory if they are contained within the same topologically associated domain) are based on binary features of Hi-C maps. Dot J is a correlation based predictor which makes enhancer-gene predictions by correlating epigenetic signals in DNA elements with gene expression across many cell types [5] 'T' represents enhancer-gene connection predicted by the TargetFinder algorithm which combines loop calls with epigenetic data. [38] 'P' represents enhancer-gene connection predicted by RNA polymerase II ChIA-PET loops [22] 'H' represents enhancer-gene connection predicted by H3K27ac Hi-ChIP loops [27] **(b)** Hypothetical calculation of the ABC Score. e1 and e2 represent two arbitrary enhancers for the gene (black arrow). The Activity of the elements is determined by quantitative DHS and H3K27ac signals. The Contact between the elements and gene promoter is determined by Hi-C. **(c)** Comparison of ABC scores (predicted effect) to observed changes in gene expression following perturbations. Each dot represents one tested element-gene pair. Red/blue dots: connections for which perturbation resulted in a significant decrease/increase in the expression of the tested gene. Gray dots: no significant effect. Dotted black line denotes 70% recall, corresponding to the big red dot in **(a)**

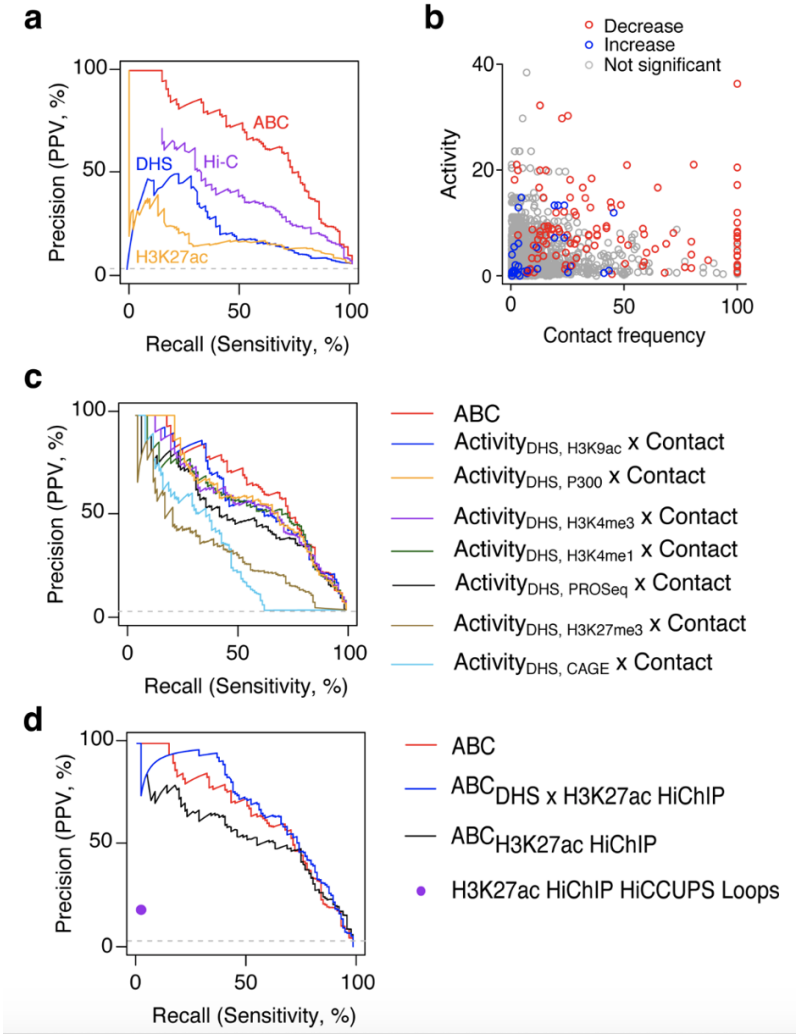

Figure 4: Figure reproduced from [14], Extended Data Fig 3 **(a)** Precision-recall curves for classifying regulatory element-gene pairs, comparing each of the components of the ABC score **(b)** Scatterplot of Activity and Contact frequency for each tested element-gene. Contact frequency is derived from KR normalized Hi-C experiments and linearly scaled to a maximum of 100 **(c)** Performance of ABC Score when H3K27ac is replaced by other epigenomic datasets. $\text{Activity}_{\text{Feature1,Feature2}} = \text{sqrt(Feature1 RPM x Feature2 RPM)}$. (ABC score corresponds to $\text{Activity}_{\text{DHS, H3K27ac x Contact}}$) **(d)** Precision-recall curves for the ABC model using H3K27ac HiChIP [27]. $\text{ABC}_{\text{DHS x H3K27ac Hi-ChIP}}$ corresponds to a predictive model whose score is proportional to the Dnase signal at the candidate element multiplied by the quantitative H3K27ac Hi-ChIP signal between the element and gene promoter. $\text{ABC}_{\text{H3K27ac Hi-ChIP}}$ is the same as above but only uses the existence of the DHS peak as opposed to the quantitative signal in the DHS peak. H3K27ac HiChIP HiCCUPS Loops is the HiCCUPS loop calls derived from the H3K27ac HiChIP experiment. ABC (red line) corresponds to the standard ABC Score using Dnase, H3K27ac and Hi-C.

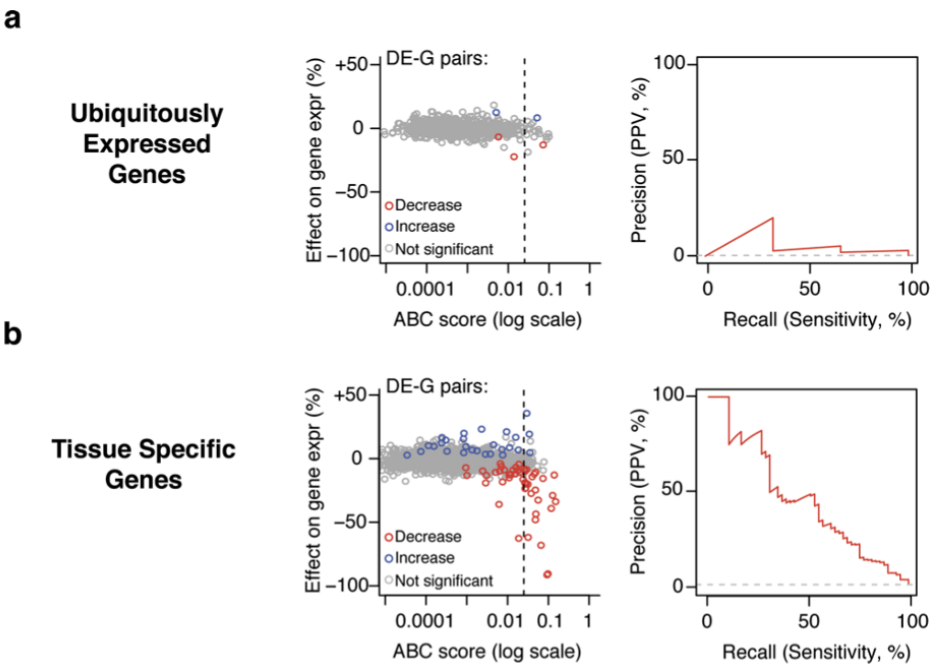

Figure 5: Figure reproduced from [14] Extended Data Fig 4 **(a)** Left: Comparison of ABC scores (predicted effect) with observed changes in gene expression upon CRISPR perturbations. Each dot represents one tested DE-G pair where G is a ubiquitously expressed gene. Right: precision-recall curve for ABC score in classifying enhancers for ubiquitously expressed genes **(b)** Same as (a) for tissue-specific genes. All panels include only the subset of our dataset for which we have CRISPRi tiling data to comprehensively identify all enhancers that regulate each gene (30 genes from this study [14], 2 from previous studies [13];

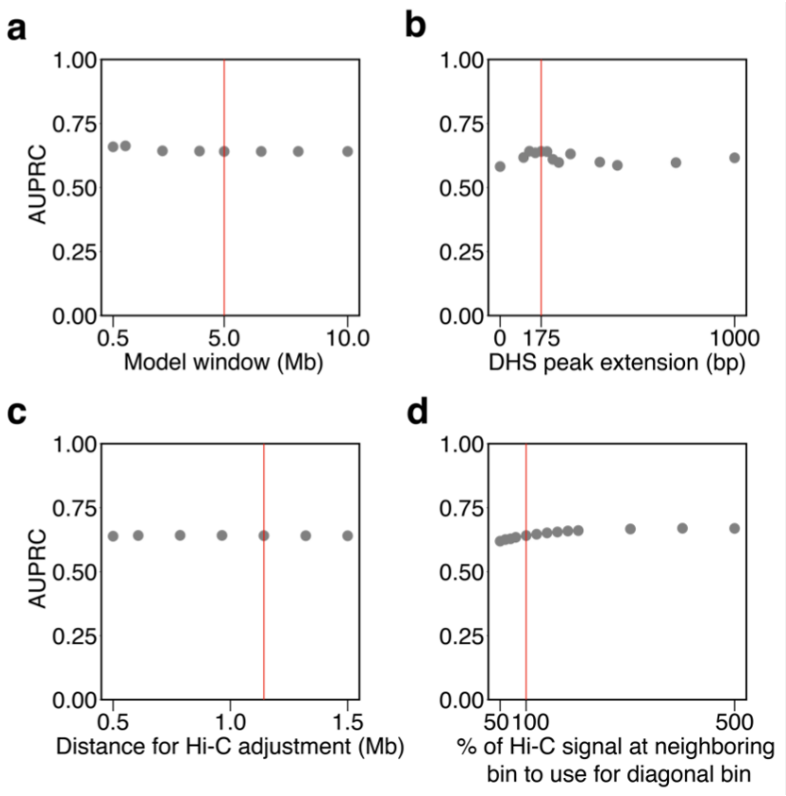

Figure 6: Figure reproduced from [14], Supplement Fig 5. Changing the data processing parameters parameters of the ABC score does not dramatically affect performance near the default values. Each panel presents the area under the precision recall curve (AUPRC) for the ABC score when changing the specified parameter. Red lines indicate the values used throughout this paper. **(a)** Genomic distance within which elements are included in the model. **(b)** Number of bases DHS peaks were extended on either side before merging to create candidate elements. **(c)** Genomic distance used to compute the pseudocount added to the Contact component. **(d)** In processing Hi-C data, each diagonal entry of the Hi-C matrix is replaced by some percentage of the maximum of its four neighboring entries

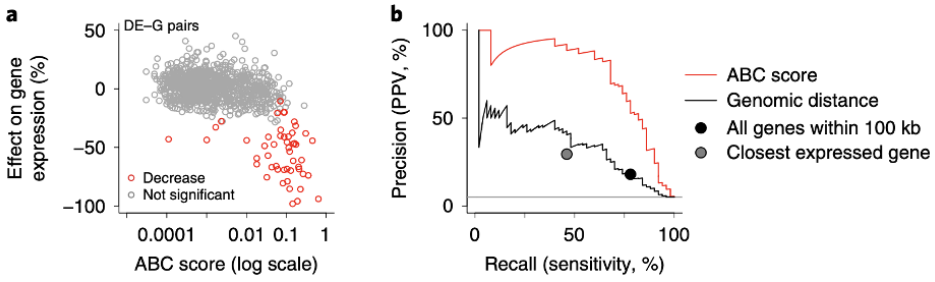

Figure 7: Figure reproduced from [14], Main 4. ABC Performance on CRISPR data in cell types other than K562

Chapter 3

The ABC Model and genome architecture

In this chapter we revisit various models for the role of genome architecture in gene regulation in light of the results from chapter 2.

3.1 Introduction

The development of the Hi-C assay has revealed many features of the 3D genome such as TADs and loops [30]. These features have been hypothesized to play important roles in gene regulation and have been investigated in some studies at individual loci [36, 15]. Yet, the precise relationship between 3D genome architecture and gene regulation is still unclear, and general principles that apply across the genome have yet to be elucidated.

We begin by formalizing two common notions of the role of 3D architecture into enhancer prediction models. We then show that these models are incompatible with the results described in chapter 2. We then describe alternative models suggested by the results from chapter 2.

3.2 TAD and loop hypotheses

We begin by stating what we call the TAD and loop enhancer-prediction models:

- The TAD predictor: this model proposes that enhancer-gene communication is non-specific within TADs and constrained to within TADs.
- The Loop predictor: this model proposes that enhancer-gene connections can be identified as loops or peaks in a Hi-C experiment

In order to test the TAD and loop predictors, we need to convert these conceptual models into concrete predictive models. We note that the CRISPRi-FlowFISH experiments were not specifically designed to test such hypotheses (an ideal such experiment would directly perturb genome architecture), however, we can use the CRISPR dataset to test whether these models are consistent with the experimental data.

We define the TAD predictor as classifying all element-gene pairs that are within a TAD as positives and all element-gene pairs not contained in the same TAD as negatives (ie, all DHS peaks regulate all genes within its own TAD, but no genes outside of its TAD). We observe that this predictor has high recall, but low precision: most regulatory enhancer-gene pairs are contained within a TAD, but the vast majority of putative enhancers (Dnase-peaks) do not regulate a gene within its

TAD (Fig 3a). We also tested whether the TAD hypothesis performed better by incorporating enhancer activity: we tested a model in which the Contact component of the ABC Score is equal to 1 if the element and gene promoter are in the same TAD and 0 otherwise. We find that this model performs substantially worse than using Hi-C data (Fig 8a).

We define the loop predictor as classifying all element-gene pairs that are at opposite ends of Hi-C loops (as called by the HiCCUPS algorithm on K562 Hi-C data) as positives and all other element-gene pairs as negatives. We observe that this predictor has modest precision and low recall: the vast majority of regulatory enhancer-gene pairs are not identified with loops, and more than half of all element-gene pairs associated with loops are not regulatory (Fig 3a). We note that this analysis is contingent on the HiCCUPS loop caller. It is possible that the conceptual notion of identifying enhancer-gene connections as focal amplifications in Hi-C experiments is valid, but a different loop calling method tuned more precisely for this task would be needed. In particular, the median distance between anchors at ends of HiCCUPS loops in K562 is 260kb, whereas the median distance between regulatory enhancer-gene pairs in the CRISPRi-FlowFISH dataset is 22kb. It is also possible that chromatin confirmation capture methods with higher resolution (such as micro-C or promoter-capture Hi-C) may provide better experimental technologies to identify enhancer-gene connections.

3.3 Using a powerlaw relationship as opposed to Hi-C data

Hi-C experiments have demonstrated that contact frequency is well approximated by a power-law function of genomic distance [32]. We fit a power-law to K562 Hi-C data and found the power-law relationship with exponent .7 explained about 70% percent of the variance in Hi-C data. We thus tested a version of ABC in which the Contact component of the ABC Score is derived from the power-law relationship as opposed to the Hi-C data itself. We find that this predictor has similar performance to the ABC Score based on cell-type specific Hi-C data (Fig 8a,b). How is it possible to accurately make cell-type specific predictions without using cell-type specific Hi-C data? We recall that the median distance between regulatory pairs in our dataset is 22kb, whereas many of the cell-type specific features of the 3D genome that are not captured in a powerlaw (loops, TADS, stripes) appear at much

larger distances.

To investigate this further we computed the power-law based version of ABC with a wide range of powerlaw exponents. We find that ABC performance is high for exponents that match Hi-C data, but is lower for more extreme exponents (Fig 8c). We also find that a version of ABC in which the Contact component decays linearly performs poorly (Fig 8a,b). We further note that even within TADs there is a decay rate with distance (Fig 8d). The ABC Model thus makes the prediction that distances within TADs are important, which was subsequently observed in a recent experiment [40]. A similar relationship where contact frequency decays with distance holds for loops (Fig 8f). For example the strongest loops at a distance of 500kb have the same contact frequency as typical non-loop loci at a distance of 50kb.

We thus come to our main conclusion in this section: the salient feature of the 3D genome that appears to matter for enhancer-gene prediction is contact frequency, not binary features of the 3D genome.

3.4 Construction of an average Hi-C dataset

The performance of the power-law version of ABC suggests that making cell-type specific enhancer-gene predictions may not require cell-type specific Hi-C data. However, we know there are certain features of Hi-C data, which in some cases may be important for enhancer-gene prediction, that are not apparent in the powerlaw distribution. Interestingly, many of the TADs and loops which have been identified in Hi-C experiments are actually not cell type specific. For example, [30] found that 55-75% of loops are shared between distinct cell types. As such we suggest that by averaging together many different Hi-C maps, we can form a reference Hi-C map which is more informative than the power-law and can serve as a reference for making cell-type specific enhancer-gene predictions in the absence of cell-type specific Hi-C data.

We construct such a map by averaging together Hi-C data from 10 cell lines. We find that such a map explains 80% of the variance in Hi-C data in a particular cell line. We use this average Hi-C map to predict enhancer-gene connections in a variety of cell lines and find a high level of performance when comparing to CRISPR data (Fig 8a).

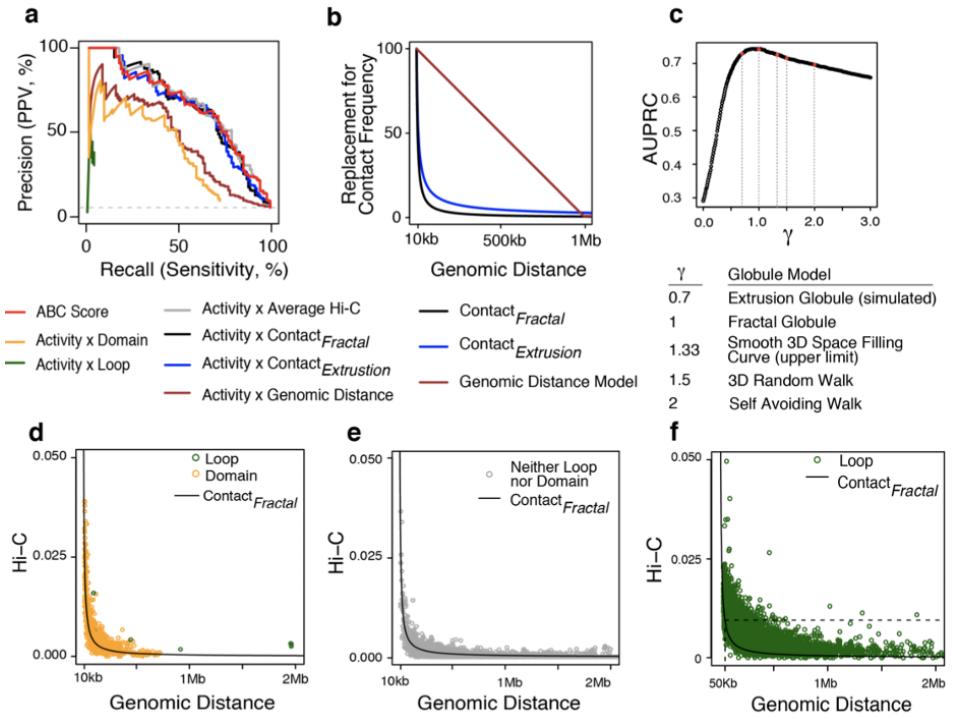

Figure 8: Figure reproduced from [14], Supplementary figure 6] **(a)** PR curves for various versions of the ABC model in which the contact component (derived from K562 Hi-C data) is replaced by binary features of Hi-C experiments (loops or domains) or decreasing functions of distance. Yellow: Contact component is equal to 1 if the element and gene promoter are in the same TAD (based on K562 Hi-C data) and 0 otherwise. Green: Contact component is equal to 1 if the element and gene promoter are at opposite ends of the same HiCCUPS loop (based on K562 Hi-C data) and 0 otherwise. Black/blue/brown: Contact component is replaced with decreasing functions of linear distance. Brown represents a linear decay. Blue represents a powerlaw decay with exponent .7 and black represents a powerlaw decay with exponent 1. Gray: Contact component is replaced with Hi-C data computed as the average of Hi-C datasets from 10 different cell types. **(b)** Visualization of functions used to replace contact frequency in (a) **(c)** AUPRC for models in which the Contact component of the ABC Score is replaced by a powerlaw function with varying exponent. Values of exponent derived from various polymer models of DNA are highlighted in red. **(d,e)** Scatterplot of genomic distance vs Contact frequency (from K562 Hi-C) for all experimental tested element-genea pairs in K562 at distance >10kb. Colors represent membership in the same TAD (yellow) loop (green) or neither (gray) **(f)** Scatterplot of genomic distance vs contact frequency for all HiCCUPS loops in K562. Although contact frequency is greater than expected under the powerlaw model (with exponent .7) the absolute increase is modest. For example, the loops with the highest contact frequency separated by 500kb have the same contact frequency as non-loop loci at 50kb

Chapter 4

A database of ABC predictions with applications to human genetics

In this chapter we describe how we built a database of ABC Predictions across 131 cell types. We then describe how this database can be used in conjunction with human genetic studies to investigate the genetic basis of complex traits and diseases.

4.1 Building a database of ABC predictions

Given our ability to make cell type specific predictions using the average Hi-C dataset, the data required to generate ABC predictions in a particular cell type are just ATAC-Seq and H3K27ac ChIP-Seq. Accordingly, we gathered these two epigenomic datasets for a total of 131 cell types and states (hereafter termed biosamples). These biosamples included a variety of immortalized cell lines and primary tissue from the ENCODE and Roadmap consortia, resting and stimulated immune cell types from the Engreitz lab, and a variety of other samples from the literature. We then generated ABC predictions in each of these biosamples.

One particular challenge in building this database of ABC predictions is that the ABC Score is quite sensitive to the signal to noise ratio of the epigenetic data. As the signal to noise ratio of the epigenetic data decreases, ABC Scores also tend to become more uniform. As such ABC Scores computed in two different biosamples, where the epigenetic data in each biosample has different signal to noise ratios, are not directly comparable (although the relative ranking of each element-gene pair within each biosample is still valid). In order to mitigate this, we quantile normalized the counts of ATAC-Seq and H3K27ac ChIP-Seq across the candidate elements in each cell type to a reference (chosen to be the counts in K562 cells). This procedure allows ABC scores to be comparable across cell types and facilitates the creation of a reference database. In our initial testing we also noted the sensitivity of the ABC Score to various properties of the Hi-C datasets. However, the use of the Average Hi-C dataset for the entire database sidesteps these issues.

We note some of the basic properties of our predictions. Each expressed gene is predicted to be regulated by 2.8 enhancers while each enhancer is predicted to regulate 2.7 genes. We also note the cell type specificity of our predictions - on average only 19% of connections are shared between cell types, with cell types from similar tissue types or cell lineages having more shared connections (Fig 9a,b).

We employed various checks to evaluate the self consistency of the predictions. For biosamples in which independent replicates of the epigenomic data were

available, we computed ABC Scores separately in each replicate. We found ABC Scores to be quantitatively well correlated (Fig 9c). We also compared the number of shared connections between the replicates. We found that on average 90% of connections are shared between replicates, with the proportion of shared connections increasing as the ABC Score cutoff increased (Fig 9d) or the correlation of the underlying epigenetic data increased (Fig 9e).

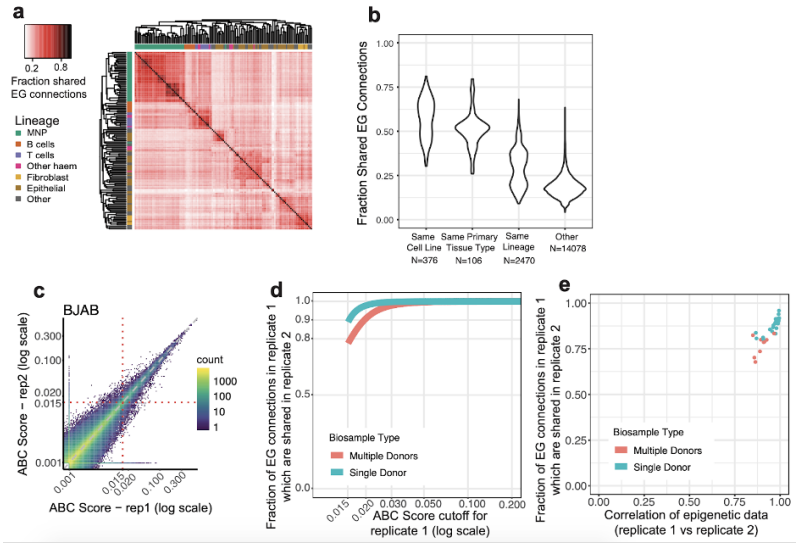

Figure 9: Figure reproduced from [[28], Supplementary figure 6] **(a)** Heatmap of fraction of shared connections among 131 biosamples. For each pair of biosamples we computed the fraction of predicted EG connections shared between the biosamples. Median of row medians is 19% **(b)** Distribution of shared predicted connections stratified by relatedness of the two biosamples **(c)** Quantitative reproducibility of ABC Scores - example from the BJAB cell line. Each axis represents ABC Scores computed using independent epigenomic experiments **(d)** Fraction of shared connections increases as ABC Score cutoff increases. Single Donor refers to biosamples for which the epigenomic data is derived from a single donor, Multiple Donors refers to biosamples for which the epigenomic data is derived from multiple donors **(e)** Fraction of shared connections increases as correlation of underlying epigenomic data increases

4.2 Applications to human genetics

We next evaluated whether this database could be used to interpret disease-associated non-coding genetic variation. The central idea is that a disease-associated variant contained within a predicted ABC enhancer yields a hypothesis by which the vari-

ant mediates disease risk - the variant is hypothesized to influence enhancer activity which in turn affects gene expression in a particular cell type or state. The approach is then to check if any fine-mapped disease associated variants lie within a predicted ABC enhancer within the entire database of predictions.

We initially evaluated this approach in the case of inflammatory bowel disease (IBD). IBD was chosen because there existed accurate fine-mapping and many of the cell types known to be relevant for IBD were contained in our ABC database. We found that ABC enhancers in IBD relevant cell types were enriched for IBD associated variants [28]. Additionally, we found that the ABC Score successfully connected these variants to genes known to be relevant to IBD (a gold standard set of genes was curated based on coding variation, therapeutic targets etc). Finally, the ABC model predicted that the target of a specific variant, rs1250566 was the gene PPIF instead of ZMIZ1 (which was previously hypothesized to be the right gene since the variant resides in an intron of ZMIZ1). This prediction was verified using CRISPRi experiments and the impact of PPIF on cellular function was studied [28]

We also applied this approach to a recent study of clonal hematopoiesis. We found that a fine-mapped non-coding variant was predicted to regulate a gene TET2 [2]. A CRISPR deletion experiment validated this connection which yielded a regulatory mechanism hypothesis for the means by which this variant mediated disease risk.

We do note some limitations to the approach of using the ABC prediction database to interpret non-coding genetic variation.

- ABC makes predictions at the enhancer level, not the variant level.
- A predicted ABC connection may actually be real (ie, the enhancer does regulate the gene) but it may not be relevant for disease.
- The ABC Score is only designed to predict the effect of enhancers, it does not predict other non-coding mechanisms such as splicing, UTR, CTCF etc.
- Success of the approach is fundamentally limited by the size and content of the ABC database.
- The ABC Model itself is not perfect

Despite these limitations, the success of the approach for IBD and CHIP suggests it could be more widely applicable.

Chapter 5

Single-cell screen power calculations

This is a brief chapter describing my contributions to [23]. This study describes a new technology, Hybridization of Probes to RNA for sequencing (HyPR-seq), to measure the expression level of a select set of RNAs in single cells. This technology could be used in combination with CRISPRi to design enhancer screens that test thousands of putative enhancer-gene connections in a single experiment. Such sample sizes will be necessary in order to further refine the ABC Model and other models of gene regulation.

My contributions to this study were to conduct a power calculation to estimate the statistical power to detect certain effect sizes based on many parameters of the screen. These parameters include the effect size of the enhancer on gene expression, the expression of the gene, the number of cells needed to be profiled, the number of guide RNAs per enhancer element etc. Simulations were conducted to assess statistical power in a hypothetical HyPR screen. These simulations were compared against the observed power in an actual single-cell CRISPR screen [16] which used full transcriptome 10x sequencing (as opposed to target sequencing with HyPR).

HyPR technology allows the sequencing reads to be concentrated in the genes of interest which greatly increases the statistical power of the experiment. Considering a hypothetical enhancer screen consisting of 1,000 gRNAs tested against 50 genes, we find that approximately 25,000 single cells would need to be profiled with HyPR (assuming 5,000 reads per cell) to have 90% power to detect 25% effects on gene expression. To achieve the same power with 10x whole transcriptome sequencing would require over 1,000,000 cells at 20,000 reads per cell. This analysis does not model guide to guide variability between guides targeting the same enhancer element.

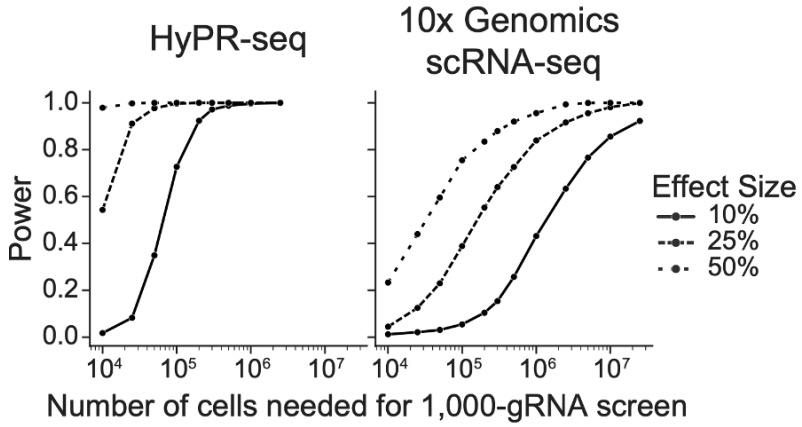

Figure 10: Figure reproduced from [[23], Main figure 2d] Power to detect changes in gene expression of various magnitudes as a function of the total number of cells profiled. Power is shown for a hypothetical screen of 1,000 CRISPR perturbations and is averaged over 37 genes expressed at $>$ 1.5 UMIs per cell in a HyPR-seq dataset (Left) and the same genes in a 10X Genomics Chromium 3 scRNA-seq dataset (Right) sequenced to 18,000 UMIs per cell [16]

Chapter 6

Mathematical formalization of ABC Model

The ABC model as described in Chapter 2 is considered informal in the sense that the model is presented as a cartoon. The ABC Score is derived from the cartoon using various assumptions and heuristics. In this chapter we show how the informal representation of the ABC model can be translated into a formal mathematical model of transcription. We show how the ABC Score can be formally derived from the mathematical model.

6.1 The linear framework

The mathematical modeling framework we employ is known as the linear framework [17]. The linear framework is a graph based approach to Markov processes. The starting point of the linear framework consists of a finite, directed, labeled graph, G. In the context of gene regulation, each vertex of G represents a regulatory state of a gene and the edge labels represent transition rates between the regulatory states. The ABC Score is fundamentally a statement about mRNA counts, as such we need a framework that models the number of mRNA present in the cell as well as the regulatory state of the gene. To do so, we define the copy-number graph $C(G)$ of any regulatory graph. $C(G)$ is an infinite graph which chains together countably many copies of G by the production and degradation of mRNA. An example conversion of a finite regulatory graph to an infinite copy number graph is given in Figure 11.

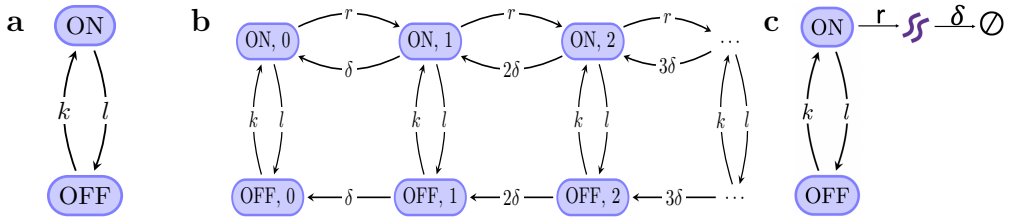

Figure 11: Conversion of finite underlying graph to infinite copy-number graph. **(a)** Underlying regulatory graph of a two state model. **(b)** Infinite copy number graph of two state model in (a) in which ON is the only production state with production rate r and degradation rate δ. The state (ON, s) indicates that the gene is in the ON state and there are s copies of mRNA in the cell. **(c)** Compact representation of infinite copy number graph in (b). The purple squiggles represent mRNA molecules. The arrow from ON with label r indicates that ON is a production state.

6.2 The multi-ON model

Given a gene with m enhancers, the simplest possible linear framework representation of the ABC model consists of graph with $m + 1$ states (Fig 12c). This model includes one state in which mRNA is not produced and m production states. We call this model the 'multi-ON model'. The multi-ON model does not explicitly specify anything about the biology of gene regulation - each ON state in the multi-on model is an abstract state. In this case we consider the state ON_i to refer to a state in which the gene is being transcribed due to the effects of enhancer i. In the context of the ABC model, this means that ON_i represents a state in which enhancer i is active and is contacting the gene promoter.

However, even with this idea in mind, it is unclear how the notions of Activity and Contact should be incorporated into the transitions rates of the multi-ON model. One possibility is that the notion of Contact should be incorporated into the rates k_i and l_i and Activity incorporated into the production rates r_i. Another possibility is that both Activity and Contact should be incorporated into the transitions rates k_i and l_i and that the production rates r_i are all equal and reflect a property of the gene and not the enhancers. Although the precise mapping between the conceptual ABC Model and the multi-ON model is not completely clear, we do show that, under certain interpretations, the ABC Score can be formally derived from a mathematical analysis of the multi-ON model.

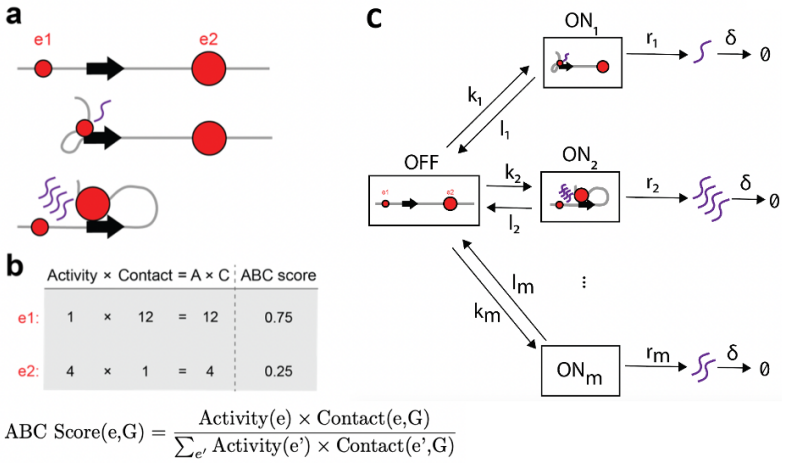

Figure 12: A mathematical formalization of the ABC Model **(a)** Cartoon description of ABC model. Same as in Fig 3 **(b)** Computation of ABC Score. Same as in Fig 3 **(c)** Simplest possible formulation of the ABC Model using Markov processes.

In order to derive the ABC Score from the multi-ON model, we must first analyze the steady-state behavior of the multi-ON model. We employ a seminal result of Sanchez and Kondev [33] which states that the mean steady-state mRNA copy number in any copy number graph is given by:

$$\text{Mean} = \frac{r \cdot \mu}{\delta} \tag{6.1}$$

where r is the vector of production rates, δ is the mRNA degradation rate and μ is the steady-state probabilities of the underlying regulatory graph G. In order to compute μ in the multi-ON model we use a general result known as the Matrix Tree Theorem (MTT) [17]. The MTT states that μ can be calculated as

$$\mu_i := \frac{\rho_i}{\rho_1 + \rho_2 + \cdots + \rho_N} \tag{6.2}$$

where

$$\rho_i = \sum_{T \in \Theta_i(G)} \prod_{j \to k \in T} e(j \to k) \tag{6.3}$$

Here $\Theta_i(G)$ denotes the set of all spanning trees of G rooted at vertex i and the product is over all edge labels of the tree. A spanning tree in a directed graph G is a tree (a tree is a connected and has no cycles) which includes all vertices of the graph. A spanning tree is rooted at vertex i, if i is the only vertex in the tree that has no outgoing edges.

The advantage of the MTT is that in certain graphs enumerating spanning trees is very easy. The multi-ON graph is one such example, there is only 1 spanning tree rooted at each vertex, making the computation trivial (Fig 13)

Indexing OFF as state 1 and ON_i as state $i+1$ and applying the MTT we have:

$$\rho = \begin{bmatrix} l_1 l_2 \dots l_m \\ k_1 l_2 \dots l_m \\ \vdots \\ l_1 l_2 \dots k_m \end{bmatrix}$$

Let $\lambda_i := \frac{k_i}{l_i}$ and $Z = 1 + \sum_{i=1}^{m} \lambda_i$. Dividing ρ by the scalar factor $l_1 l_2 \dots l_m$ and normalizing yields the steady-state probability distribution

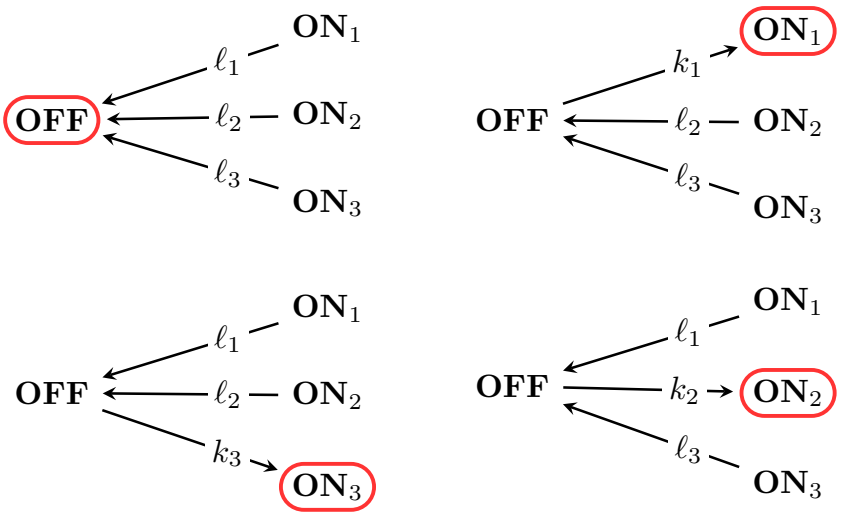

Figure 13: Spanning trees in the multi-ON regulatory graph with 3 enhancers. The root of each tree is outlined in red. There is only one spanning tree rooted at each vertex.

$$\mu = \frac{1}{Z}\begin{bmatrix} 1 \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{bmatrix} \tag{6.4}$$

Applying (6.1) gives the mean mRNA copy number in the multi-ON model as

$$\text{Mean} = \frac{1}{\delta}\frac{1}{Z}\sum_{i=1}^{m} r_i \frac{k_i}{l_i} \tag{6.5}$$

6.3 Deriving the ABC Score

We now show how the ABC Score can be derived from the mean mRNA copy number of the multi-ON model. We recall that in the experiments described in [14], the mean mRNA copy number, $\langle G \rangle$, of a particular gene is measured in a control condition and the mean is measured in a condition in which enhancer j of the gene is perturbed, which we denote $\langle G^{\Delta_j} \rangle$. The ABC Score intends to predict the fractional change in expression upon perturbation:

$$\text{Fraction Change in Expression} = \frac{\langle G \rangle - \langle G^{\Delta_j} \rangle}{\langle G \rangle} \tag{6.6}$$

The method used to perturb enhancers in [14] is CRISPR interference (CRISPRi). We must now consider how to apply the effects of a CRISPRi perturbation to enhancer j in the mutli-ON model. This choice is subtle because the relationship between the biology of CRISPRi and the multi-ON model is not clear. The potential options would be to change one of the rates associated to state ON_j (k_j, l_j or r_j) or to just remove vertex ON_j from the graph. Below we proceed by setting r_j equal to 0 which eliminates any transcription due to enhancer j. Removing ON_j from the graph (or equivalently letting $k_j \to 0$ or $l_j \to \infty$) will result in similar qualitative behavior with some technical caveats not discussed here.

Using (6.5) and (6.6) and making the assumption that r_j is the only rate parameter that changes in the Δ_j condition leads to

$$\text{Fraction change in expression upon CRISPRi of enhancer j} = \frac{r_j \lambda_j}{\sum_i r_i \lambda_i} \tag{6.7}$$

We note that this expression has the identical mathematical form as the ABC Score (Fig 12b). Equations (6.6) and (6.7) also highlight the power of conducting perturbational experiments. By forming the ratio in (6.6), the dependence on the degradation rate δ cancels out. This is advantageous as it becomes possible to make the fold-change predictions without measuring the mRNA decay rate.

A much more extensive discussion of these formal models and their relation to the ABC model is forthcoming [Nasser et al, in preparation].

Chapter 7

Discussion

Understanding the role of enhancers in gene regulation is an active field of modern research. There are many broad open questions that are being investigated: What are the biophyiscal and biochemical mechanisms by which enhancers regulate their target genes? How does enhancer-gene specificity arise? Can we experimentally map and computationally predict enhancer-gene conncetions? What is the role of enhancers in development, signaling, disease and other cellular or physiological phenomena?

The ABC model provides some insight into these questions. On its most basic level, as a practical predictor, the ABC model provides a means by which to computationally predict enhancer-gene connections. Our results suggest that cell-type specific Hi-C data is not needed in order to make cell-type specific enhancer-gene predictions (at least for some subset of enhancer-gene connections). As such the requirements of the model are ATAC-Seq and H3K27ac ChIP-Seq. These requirements are modest and will facilitate the generation of large scale databases of predicted enhancer-gene connections. These databases can then be mined to interpret noncoding genomic variation as described in Chaper 4.

Can we learn anything new about enhancer biology from the predictive ability of the ABC Model? This is a difficult question to answer. The initial conceptualization of ABC was based on seminal advances in our understanding of enhancer biology. Early investigations into enhancers showed the importance of chromatin conformation, chromatin accessibility and suggested that histone modifications were correlated with enhancer activity. ABC was a simple way to put all these ideas together into a predictive model and provides further support for these conceptual ideas.

One important conclusion from the ABC story (so far!) is that a complicated process such as eukaryotic transcription can be effectively described using a simple model. We did not need to know the structural biology and interactions between each component of the transcriptional machinery in order to identify enhancers in the genome. Instead, we have settled for a higher level representation of the system. We have arrived at a model which is just precise enough to describe the experiments: CRISPRi-FlowFISH experiments perturb the enhancer as a whole, and our model is represented at the level of the enhancer. An experiment with a more targeted perturbation (such as a single base-pair change to an enhancer element, or a single amino-acid substitution to a transcription factor) would require a model that ex-

plicitly contains this perturbation component.

The results presented herein describe the first iteration of the ABC model. While the performance of the model is sufficiently high to be used as a practical predictor, there is still substantial room for improvement. Specific open questions include

- Do the results for ubiquitously expressed genes hold in general? Can we identify a set of genes that are not regulated by distal enhancers?
- To what extent is cell-type specific data needed in order to make cell type specific predictions? How well does the using the average Hi-C dataset do? Are the enhancer-gene connections that need cell-type specific Hi-C data to be identified special in some way?
- The ABC Score is currently mainly being used as a binary predictor. However, our results do show that the ABC score is correlated with the effect size observed in the CRISPR screens. Can we actually predict effect size in a parameter free manner? Can we predict higher order features of transcription such as gene expression variance or bursting kinetics?
- ABC currently operates at the level of 500bp enhancer elements. Is it possible to get a sequence-level understanding of enhancers?
- What is the best path forward? Should we try to develop better ways to measure 'Activity' (say by combining multiple epigenomic datasets or through MPRA reporter assays) and 'Contact' (with higher resolution Hi-C)? Or is an entirely new modeling framework needed?

Tackling such questions will require much more experimental data. Larger experimentally derived CRISPR datasets will need to be generated which can be used as gold-standard datasets to develop further improvements to ABC. Improvements in enhancer screening technologies may facilitate the generation of such datasets.

In addition to more enhancer screens, the best way to improve enhancer prediction methods is to increase our understanding of enhancer biology. Open questions remain such as (i) how do enhancers regulate gene expression over large physical distances? (ii) what is the role of enhancer RNAs? (iii) what step in the transcription process do enhancers regulate? (iv) how do multiple transcription factors work together at the level of a single enhancer element? (v) how do multiple

enhancers work together to regulate gene expression? Progress on each of these topics can lead to better enhancers models. But how do we actually do this? How do we convert basic experimental advances in enhancer biology into a quantitative models of enhancer regulation?

We suggest that the results of Chapter 6 may form the basis for such a strategy. Such an approach is a way to convert informal models (depicted as cartoons or described using words) into formal mathematical models that make testable predictions. Such formalizations may not always be useful. However, the motivating idea is that if we can convert conceptual ideas into quantitative models, and if these models make quantitative predictions, and if these predictions are confirmed by experiment, then this is good evidence that we have achieved a reasonable understanding of the system which is being investigated. Such an approach has been conducted in the context of prokaryotic gene regulation [29]. It may be worthwhile to attempt such efforts in the eukaryotic paradigm.

The field of eukaryotic gene regulation is at an exciting moment. Experimental advances have enabled new ways to observe and perturb the components involved in gene regulation. A central challenge will be to convert these experimental observations into a comprehensive understanding of gene regulation. We suggest modeling can be a way to synthesize our understanding at the conceptual and quantitative level. The work in this thesis provides a case study for such an approach in the context of enhancers and suggests this approach may be more widely applicable.

References

[1] J. Banerji, L. Olson, and W. Schaffner. "A Lymphocyte-Specific Cellular Enhancer Is Located Downstream of the Joining Region in Immunoglobulin Heavy Chain Genes". In: *Cell* 33.3 (July 1983), pp. 729–740. ISSN: 0092-8674. DOI: 10.1016/0092-8674(83)90015-6.

[2] Alexander G. Bick et al. "Inherited Causes of Clonal Haematopoiesis in 97,691 Whole Genomes". In: *Nature* 586.7831 (Oct. 2020), pp. 763–768. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2819-2.

[3] Hugo B. Brandão, Michele Gabriele, and Anders S. Hansen. "Tracking and Interpreting Long-Range Chromatin Interactions with Super-Resolution Live-Cell Imaging". In: *Current Opinion in Cell Biology*. Cell Nucleus 70 (June 2021), pp. 18–26. ISSN: 0955-0674. DOI: 10.1016/j.ceb.2020.11.002.

[4] Eliezer Calo and Joanna Wysocka. "Modification of Enhancer Chromatin: What, How and Why?" In: *Molecular cell* 49.5 (Mar. 2013), 10.1016/j.molcel.2013.01.038. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2013.01.038.

[5] Qin Cao et al. "Reconstruction of Enhancer-Target Networks in 935 Samples of Human Primary Cells, Tissues and Cell Lines". In: *Nature Genetics* 49.10 (Oct. 2017), pp. 1428–1436. ISSN: 1546-1718. DOI: 10.1038/ng.3950.

[6] Melina Claussnitzer et al. "A Brief History of Human Disease Genetics". In: *Nature* 577.7789 (Jan. 2020), pp. 179–189. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1879-7.

[7] M C Driscoll, C S Dobkin, and B P Alter. "Gamma Delta Beta-Thalassemia Due to a de Novo Mutation Deleting the 5' Beta-Globin Gene Activation-Region Hypersensitive Sites." In: *Proceedings of the National Academy of Sciences of the United States of America* 86.19 (Oct. 1989), pp. 7470–7474. ISSN: 0027-8424.

[8] Eli Eisenberg and Erez Y. Levanon. "Human Housekeeping Genes, Revisited". In: *Trends in genetics: TIG* 29.10 (Oct. 2013), pp. 569–574. ISSN: 0168-9525. DOI: 10.1016/j.tig.2013.05.010.

[9] Jesse M. Engreitz, Jenna E. Haines, Elizabeth M. Perez, Glen Munson, Jenny Chen, Michael Kane, Patrick E. McDonel, Mitchell Guttman, and Eric S. Lander. "Local Regulation of Gene Expression by lncRNA Promoters, Transcription and Splicing". In: *Nature* 539.7629 (Nov. 2016), pp. 452–455. ISSN: 1476-4687. DOI: 10.1038/nature20149.

[10] Jason Ernst et al. "Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types". In: *Nature* 473.7345 (May 2011), pp. 43–49. ISSN: 1476-4687. DOI: 10.1038/nature09906.

[11] Alistair R. R. Forrest et al. "A Promoter-Level Mammalian Expression Atlas". In: *Nature* 507.7493 (Mar. 2014), pp. 462–470. ISSN: 1476-4687. DOI: 10.1038/nature13182.

[12] Daniel R Fuentes, Tomek Swigut, and Joanna Wysocka. "Systematic Perturbation of Retroviral LTRs Reveals Widespread Long-Range Effects on Human Gene Regulation". In: *eLife* 7 (Aug. 2018). Ed. by Edith Heard and Detlef Weigel, e35989. ISSN: 2050-084X. DOI: 10.7554/eLife.35989.

[13] Charles P. Fulco, Mathias Munschauer, Rockwell Anyoha, Glen Munson, Sharon R. Grossman, Elizabeth M. Perez, Michael Kane, Brian Cleary, Eric S. Lander, and Jesse M. Engreitz. "Systematic Mapping of Functional Enhancer–Promoter Connections with CRISPR Interference". In: *Science* 354.6313 (Nov. 2016), pp. 769–773. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aag2445.

[14] Charles P. Fulco et al. "Activity-by-Contact Model of Enhancer–Promoter Regulation from Thousands of CRISPR Perturbations". In: *Nature Genetics* 51.12 (Dec. 2019), pp. 1664–1669. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0538-0.

[15] Eileen E. M. Furlong and Michael Levine. "Developmental Enhancers and Chromosome Topology". In: *Science (New York, N.Y.)* 361.6409 (Sept. 2018), pp. 1341–1345. ISSN: 0036-8075. DOI: 10.1126/science.aau0320.

[16] Molly Gasperini et al. "A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens". In: *Cell* 176.1-2 (Jan. 2019), 377–390.e19. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.11.029.

[17] Jeremy Gunawardena. "A Linear Framework for Time-Scale Separation in Nonlinear Biochemical Systems". In: *PLOS ONE* 7.5 (May 2012), e36321. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0036321.

[18] Farhad Hormozdiari et al. "Functional Disease Architectures Reveal Unique Biological Role of Transposable Elements". In: *Nature Communications* 10.1 (Sept. 2019), p. 4054. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11957-5.

[19] Martin Kampmann. "CRISPRi and CRISPRa Screens in Mammalian Cells for Precision Biology and Medicine". In: *ACS chemical biology* 13.2 (Feb. 2018), pp. 406–416. ISSN: 1554-8929. DOI: 10.1021/acschembio.7b00657.

[20] Matthew H. Larson, Luke A. Gilbert, Xiaowo Wang, Wendell A. Lim, Jonathan S. Weissman, and Lei S. Qi. "CRISPR Interference (CRISPRi) for Sequence-Specific Control of Gene Expression". In: *Nature Protocols* 8.11 (Nov. 2013), pp. 2180–2196. ISSN: 1750-2799. DOI: 10.1038/nprot.2013.132.

[21] Laura A. Lettice, Simon J. H. Heaney, Lorna A. Purdie, Li Li, Philippe de Beer, Ben A. Oostra, Debbie Goode, Greg Elgar, Robert E. Hill, and Esther de Graaff. "A Long-Range Shh Enhancer Regulates Expression in the Developing Limb and Fin and Is Associated with Preaxial Polydactyly". In: *Human Molecular Genetics* 12.14 (July 2003), pp. 1725–1735. ISSN: 0964-6906. DOI: 10.1093/hmg/ddg180.

[22] Guoliang Li et al. "Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation". In: *Cell* 148.1-2 (Jan. 2012), pp. 84–98. ISSN: 1097-4172. DOI: 10.1016/j.cell.2011.12.014.

[23] Jamie L. Marshall et al. "HyPR-seq: Single-cell Quantification of Chosen RNAs via Hybridization and Sequencing of DNA Probes". In: *Proceedings of the National Academy of Sciences* 117.52 (Dec. 2020), pp. 33404–33413. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2010738117.

[24] Jamie L. Marshall et al. "HyPR-seq: Single-cell Quantification of Chosen RNAs via Hybridization and Sequencing of DNA Probes". In: *Proceedings of the National Academy of Sciences* 117.52 (Dec. 2020), pp. 33404–33413. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2010738117.

[25] Matthew T. Maurano et al. "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA". In: *Science (New York, N.Y.)* 337.6099 (Sept. 2012), pp. 1190–1195. ISSN: 1095-9203. DOI: 10.1126/science.1222794.

[26] P. Moreau, R. Hen, B. Wasylyk, R. Everett, M.P. Gaub, and P. Chambon. "The SV40 72 Base Repair Repeat Has a Striking Effect on Gene Expression Both in SV40 and Other Chimeric Recombinants". In: *Nucleic Acids Research* 9.22 (Nov. 1981), pp. 6047–6068. ISSN: 0305-1048. DOI: 10.1093/nar/9.22.6047.

[27] Maxwell R. Mumbach, Adam J. Rubin, Ryan A. Flynn, Chao Dai, Paul A. Khavari, William J. Greenleaf, and Howard Y. Chang. "HiChIP: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture". In: *Nature Methods* 13.11 (Nov. 2016), pp. 919–922. ISSN: 1548-7105. DOI: 10.1038/nmeth.3999.

[28] Joseph Nasser et al. "Genome-Wide Enhancer Maps Link Risk Variants to Disease Genes". In: *Nature* 593.7858 (May 2021), pp. 238–243. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03446-x.

[29] Rob Phillips, Nathan M. Belliveau, Griffin Chure, Hernan G. Garcia, Manuel Razo-Mejia, and Clarissa Scholes. "Figure 1 Theory Meets Figure 2 Experiments in the Study of Gene Expression". In: *Annual Review of Biophysics* 48.1 (2019), pp. 121–163. DOI: 10.1146/annurev-biophys-052118-115525.

[30] Suhas S.P. Rao et al. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping". In: *Cell* 159.7 (Dec. 2014), pp. 1665–1680. ISSN: 00928674. DOI: 10.1016/j.cell.2014.11.021.

[31] Yakir A. Reshef et al. "Detecting Genome-Wide Directional Effects of Transcription Factor Binding on Polygenic Disease Risk". In: *Nature Genetics* 50.10 (Oct. 2018), pp. 1483–1493. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0196-7.

[32] Adrian L. Sanborn et al. "Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-Type and Engineered Genomes". In: *Proceedings of the National Academy of Sciences* 112.47 (Nov. 2015), E6456–E6465. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1518552112.

[33] Álvaro Sánchez and Jané Kondev. "Transcriptional Control of Noise in Gene Expression". In: *Proceedings of the National Academy of Sciences* 105.13 (Apr. 2008), pp. 5081–5086. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0707904105.

[34] Stefan Schoenfelder and Peter Fraser. "Long-Range Enhancer–Promoter Contacts in Gene Expression Control". In: *Nature Reviews Genetics* 20.8 (Aug. 2019), pp. 437–455. ISSN: 1471-0064. DOI: 10.1038/s41576-019-0128-0.

[35] François Spitz and Eileen E. M. Furlong. "Transcription Factors: From Enhancer Binding to Developmental Control". In: *Nature Reviews Genetics* 13.9 (Sept. 2012), pp. 613–626. ISSN: 1471-0064. DOI: 10.1038/nrg3207.

[36] Quentin Szabo, Frédéric Bantignies, and Giacomo Cavalli. "Principles of Genome Folding into Topologically Associating Domains". In: *Science Advances* 5.4 (), eaaw1668. DOI: 10.1126/sciadv.aaw1668.

[37] Bas Tolhuis, Robert-Jan Palstra, Erik Splinter, Frank Grosveld, and Wouter de Laat. "Looping and Interaction between Hypersensitive Sites in the Active β-Globin Locus". In: *Molecular Cell* 10.6 (Dec. 2002), pp. 1453–1465. ISSN: 1097-2765. DOI: 10.1016/S1097-2765(02)00781-5.

[38] Sean Whalen, Rebecca M. Truty, and Katherine S. Pollard. "Enhancer-Promoter Interactions Are Encoded by Complex Genomic Signatures on Looping Chromatin". In: *Nature genetics* 48.5 (May 2016), pp. 488–496. ISSN: 1061-4036. DOI: 10.1038/ng.3539.

[39] Jiang Zhu, Fuhong He, Shuhui Song, Jing Wang, and Jun Yu. "How Many Human Genes Can Be Defined as Housekeeping with Current Expression Data?" In: *BMC Genomics* 9.1 (Apr. 2008), p. 172. ISSN: 1471-2164. DOI: 10.1186/1471-2164-9-172.

[40] Jessica Zuin et al. *Nonlinear Control of Transcription through Enhancer-Promoter Interactions*. Apr. 2021. DOI: 10.1101/2021.04.22.440891.

Appendix A

Work published by the author and co-author declarations

This thesis is based on four published peer reviewed articles. Some unpublished work is described in chapter 6. The four published articles are listed below with my contributions to them. Attestations from co-authors follow below.

1. Charles P. Fulco, Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, Rockwell Anyoha, Benjamin R. Doughty, Tejal A. Patwardhan, Tung H. Nguyen, Michael Kane, Elizabeth M. Perez, Neva C. Durand, Caleb A. Lareau, Elena K. Stamenova, Erez Lieberman Aiden, Eric S. Lander, and Jesse M. Engreitz. "Activity-by-Contact Model of Enhancer–Promoter Regulation from Thousands of CRISPR Perturbations". In: *Nature Genetics* 51.12 (Dec. 2019), pp. 1664–1669. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0538-0

   I was the lead computationalist on this paper and made major contributions to all aspects of ABC model development, analysis/interpretation of experimental data, codebase development and manuscript preparation. I led the development of a github repository which implements the ABC model.

2. Joseph Nasser, Drew T. Bergman, Charles P. Fulco, Philine Guckelberger, Benjamin R. Doughty, Tejal A. Patwardhan, Thouis R. Jones, Tung H. Nguyen, Jacob C. Ulirsch, Fritz Lekschas, Kristy Mualim, Heini M. Natri, Elle M. Weeks, Glen Munson, Michael Kane, Helen Y. Kang, Ang Cui, John P. Ray, Thomas M. Eisenhaure, Ryan L. Collins, Kushal Dey, Hanspeter Pfister, Alkes L. Price, Charles B. Epstein, Anshul Kundaje, Ramnik J. Xavier, Mark J. Daly, Hailiang Huang, Hilary K. Finucane, Nir Hacohen, Eric S. Lander, and Jesse M. Engreitz. "Genome-Wide Enhancer Maps Link Risk Variants to Disease Genes". In: *Nature* 593.7858 (May 2021), pp. 238–243. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03446-x

   I was a lead computationalist on this paper and made major contributions to project conceptualization, analysis/interpretation of experimental data, codebase development and manuscript preparation. I laid the foundations for this project by generating a database of enhancer-gene predictions using the ABC model in 131 biosamples. My contributions then enabled the rest of the paper and his results are featured in all other main and supplemental figures. I also had a significant role in leading the overall conceptualization and direction of this project and conducted many of the initial/pilot analysis.

3. Jamie L. Marshall, Benjamin R. Doughty, Vidya Subramanian, Philine Guckelberger, Qingbo Wang, Linlin M. Chen, Samuel G. Rodriques, Kaite Zhang, Charles P. Fulco, Joseph Nasser, Elizabeth J. Grinkevich, Teia Noel, Sarah Mangiameli, Drew T. Bergman, Anna Greka, Eric S. Lander, Fei Chen, and Jesse M. Engreitz. "HyPR-seq: Single-cell Quantification of Chosen RNAs via Hybridization and Sequencing of DNA Probes". In: *Proceedings of the National Academy of Sciences* 117.52 (Dec. 2020), pp. 33404–33413. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2010738117

   My main contribution to this paper was in the experimental design of several pilot experiments. I performed power calculations to determine the size of the experiment that would be required (number of cells, number of reads per cell, number of guide RNAs, etc) in order to have sufficient statistical power to detect the expected effect sizes. These calculations were then used to actually design the experiment

4. Alexander G. Bick, Joshua S. Weinstock, Satish K. Nandakumar, Charles P. Fulco, Erik L. Bao, Seyedeh M. Zekavat, Mindy D. Szeto, Xiaotian Liao, Matthew J. Leventhal, Joseph Nasser, Kyle Chang, Cecelia Laurie, Bala Bharathi Burugula, Christopher J. Gibson, Abhishek Niroula, Amy E. Lin, Margaret A. Taub, Francois Aguet, Kristin Ardlie, Braxton D. Mitchell, Kathleen C. Barnes, Arden Moscati, Myriam Fornage, Susan Redline, Bruce M. Psaty, Edwin K. Silverman, Scott T. Weiss, Nicholette D. Palmer, Ramachandran S. Vasan, Esteban G. Burchard, Sharon L. R. Kardia, Jiang He, Robert C. Kaplan, Nicholas L. Smith, Donna K. Arnett, David A. Schwartz, Adolfo Correa, Mariza de Andrade, Xiuqing Guo, Barbara A. Konkle, Brian Custer, Juan M. Peralta, Hongsheng Gui, Deborah A. Meyers, Stephen T. McGarvey, Ida Yii-Der Chen, M. Benjamin Shoemaker, Patricia A. Peyser, Jai G. Broome, Stephanie M. Gogarten, Fei Fei Wang, Quenna Wong, May E. Montasser, Michelle Daya, Eimear E. Kenny, Kari E. North, Lenore J. Launer, Brian E. Cade, Joshua C. Bis, Michael H. Cho, Jessica Lasky-Su, Donald W. Bowden, L. Adrienne Cupples, Angel C. Y. Mak, Lewis C. Becker, Jennifer A. Smith, Tanika N. Kelly, Stella Aslibekyan, Susan R. Heckbert, Hemant K. Tiwari, Ivana V. Yang, John A. Heit, Steven A. Lubitz, Jill M. Johnsen, Joanne E. Curran, Sally E. Wenzel, Daniel E. Weeks, Dabeeru C. Rao, Dawood Darbar, Jee-Young Moon, Russell P. Tracy, Erin J. Buth, Nicholas Rafaels, Ruth J. F. Loos, Peter Durda, Yongmei Liu, Lifang Hou, Jiwon Lee, Priyadarshini

Kachroo, Barry I. Freedman, Daniel Levy, Lawrence F. Bielak, James E. Hixson, James S. Floyd, Eric A. Whitsel, Patrick T. Ellinor, Marguerite R. Irvin, Tasha E. Fingerlin, Laura M. Raffield, Sebastian M. Armasu, Marsha M. Wheeler, Ester C. Sabino, John Blangero, L. Keoki Williams, Bruce D. Levy, Wayne Huey-Herng Sheu, Dan M. Roden, Eric Boerwinkle, JoAnn E. Manson, Rasika A. Mathias, Pinkal Desai, Kent D. Taylor, Andrew D. Johnson, Paul L. Auer, Charles Kooperberg, Cathy C. Laurie, Thomas W. Blackwell, Albert V. Smith, Hongyu Zhao, Ethan Lange, Leslie Lange, Stephen S. Rich, Jerome I. Rotter, James G. Wilson, Paul Scheet, Jacob O. Kitzman, Eric S. Lander, Jesse M. Engreitz, Benjamin L. Ebert, Alexander P. Reiner, Siddhartha Jaiswal, Gonçalo Abecasis, Vijay G. Sankaran, Sekar Kathiresan, and Pradeep Natarajan. "Inherited Causes of Clonal Haematopoiesis in 97,691 Whole Genomes". In: *Nature* 586.7831 (Oct. 2020), pp. 763–768. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2819-2

My contributions to this publication center on the interpretation of the signal at the TET2 locus. I was responsible for the bioinformatic analysis leading to the predicted function of the causal variant at the TET2 locus. I generated enhancer-gene predictions using the ABC model and noted that the causal variant at the TET2 locus overlapped a predicted enhancer in haematopoietic progenitor cells. This prediction was subsequently validated experimentally which constituted a major contribution to the publication as a whole. I also assisted in writing and editing the portion of the manuscript relating to these analyses.

Stanford University

Jesse M. Engreitz, Ph.D.

Assistant Professor
Department of Genetics
Stanford University

Biomedical Innovation Building
240 Pasteur Drive, Room 3753
Stanford, CA 94305-5162
Phone: +1 (650) 497-2434
Email: engreitz@stanford.edu

BASE Research Initiative
Betty Irene Moore Children's Heart Center
Lucile Packard Children's Hospital

December 10, 2021

To whom it may concern:

The purpose of this letter is to document the contributions of Joseph (Joe) Nasser to the publications listed below.

Joe joined my group and the lab of Dr. Eric Lander at the Broad Institute of MIT and Harvard, located in Cambridge, Massachusetts, USA, as a staff computational biologist in June 2017. (There, I was a Junior Fellow at the Harvard Society of Fellows and led a research team for 4 years. I am now an Assistant Professor at Stanford University as of May 2020.)

Charles P. Fulco*, **Joseph Nasser*** et al, Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. ***Nature Genetics.*** 51, 1664–1669 (2019). https://www.nature.com/articles/s41588-019-0538-0
*Denotes equal contribution

Joe was the lead computational author on this paper and made major contributions to all aspects of ABC model development, analysis/interpretation of experimental data, codebase development and manuscript preparation. Joe conducted analysis related to and created main figures 2c-e, 3 and 4a-b, extended data figures 3 and 4, and supplementary figures 3, 4, 6, 7, 9 and 11. Joe led the development of a github repository which implements the ABC model.

Joseph Nasser*, Drew T. Bergman*, Charles P. Fulco*, Philine Guckelberger*, Benjamin R. Doughty* et al, Genome-wide enhancer maps link risk variants to disease genes. **Nature** 593, 238–243 (2021). https://www.nature.com/articles/s41586-021-03446-x

Joe was a lead computational author on this paper and made major contributions to project conceptualization, analysis/interpretation of experimental data, codebase development and manuscript preparation. Joe laid the foundations for this project by optimizing methods to apply ABC robustly to diverse datasets, including standardizing scores across cell types and datasets of varying qualities. Joe applied this approach to build a database of enhancer-gene predictions using the ABC model in 131 cell types and tissues. These predictions were used throughout the paper are featured in all other main and supplemental figures. Joe also had a significant role in the conceptualization of this project and conducted analyses to assess the properties of ABC predictions across cell types and tissues.

Jamie L. Marshall*, Benjamin R.Doughty*, Vidya Subramanian, Qingbo Wang, Linlin M. Chen, Samuel G. Rodriques, Kaite Zhang , Philine Guckelberger, Charles P. Fulco , **Joseph Nasser**, Elizabeth J. Grinkevich, Teia Noel, Sarah Mangiameli, Anna Greka, Eric S. Lander,

Fei Chen, Jesse M. Engreitz, HyPR- seq: Single-cell quantification of chosen RNAs via hybridization and sequencing of DNA probes, **PNAS** 2020 117 (52) https://www.pnas.org/content/117/52/33404

Joe's contributed to the experimental design of several pilot experiments (Figure 2). Specifically, Joe designed and conducted power calculations to determine the size of the experiment that would be required (number of cells, number of reads per cell, and number of guide RNAs) in order to have sufficient statistical power to detect the expected effect sizes.

Sincerely,

Jesse Engreitz, Ph.D.
Assistant Professor
Department of Genetics, Stanford University School of Medicine
BASE Research Initiative, Betty Irene Moore Children's Heart Center

Pradeep Natarajan, MD MMSc
Director of Preventive Cardiology, Massachusetts General Hospital
Associate Member, Broad Institute of Harvard & MIT
Assistant Professor of Medicine, Harvard Medical School
185 Cambridge St, CPZN 3.184, Boston, MA 02114
pnatarajan@mgh.harvard.edu
617-724-3526

2021-03-18

To Whom It May Concern:

The purpose of this letter is to document the contributions of Joseph (Joe) Nasser to the publication:

Bick, A.G., Weinstock, J.S., Nandakumar, S.K. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586,** 763–768 (2020)

This publication investigates genetic risk for clonal haematopoiesis (CHIP). In this study, whole genome sequencing is performed in 97,691 individuals which is then used to conduct a genome-wide association study for CHIP. The publication identifies three genomic loci associated with CHIP status, including a non-coding signal at the TET2 locus.
Joe's contributions to this publication center on the interpretation of the signal at the TET2 locus. Joe was responsible for the bioinformatic analysis leading to the predicted function of the causal variant at the TET2 locus (Figure 3a-c). Specifically, Joe generated enhancer-gene predictions using the ABC model and noted that the causal variant at the TET2 locus overlapped a predicted enhancer in haematopoietic progenitor cells. This prediction was subsequently validated experimentally which constituted a major contribution to the publication as a whole. Joe also assisted in writing and editing the portion of the manuscript relating to these analyses.

For your reference, I am the Director of Preventive Cardiology at Massachusetts General Hospital (MGH), Assistant Professor of Medicine at Harvard Medical School, and Associate Member of the Broad Institute. I oversee clinical and training programs in primary and secondary prevention of atherosclerotic cardiovascular disease at MGH. In addition, I direct a research program across the MGH Cardiovascular Research Center and the Broad Institute Cardiovascular Disease Initiative supported primarily by the NIH, American Heart Association, and Fondation Leducq. My research program focuses on using germline genetics and somatic genetics from blood cells to understand the causes of atherosclerotic cardiovascular disease to

improve prevention. We use conventional epidemiology, statistical genetics, biomedical informatics, and human investigation to address these issues.

Sincerely,

Pradeep Natarajan, MD MMSc

Bristol Myers Squibb™

November 26, 2021

To whom it may concern,

The purpose of this letter is to verify and document the scientific contributions of Joseph Nasser to the paper: Charles P. Fulco*, Joseph Nasser* *et al*. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. ***Nature Genetics***. 51, 1664-1669 (2019). https://doi.org/10.1038/s41588-019-0538-0

Joe was the lead computational biologist on this project and made major contributions to all aspects of ABC model development, analysis and interpretation of experimental data, codebase development and manuscript preparation. Joe conducted analysis for and created Figures 2c-e, 3, and 4a-b; Extended Data Figures 3 and 4; and Supplementary Figures 3, 4, 6, 7, 9, and 11.

I conducted research on this study as a graduate student in the laboratory of Eric Lander at the Broad Institute of MIT and Harvard. I am now a Principal Scientist at Bristol Myers Squibb.

Sincerely,

Charles Fulco

Charles Fulco, Ph.D.
Principal Scientist | Early Target Biology
Research and Early Development

100 Binney Street
Cambridge MA 02142

Pronouns He/His
charles.fulco@bms.com

BROAD INSTITUTE

415 Main Street
Cambridge, MA 02142
T 617-714-7000 F 617-714-8972
www.broadinstitute.org

November 29, 2021

To whom it may concern:

The purpose of this letter is to document the contributions of Joseph (Joe) Nasser to the publication:

Jamie L. Marshall, Benjamin R.Doughty, Vidya Subramanian,Qingbo Wang, Linlin M. Chen, Samuel G. Rodriques, Kaite Zhang , Philine Guckelberger, Charles P. Fulco , **Joseph Nasser**, Elizabeth J. Grinkevich, Teia Noel, Sarah Mangiameli, Anna Greka, Eric S. Lander, Fei Chen, Jesse M. Engreitz, HyPR- seq: Single-cell quantification of chosen RNAs via hybridization and sequencing of DNA probes, **PNAS** 2020 117 (52) https://www.pnas.org/content/117/52/33404

This paper presents a novel experimental approach (HyPR-seq) to sensitively detect selected RNAs in single cells.

Joe's main contribution to this paper was in the experimental design of several pilot experiments (Figure 2). Joe performed various power calculations to determine the size of the experiment that would be required (number of cells, number of reads per cell, number of guide RNAs, etc) in order to have sufficient statistical power to detect the expected effect sizes. These calculations were then used to actually design the experiment. Joe also analyzed the results of the experiment to assess the suitability of various assumptions made in the power calculations.

I have been a Senior Group Leader at the Broad Institute of MIT and Harvard in the Kidney Disease Initiative since 2018 leading a group focused on the development of new technologies for quantifying RNA expression and determining the spatial location of RNA in the kidney in health and disease.

Sincerely,

[redacted]

Jamie L Marshall, PhD
Senior Group Leader
Kidney Disease Initiative

VANDERBILT UNIVERSITY MEDICAL CENTER

Genetic Medicine

11/21/21

To Whom It May Concern:

The purpose of this letter is to document the contributions of Joseph (Joe) Nasser to the publication:

Bick, A.G., Weinstock, J.S., Nandakumar, S.K. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586,** 763–768 (2020)

This publication investigates genetic risk for clonal haematopoiesis (CHIP). In this study, whole genome sequencing is performed in 97,691 individuals which is then used to conduct a genome-wide association study for CHIP. The publication identifies three genomic loci associated with CHIP status, including a non-coding signal at the TET2 locus.

Joe's contributions to this publication center on the interpretation of the signal at the TET2 locus. Joe was responsible for the bioinformatic analysis leading to the predicted function of the causal variant at the TET2 locus (Figure 3a-c). Specifically, Joe generated enhancer-gene predictions using the ABC model and noted that the causal variant at the TET2 locus overlapped a predicted enhancer in haematopoietic progenitor cells. This prediction was subsequently validated experimentally which constituted a major contribution to the publication as a whole. Joe also assisted in writing and editing the portion of the manuscript relating to these analyses.

Sincerely,

Alexander Bick, MD PhD
Assistant Professor of Medicine
Vanderbilt University School of Medicine
Alexander.Bick@VUMC.org

507 Light Hall
Nashville, TN 37232-0275

tel 615.936.2660
fax 615.936.2510
https://medicine.mc.vanderbilt.edu/genetic-medicine-division-home

Appendix B

Full bibliography of works published by the author

1. Charles P. Fulco et al. “Activity-by-Contact Model of Enhancer–Promoter Regulation from Thousands of CRISPR Perturbations”. In: *Nature Genetics* 51.12 (Dec. 2019), pp. 1664–1669. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0538-0

2. Joseph Nasser et al. “Genome-Wide Enhancer Maps Link Risk Variants to Disease Genes”. In: *Nature* 593.7858 (May 2021), pp. 238–243. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03446-x

3. Jamie L. Marshall et al. “HyPR-seq: Single-cell Quantification of Chosen RNAs via Hybridization and Sequencing of DNA Probes”. In: *Proceedings of the National Academy of Sciences* 117.52 (Dec. 2020), pp. 33404–33413. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2010738117

4. Alexander G. Bick et al. “Inherited Causes of Clonal Haematopoiesis in 97,691 Whole Genomes”. In: *Nature* 586.7831 (Oct. 2020), pp. 763–768. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2819-2

5. Yakir A. Reshef et al. “Detecting Genome-Wide Directional Effects of Transcription Factor Binding on Polygenic Disease Risk”. In: *Nature Genetics* 50.10 (Oct. 2018), pp. 1483–1493. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0196-7

6. Farhad Hormozdiari et al. “Functional Disease Architectures Reveal Unique Biological Role of Transposable Elements”. In: *Nature Communications* 10.1 (Sept. 2019), p. 4054. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11957-5

Appendix C

Publications included in this thesis