Validating Functional Mechanisms For Non-Coding Genetic Variants Associated With Complex Traits

Cynthia Ann Kalita

Wayne State University, cakalita@gmail.com
VALIDATING FUNCTIONAL MECHANISMS FOR NON-CODING GENETIC VARIANTS ASSOCIATED WITH COMPLEX TRAITS

by

CYNTHIA A. KALITA

DISSERTATION

Submitted to the Graduate School,

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2018

MAJOR: MOLECULAR GENETICS AND GENOMICS

Approved By:

________________________________________
Advisor Date

________________________________________
________________________________________
________________________________________
________________________________________
DEDICATION

I would like to dedicate this to my family,

who have been nothing but supportive to me during this process.
ACKNOWLEDGEMENTS

I cannot thank Dr. Francesca Luca and Dr. Roger Pique-Regi enough for allowing me to pursue my Doctoral degree under their mentorship. Their guidance and enthusiasm for science kept me proceeding forward with my thesis even when obstacles occurred. I am grateful to all members of the LucaPique group, especially past members Dr. Greg Moyerbrailean and Dr. Allison Richards who would let me brainstorm my project with them. In addition, I would like to thank Andrew Freiman and Camilla Cascardo for their help in the lab. I would also like to thank my committee members who gave me their valuable guidance and insight.
# TABLE OF CONTENTS

Dedication ................................................................. ii

Acknowledgements ....................................................... iii

Chapter 1 Introduction ................................................... 1

Chapter 2 Computational predictions and analysis of non-coding variants for allele-specific expression ................................................................. 11

  Introduction ................................................................. 11
  Methods ................................................................. 15
  Results ................................................................. 26
  Discussion ............................................................... 38

Chapter 3 High throughput characterization of genetic effects on DNA:protein binding and gene transcription ................................................................. 41

  Introduction ................................................................. 41
  Methods ................................................................. 43
  Results ................................................................. 45
  Discussion ............................................................... 55

Chapter 4 Rapid profiling of regulatory regions and variants to study GxE in complex traits ................................................................. 58

  Introduction ................................................................. 58
  Methods ................................................................. 60
  Results ................................................................. 68
  Discussion ............................................................... 77

Chapter 5 Conclusion ..................................................... 79

Appendix A ................................................................. 82
LIST OF TABLES

Table 1  Statistical methods for ASE and MPRA analysis. ....................... 14
Table 2  Primers used in BiT-STARR-seq. ........................................ 64
Table 3  GEMMA per SNP heritability estimates relative to the genomic av-
erage. ................................................................. 70
Table 4  Number of identified factors/differential factors from ATAC-seq in
LCLs. ........................................................................ 75
Table 5  Summary of GWAS meta analysis traits examined. ...................... 84
Table 6  Factor binding sites enriched for GWAS SNPs. ............................ 84
Table 7  SNPs associated with GWAS traits that fall in Centipede-predicted
TF binding sites. .......................................................... 84
Table 8  Reporter gene assay results. .................................................... 84
Table 9  Annotations Used. ............................................................... 101
Table 10 QuASAR Results for LCLs. .................................................... 101
Table 11 QuASAR Results for HepG2. ................................................ 101
Table 12 Logistic Model predictions for LCLs. ....................................... 101
Table 13 Logistic Model predictions for HepG2. .................................... 101
Table 14 BiT-STARR-seq results. ....................................................... 113
Table 15 Annotations Used: CentiSNPs. .............................................. 113
Table 16 Annotations Used: GWAS. ................................................... 113
Table 17 Annotations Used: eQTL. ..................................................... 113
Table 18 Distribution of ASE Z-scores. ................................................. 114
Table 19 Transcription factors in BiT-STARR-seq. .................................. 114
Table 20 DEseq results: Combined concentrations. ............................... 114
Table 21 DEseq results: Low concentration. ........................................ 114
Table 22 DEseq results: Mid. ............................................................. 114
Table 23  DEseq results: High. ................................................. 114
Table 24  BiT-BUNDLE-seq results. ........................................ 114
Table 25  Designed Regulatory Category Content. .................... 115
Table 26  Enrichment for ASB and ASE variants in TFs. .............. 115
Table 27  ASB and complex traits. .......................................... 115
Table 28  ASE and complex traits. .......................................... 115
Table 29  Designed Regulatory Category Content. .................... 115
Table 30  Primers used in BiT-STARR-seq. ............................... 115
Table 31  Sequencing Depth. .................................................. 116
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>GWAS study outline and example hit.</td>
<td>2</td>
</tr>
<tr>
<td>Figure 2</td>
<td>High-throughput regulatory assays.</td>
<td>4</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Integration of annotations into GWAS results.</td>
<td>28</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Reporter gene assay validation of allelic regulatory activity.</td>
<td>30</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Comparing ASE testing methods in LCLs from Tewhey et al. (2016).</td>
<td>32</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Comparing ASE testing methods in Ulirsch et al. (2016).</td>
<td>33</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Exploring the performance across multiple simulated conditions.</td>
<td>35</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Validating experimental annotations in LCLs.</td>
<td>36</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Validating computational genomic annotations in LCLs.</td>
<td>38</td>
</tr>
<tr>
<td>Figure 10</td>
<td>BiT-STARR-seq and BiT-BUNDLE-seq identify regulatory variants in non-coding regions.</td>
<td>48</td>
</tr>
<tr>
<td>Figure 11</td>
<td>ASE for individual transcription factors.</td>
<td>50</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Allele-specific binding for NFKB1.</td>
<td>52</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Overlap with GWAS.</td>
<td>54</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Integration with GWAS.</td>
<td>73</td>
</tr>
<tr>
<td>Figure 15</td>
<td>Differential hypersensitive ATAC-seq.</td>
<td>74</td>
</tr>
<tr>
<td>Figure 16</td>
<td>ASE in treatments identified from BiT-STARR-seq.</td>
<td>76</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Enrichment for ASE in TF footprints matched by treatment.</td>
<td>77</td>
</tr>
<tr>
<td>Figure 18</td>
<td>Enrichment of transcription factors motifs from fgwas.</td>
<td>85</td>
</tr>
<tr>
<td>Figure 19</td>
<td>Association plots identifying SNPs in footprints.</td>
<td>86</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Association plots identifying SNPs in footprints from fgwas.</td>
<td>87</td>
</tr>
<tr>
<td>Figure 21</td>
<td>Comparing ASE testing methods in HepG2 from Tewhey et al. (2016).</td>
<td>103</td>
</tr>
<tr>
<td>Figure 22</td>
<td>Validating experimental annotations in HepG2.</td>
<td>103</td>
</tr>
<tr>
<td>Figure 23</td>
<td>Validating computational genomic annotations in HepG2.</td>
<td>103</td>
</tr>
</tbody>
</table>
Figure 24 Identifying cell type effects in LCLs. ........................................ 104
Figure 25 Identifying cell type effects in HepG2. ................................. 104
Figure 26 Comparing QuASAR and QuASAR-MPRA ASE testing methods in LCLs. ................................................................. 104
Figure 27 Schematic of oligos in BiT-STARR-seq and BiT-BUNDLE-seq. ........ 117
Figure 28 Correlation of DNA libraries. ............................................ 117
Figure 29 Correlation of RNA libraries. ............................................. 118
Figure 30 BiT-STARR-seq effect by regulatory category. ....................... 119
Figure 31 DNase window centering of BiT-STARR-seq variants. ............... 119
Figure 32 DNase peak size of BiT-STARR-seq variants. ......................... 120
Figure 33 Enrichment of NF-κB complex footprints in BiT-BUNDLE-seq bound regions. ................................................................. 120
Figure 34 Overlap between constructs with significant ASB for each concentration in BiT-BUNDLE-seq. ................................. 121
Figure 35 Overlap between constructs with significant ASB or ASE. ......... 121
Figure 36 Depletion of ASE with TFs that repress NFKB1 binding. .......... 122
CHAPTER 1 INTRODUCTION

GWAS and functional genomics: uncovering the molecular mechanisms behind association

Genome wide association studies (GWAS) have identified thousands of common genetic variants associated with complex traits, including normal traits and common diseases. For these studies, case and control individuals are genotyped, then differences in allele frequency at each single nucleotide polymorphism (SNP) are quantified. SNPs are determined to be associated with the trait if the case allele frequency is significantly different from that of the control group (Figure 1A). The end goal is to discover the genetic basis of complex traits in order to predict case conditions, for intervention, and for therapeutics. GWAS work by exploiting linkage disequilibrium (LD), with genotyping arrays sampling just enough SNPs to determine haplotype. These haplotypes are a result of historical evolutionary forces such as population size, mutation, recombination rate, and natural selection.

Despite many successful findings, GWAS have failed to explain the majority of estimated heritability. Part of this problem comes from the fact that GWAS identify large regions of association and in general, cannot directly pinpoint the true causative variant. Statistical power is an additional challenge. In order for a variant to reach genome-wide significance, it depends on sample size, effect size of the causal variant, and frequency of the variant. Despite the challenges, array style studies remain the most common due to how affordable they are. Unobserved variant frequencies can typically be recovered through statistical imputation by using haplotype information, therefore increasing statistical power for the study.
The success and failure of GWAS have spurred multiple approaches to better model population structure, identify novel variants, estimate and partition genetic covariance, or infer causality. One of the largest challenges that faces GWAS is in determining the functional relevance of human DNA sequence variants (Ward and Kellis, 2012). An estimated 80% of the genome (Consortium, 2012), and an ever larger percentage (90%) of GWAS hits (Maurano et al., 2012), lie outside of protein coding regions. This makes it difficult to infer mechanisms linking individual genetic variants with the disease trait (Figure 1B). In addition, we do not know under which environmental conditions the sequence variants have a functional impact, and whether they become one of many factors involved in complex phenotypes at the organismal level.

Figure 1: **GWAS study outline and example hit.** Source: Wellcome Trust Case Control Consortium (June 2007). A) Overview of GWAS study design. B) GWAS hit for Crohn’s Disease.
High-throughput assays to identify functional regulatory variants

Non-coding regions in the genome are largely made up of regulatory elements, for which it can be difficult to predict the functional outcome of variants within these sites. Functional genomics data collected by ENCODE (http://genome.ucsc.edu/ENCODE/), Roadmap Epigenome (http://www.roadmapepigenomics.org/), and other groups (Visel et al., 2009) have provided large amounts of information on regulatory regions of the human genome (Figure 2A). Protein binding to DNA mainly controls gene regulation and these proteins require open chromatin to access the DNA. DNA accessibility assays (e.g., DNase-seq (Song and Crawford, 2010), FAIRE-seq (Simon et al., 2012) or ATAC-seq (Buenrostro et al., 2013)) are used to identify these regions. DNase-seq uses an enzyme, DNaseI to cut DNA, where it can’t cut heterochromatin or regions bound by a protein (protected areas). The DNA is sequenced, and regions of open chromatin show peaks of increased cuts. However, even with identifying accessible regions this fails to provide regulatory mechanisms.

Transcription factors (TFs) are important gene regulators and include both enhancers and repressors or insulators. Enhancers are short DNA regions of 20-400bp that can activate promoters (Bondarenko et al., 2003) to increase gene expression. They work by stabilizing DNA-protein complexes at the enhancer and target promoter before activation or forming loops to bring the enhancer and the promoter closer together, facilitating communication between the enhancer and its target (Bondarenko et al., 2003; Ong and Corces, 2011; Krivega and Dean, 2012). Enhancers are well characterized as working position independent, and can target promoters both in cis and in trans. In contrast, insulators typically work by separating enhancers from the target promoter, creating separate domains.
Loop formation is the most common mechanism, and this is typically orientation dependent, needing the insulator and promoter in anti-parallel directions (West et al., 2002).


Chromatin immunoprecipitation sequencing (ChIP-seq) assays can directly identify where TFs are binding genome-wide, where the DNA-bound protein is immunoprecipitated using a specific antibody. There are some drawbacks to this method. While this method does capture genome wide binding, it is only for a single TF at a time, making this method low throughput. This method also relies on the availability of high quality antibodies to the TF of interest. Also, recent studies have found that 50% of the binding sites found may have
no role in transcriptional regulation (Li et al., 2008; Fisher et al., 2012; Teytelman et al., 2013; Lizio et al., 2015). This may be due to non-specific binding of TFs to the DNA, which allows for transient binding for non-functional weak interactions (Shlyueva et al., 2014).

Alternatively, binding events can be identified from DNase-seq data, which allows for identification of all potential TFs from a single assay. Centipede (Pique-Regi et al., 2011) is a program that detects these sites by integrating binding sites predicted by a motif model with genome wide DNase I footprinting data (from the ENCODE and Roadmap Epigenome projects). Transcription factors bind in regions where the chromatin is open and is more accessible to DNase I cleavage. To identify these sites, peak sites are analyzed for footprints of protection from DNase which is likely where the TF is binding. Looking at the DNA sequence at this position can provide the motif for the TF binding there (Figure 2B).

The availability of extensive functional annotations (Consortium, 2012; Pique-Regi et al., 2011; Hoffman et al., 2012) now enables the integration of functional genomic information into eQTL and GWAS analysis, which can be useful to dissect the causal variant and the functional basis of the observed associations. SNPs that fall within a transcription factor binding site represent a major mechanism underlying gene expression quantitative trait loci (eQTL) and GWAS variants are also much more likely to be expression quantitative trait loci (eQTNs) than SNPs not associated with any complex trait. (Nicolae et al 2010; Zhong et al. 2010). Linking genomic annotations to eQTLs and GWAS SNPs goes beyond genetic association analysis, and helps gain a better understanding of the underlying biological processes. Some of these issues have been discussed by previous works. For example, (Gaffney 2012, Veyrieras 2008, Lee 2009, Lappalainen 2013) have examined the enrichment of selected genomic annotations in cis-eQTLs, and (Pickrell 2014; Price 2012)
have found enrichment for functional categories in GWAS hits. In addition to predicting causal SNPs associated with GWAS traits within regulatory regions, it is possible to predict the allele specific binding (ASB) effect of any SNP within a motif, but this does not predict whether binding enhances or represses gene expression.

**High-throughput assays to identify functional regulatory variants**

There are a few methods that have been used to experimentally test for ASB. The two most common are QTL mapping and allele specific analysis. For QTL studies, large panels of individuals are genotyped, and the genotype is then matched to a quantitative trait (ie. open chromatin (dsQTL), gene expression (eQTL)). For higher throughput, allele-specific analysis can be used, where in a single individual you see the difference in allele count within the reads. However, these require heterozygous sites and in-depth sequencing to analyze.

Electrophoretic mobility shift assay (EMSA) can identify TF binding to any DNA sequence of interest. For this, the DNA sequence is combined *in-vitro* with the TF of interest and then run on a polyacrylamide gel. Where the protein binds, the complex travels more slowly through the gel, so the bound portion of DNA can be seen shifted upwards from the free or unbound DNA. The EMSA however, is more qualitative than quantitative, as well as being low throughput testing a single sequence at a time. This led to the development of the high throughput EMSA (Wong *et al.*, 2011; Stormo *et al.*, 2015; Levo *et al.*, 2015), where a pool of sequences are combined with the TF of interest, the bound and unbound portions of the gel are cut out, DNA extracted and sequenced to get a quantitative measure of the binding events (Figure 2). While BUNDLE-seq compared binding and reporter gene expression, and EMSA has been previously used to ascertain allelic effects, none of the
high-throughput EMSA methods have been previously used to determine allelic effects on binding.

Identifying ASB is necessary for assigning mechanism to expression and phenotype outcomes. However, gene regulation is highly complex, so binding changes do not always lead to a predicted outcome. Instead, effect on expression can be measured, and the gold standard for this is the traditional reporter assay. For this, the enhancer region containing either allele is cloned into a plasmid containing a minimal promoter and reporter gene (typically luciferase). This gets transfected into cells, and readout (luminescence) is a measure of the enhancer strength. However this method is both time and cost intensive to test each construct one at a time. So in the last few years, high throughput methods were developed to test the expression of thousands of oligos at once. There are two main methods for this, the massively parallel reporter assay (MPRA) and self transcribing active regulatory sequencing (STARR-seq). For MPRA (Melnikov et al., 2012; Kwasnieski et al., 2012; Patwardhan et al., 2012; Sharon et al., 2012; Kwasnieski et al., 2014), a multitude of unique synthesized DNA oligos containing a barcode at the 3′ UTR of a reporter plasmid are transfected into cells, and transcripts are isolated for RNA-seq. The number of barcode reads in the RNA over the number of barcode reads from the plasmid DNA is used as a quantitative measure of expression driven by the synthesized enhancer region. This method is fairly complex, requiring sequencing after cloning the enhancer/BC into the plasmid to match the sequences together and then another cloning and sequencing to place the reporter gene between the enhancer and BC. While the oligo being matched to multiple BCs helps with noise, it is imperfect since so many steps lie between the matching step and the final libraries. Studies have used this method to broadly characterize mutations within
enhancer and repressor regions (Melnikov et al., 2012; Kwasnieski et al., 2014) and fine-map eQTLs (Tewhey et al., 2016) or GWAS (Ulirsch et al., 2016).

STARR-seq (Arnold et al., 2013) methods involve fragmenting the genome and cloning the fragments 3′ of the reporter gene and sequencing the fragment directly instead of a BC proxy. The approach is based on the concept that enhancers can function independently of their relative positions and it's designed with putative enhancers placed downstream of a minimal promoter, such that active enhancers transcribe themselves, with their strength quantified as the amount of RNA transcripts within the cell. Because they do not use separate barcodes, STARR-seq approaches have streamlined protocols that allow for higher throughput. However, due to the lack of barcoding, these studies tend to be noisy.

In the original study, the D. melanogaster genome was fragmented and transfected into S2 cells (Arnold et al., 2013). An additional study involved treating transfected S2 cells containing the STARR-seq library with ecdysone (Shlyueva et al., 2014) (a well-studied hormone in insects) as many enhancers are controlled by external stimuli (Hurtado et al., 2011; Biddie et al., 2011; Hah et al., 2011, 2013; Shlyueva et al., 2014). A more recent adaptation to this method is Cap-STARR-seq (Vanhille et al., 2015). This method targeted specific regulatory regions whereby DNA fragments were captured on a custom-designed microarray. These methods were created to identify and validate regulatory regions, with no consideration of regulatory variants. However, placing the variant in biological context, for example by identifying the transcription factor whose binding is disrupted, often improves our understanding of disease mechanisms and opens further avenues for therapeutic intervention.
Analysis of the molecular function of non-coding variants across diverse risk environments

Aside from epidemiological studies, little attention has been paid so far in understanding how an individual’s genetic make-up interacts with his/her environmental exposures in defining disease risk. Accordingly, it is crucial to consider the role of genetic risk factors in the context of specific environments (e.g. vascular endothelium and stress for CVD-related traits). Little attention has focused on the analysis of how gene-environment interactions may affect an individual’s phenotype, because, in part, controlling for an individual’s lifetime environment is difficult. Cowper-Sal-lari et al. 2012 (Cowper-Sal et al., 2012) used functional genomics data to functionally annotate GWAS hits for breast cancer variants, by using functional genomic assays in the context of in-vitro estrogen treatments. The success of this study resides in considering the specific environmental context (estrogen treatment) relevant for the complex trait analyzed (breast cancer). These findings further support the necessity of considering the molecular function of each genetic risk variant in the context of the cellular and the organismal environment.

While it is extremely difficult to control for environmental exposure in the context of GWAS studies, they can be modeled under in-vitro settings with the cellular environment as a proxy for the organismal environment. This has been successfully done in eQTL studies looking at response to infection (Nédélec et al., 2016; Pacis et al., 2015) and drug treatment (Maranville et al., 2011; Mangrovite et al., 2013; Maranville et al., 2013), known as reQTLs. Looking at response to treatments with this method however is not viable on a large scale. Instead it is possible to use an allele-specific approach, where allelic effects are
modeled within a single individual at heterozygous sites. This method is also more controlled for confounding factors, giving it more power to detect conditional ASE (cASE) than is possible in reQTL studies, particularly in small sample sizes. For even more controlled conditions, exact sequences can be studied in the case of high throughput reporter assays, as was done in S2 cells with ecdysone (a well-studied hormone in insects) (Shlyueva et al., 2014) and A549 cells with dexamethasone (synthetic glucocorticoid) (Vockley et al., 2016).
CHAPTER 2 Computational predictions and analysis of non-coding variants for allele-specific expression

Introduction

Genetic variants in non-coding regions are responsible for inter-individual differences in molecular and complex phenotypes. Quantitative trait loci (QTLs) for molecular and cellular phenotypes (Dermitzakis, 2012) have been crucial in providing stronger evidence and a better understanding of how genetic variants in regulatory sequences can affect gene expression levels (Stranger, 2007; Gibbs et al., 2010; Melzer et al., 2008; Cheung et al., 2003; Brem et al., 2002). However, eQTL studies have severe limitations in identifying the true causal variant, due to linkage disequilibrium (LD) limiting the resolution of analysis. The availability of extensive functional annotations (Consortium, 2012; Pique-Regi et al., 2011; Hoffman et al., 2012; Moyerbrailean et al., 2016b) enables the integration of functional genomic information into eQTL analysis, which can be useful to dissect the causal variant and the functional basis of the observed associations (Gaffney et al., 2012; Veyrieras et al., 2008; Lee et al., 2009; Lappalainen et al., 2013; Kichaev et al., 2014; Wen et al., 2015; Pickrell, 2014). SNPs that fall within a transcription factor (TF) binding site (TFBS) represent a major mechanism underlying eQTLs (Degner et al., 2012). Recently, additional computational and experimental techniques have been developed to predict and detect allelic effects of SNPs in TFBS (CentiSNPs) using DNase I footprinting and ChIP-seq data (from the ENCODE and Roadmap Epigenome projects) (Moyerbrailean et al., 2016b; Lee et al., 2015a; Maurano et al., 2015; Zhou and Troyanskaya, 2015). Still, it is a challenge
to further validate if allelic effects in binding translate to effects on gene transcription.

Here we used the CentiSNPs annotation (variants predicted to affect TFBS) generated from previous work in done in the lab (Moyerbrailean et al., 2016b) in combination with a software called fgwas (Pickrell, 2014) to finemap variants associated with GWAS. The rich meta information provided by the tissue-specificity and the identity of the putative TF binding site being affected also helps in identifying the underlying mechanism supporting the association. While all these existing computational annotations are useful for predicting the causal SNP in an eQTL, they do not prove the SNP is truly causal, nor do they properly quantify its effect on gene expression. This can be validated experimentally using a traditional reporter assay, where the enhancer region is cloned into a plasmid containing a minimal promoter and reporter gene (ie. Luciferase. This gets transfected into cells, then the reporter signal is used as a correlate of the enhancer activity. Here we tested a subset (21) of the putatively causal GWAS hits identified using fgwas with the CentiSNP annotation for significant allele specific gene expression.

Traditional reporter assays, while the gold standard for the field, are both time and cost intensive. Instead, to dissect regulatory sequences and compare genetic effects on gene expression, different versions of high throughput reporter assays have emerged in the recent years. These include massively parallel reporter assays (MPRA) (Melnikov et al., 2012; Kwasnieski et al., 2012) and self transcribing active regulatory regions sequencing (STARR-seq) (Arnold et al., 2013) that can simultaneously measure the regulatory function of thousands of constructs at once. MPRAs utilize a multitude of unique synthesized DNA oligos that are associated with barcodes, cloned in a reporter plasmid and transfected into cells. The transcripts are then isolated for RNA-seq. The number of barcode reads in the
RNA over the number of barcode reads from the plasmid DNA is used as a quantitative measure of expression driven by the synthetic enhancer region (Melnikov et al., 2012; Kwasnieski et al., 2012; Patwardhan et al., 2012; Sharon et al., 2012; Kwasnieski et al., 2014). MPRA and STARR-seq were originally developed to identify and validate regulatory regions, but they can also be used to compare allelic effects of genetic polymorphisms or methylation (Lea et al., 2017). Recent studies used this technique to compare allelic variants of SNPs with the aim to dissect, at a large scale, the causal nucleotide in eQTL and Genome Wide Association Study (GWAS) signals. Specifically, Vockley et al. (2015) used a STARR-seq derived method (POP-STARR-seq) to measure allelic effects on gene expression for population based variation in 104 regulatory regions, and a more recent study by Tewhey et al. (2016) adapted MPRA to fine-map variants associated with gene expression in lymphoblastoid cell lines (LCLs) and HepG2.

The application of MPRA to quantify the allelic effects of regulatory variants is very similar to the challenge posed by allele-specific expression (ASE) in RNA-seq data. However, one key difference is that the proportion of plasmids for each allelic construct may not be in a 1:1 ratio. Few off-the-shelf statistical methods have been used for processing and analyzing these large MPRA datasets (Table 1), but they do not consider several technical issues that can lead to false positives, such as base-calling error and over-dispersion. As demonstrated in RNA-seq ASE approaches, a binomial distribution fails to account for overdispersion and results in overly optimistic p-values, while a beta-binomial distribution is a more adequate choice (Kumasaka et al., 2016; van de Geijn et al., 2015). Compared to RNA-seq ASE methods that combine all reads across haplotypes, in MPRA we do not need to accommodate for the uncertainty in phasing or haplotyping as the complete se-
Table 1: Statistical methods for ASE and MPRA analysis.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Type of test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
</tr>
<tr>
<td>Previously used in MPRA</td>
<td>X</td>
</tr>
<tr>
<td>Previously used for ASE</td>
<td>X</td>
</tr>
<tr>
<td>Requires normally distributed data</td>
<td>X</td>
</tr>
<tr>
<td>Underestimates the effect of biological variability</td>
<td>X</td>
</tr>
<tr>
<td>Handles overdispersion</td>
<td>X</td>
</tr>
<tr>
<td>Accounts for base calling error</td>
<td>X</td>
</tr>
</tbody>
</table>

sequence of each construct is known. By design we can also avoid oligonucleotides that could lead to ambiguous mapping. This is in stark contrast to using the entire human genome/transcriptome, which typically requires extensive pre-processing. This is because many genomes contain large number of repetitive and quasi-repetitive regions that are only one SNP or base-calling error away from many other paralogous regions. Here we further extend QuASAR (Harvey et al., 2014), an approach which considers both overdispersion and base-calling errors, to test for allelic imbalance in MPRA constructs when the default proportions are not equal. The new method allows for estimates of the dispersion parameter depending on variant-specific read coverage, and produces summary statistics that are easy to incorporate in downstream analyses.

Here we tested our new method on MPRA data from Tewhey et al. (2016) and we further confirmed the robustness of our method on another dataset from Ulirsch et al. (2016). First we compared our new QuASAR-MPRA statistical test to other tests employed in MPRA and ASE analyses (Table 1). We then demonstrate that the QuASAR-MPRA test better calibrates the $p$-values under the null hypothesis, without sacrificing statistical power. Finally, we used the allelic effects identified by QuASAR-MPRA to investigate whether the genetic variants that fall within genomic annotations, such as TF binding motifs, are good predictors for allele-specific regulatory function. Our study shows the potential value of using
robust allele-specific analysis in high throughput reporter assays, to improve fine mapping analysis of association signals and validate genomic annotations of regulatory variants.

**Methods**

**Integrating high-resolution functional annotations with GWAS and fine-mapping**

To integrate functional annotations and GWAS results, we used the fgwas command line tool (Pickrell, 2014). fgwas computes association statistics genome wide using all common SNPs from European populations in the 1KG Project, splitting the genome into blocks larger than LD. Summary statistics were imputed with ImpG using $Z$-scores from meta-analysis data. Using an empirical Bayesian framework implemented in the fgwas software, GWAS data were then combined with functional annotations. We then compared the informativeness of these annotations from each of the 1891 motifs with Centipede predicted regulatory sites to a baseline model (see section: Adding annotations to SNPs associated with complex traits in Appendix A) consisting of previously used genomic annotations identified as relevant (Pickrell, 2014). For each locus that contains at least one SNP with a PPA $> 0.2$, we only consider the SNP with the highest $p$-value or PPA from fgwas. Rather than look at a credible set, we pick a single SNP most likely to be causal and see if that SNP has a higher PPA with the annotation than without it. While reduction in size of the credible set is very important for assessing fine-mapping methodologies, here our focus is on combining annotations to identify the single most likely causal SNP per GWAS locus.
Validation of GWAS-relevant effect-SNPs

GWAS-relevant effect-SNPs located in active footprints in LCLs (the cell line used for transfection) were ranked on the Spearman correlation coefficient in Moyerbrailean et al. (2016a). We initially selected the top 25 SNPs with a positive correlation, but the assays for 4 of them failed for several technical reasons (e.g., cloning step failed). To test allele-specific effects on expression for the remaining 21 SNPs, we first constructed inserts containing the reference or alternate allele for each SNP of interest (see section: Validating putative causal SNPs by reporter gene assays in Appendix A). Cloning of these inserts in the pGL4.23 vector was performed using the Infusion Cloning HD kit (Clontech) and DNA was extracted using the PureYield kit (Promega). Transfections were performed into GM18507 using the standard protocol for the Nucleofector electroporation (Lonza). Luciferase activity was measured for up to 20 replicate experiments using the Dual-Glo Luciferase Assay Kit (Promega). We contrasted the activity of each construct to the pGL4.23 vector, to assess enhancer/repressor activity of each region. To evaluate allele-specific effects, we contrasted the activity of the reference allele to the alternate allele for each region and we used a t-test to assess significance at a $p < 0.05$ threshold. We used the Benjamini-Hochberg (Benjamini and Hochberg, 1995) procedure to assess FDR across all 21 SNPs tested.

Data source and pre-processing

We downloaded processed read counts from GEO (GSE75661) ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE75nnn/GSE75661/suppl/GSE75661_79kCollapsed_counts.txt.gz (Tewhey et al., 2016). This MPRA study was designed to look at ASE in 39,479 oligo
pairs representing 3,642 eQTLs from the GEUVADIS RNA-seq dataset of lymphoblastoid cell lines (LCLs) from European and African individuals (Lappalainen et al., 2013). It has a large number of experimental replicates (8 LCL replicates), and makes use of barcodes (an average of 73 unique barcodes per oligo per replicate) to remove PCR duplicates, making this an ideal dataset to work with. We considered separately sequences in the forward and reverse strand direction in the library, as direction of the regulatory region could potentially affect reporter gene and therefore barcode expression. Tewhey et al. (2016) found that filtering the data to remove variants with low coverage greatly reduced the variability between replicates. Higher variance could then lead to falsely identifying ASE. We therefore began processing the dataset by applying a counts filter. For each direction we removed all cases with less than five reads on the reference and alternate allele, and where the sum of two alleles was \( \leq 100 \). This gave us a total of 33,664 SNPs in the DNA library as input to the RNA library.

For the RNA library, we first separated the library into forward and reverse directions, and then required that RNA constructs were in the DNA library. We used a counts filter of 5 for both reference and alternate alleles so that we were only looking at variants that had sufficient reads covering both alleles to test for allele-specific effects on expression. This left us with 19,287 SNPs in the forward library and 19,748 SNPs in the reverse library or 33,653 SNPs total represented.

We additionally applied the QuASAR-MPRA method to a separate dataset by Ulirsch et al. (2016). We downloaded processed read counts from http://www.bloodgenes.org/RBC_MPRA/. This dataset comprised of 2,756 variants in strong linkage disequilibrium with 75 sentinel variants associated with red blood cell traits, with reference and alternate alle-
les represented in the pool of constructs. Each variant has 3 sliding windows of coverage, which we treated as separate constructs (rather than combining counts per variant). This dataset comprised of 2 DNA and 6 RNA replicates (from K562 cells). The data was processed using the same steps as with the Tewhey et al. (2016) data, resulting in 2,669 SNPs in total.

**Baseline statistical methods for comparison**

To test for ASE there are several different methods available (Table 1). The \( t \)-test, Fisher’s exact test and binomial test are classical tests remarkably appealing due to their simplicity. However, they have several limitations, as they cannot be tuned to the context of the experiment, such as levels of overdispersion (eg. from biological and technical variability) which are known to exist in ASE data (Castel et al., 2015; Skelly et al., 2011; Anders et al., 2010). A paired Student’s \( t \)-test for ASE can be used to test whether the mean expression of the reference allele is equal to the mean expression of the alternate allele. This test requires multiple replicates in order to calculate a mean for each allelic expression group that has little variance, otherwise the test will not have the power to detect differences. Fisher’s exact test has been used previously to identify ASE (Romanel et al., 2015), by testing whether the reference and alternate allele counts’ proportions are the same. Rejection of the null hypothesis, however, only informs us that the difference between the average counts in the two samples is larger than one would expect between technical replicates. In the binomial test, the null hypothesis is that observed values for two categories do not deviate from the theoretically expected distribution of observations. In ASE, the binomial test is used to determine whether the ratio of the two alleles is significantly different from the expected proportion (e.g. 0.5). This is the classic test that
has been employed previously to detect ASE in RNA-seq studies, and assumes that read
counts within each gene are binomially distributed (Kilpinen et al., 2013; Consortium et al.,
2015; Lappalainen et al., 2013; Buil et al., 2014). Even accounting for reference mapping
bias in RNA-seq reads, p-values have been found to remain inflated (Castel et al., 2015).
Other methods handling ASE such as WASP, RASQUAL, EAGLE (Kumasaka et al., 2016;
van de Geijn et al., 2014; Knowles et al., 2017) use a per SNP overdispersion parameter
and give well calibrated p-values. However these methods perform ASE QTL mapping and
their application to MPRA would require a large number of replicates (> 15).

To reproduce the Student’s t-test performed by Tewhey et al. (2016), we calculated the
log2 ratio for the reference and alternate allele constructs (RNA/DNA) for each replicate.
These values were used as input for a paired t-test in R. To perform the Fisher’s exact test
on the MPRA counts data, we first added a pseudocount of 1 (Vockley et al., 2015) to
each RNA and DNA reference and alternate allele counts and then used the fisher.test
function in R. To perform the binomial test on the MPRA counts data, we compared the
reference and alternate allele counts to the DNA proportion (reference allele/ reference
allele + alternate allele). To combine the p-values for the two LCL individuals, we used
Fisher’s method (Tewhey et al., 2016).

QuASAR Approach

QuASAR by default assumes that under the null hypothesis of no allelic imbalance the
reference and alternate allele read counts should be at 1:1 ratio. However, in MPRA, the
proportion $r_1$ of the reference reads is not necessarily 0.5 across all the $l$ genetic variants,
due to differences in PCR amplification, as well as cloning and transformation efficiencies.
Here, we have extended QuASAR to test for differences between the proportion of refer-
ence reads in DNA \( r_l \) and the proportion obtained from RNA reads \( \rho_l \). To reject the null hypothesis \( \rho_l = r_l \), we extend QuASAR’s beta-binomial model. The observed reference \( R_l \) and alternate \( A_l \) allele read counts at a given \( l \) are modeled as:

$$
\Pr (R_l|N_l, \psi_l, M_b) = \frac{(N_l)^{M_b} \Gamma (R_l + \psi_l M_b) \Gamma (A_l + (1 - \psi_l) M_b)}{(R_l)^{N_l + M_b} \Gamma (N_l + M_b) \Gamma (\psi_l M_b) \Gamma ((1 - \psi_l) M_b)}
$$

(0.1)

$$
\psi_l = [\rho_l (1 - \epsilon) + (1 - \rho_l) \epsilon]
$$

(0.2)

where \( N_l = R_l + A_l \) is the total read count at \( l \), and \( M_b \) is the concentration parameter that controls over-dispersion of the mean proportion centered around \( \psi_l \), which also incorporates in the model a base-calling error \( \epsilon \) and the allelic ratio \( \rho_l \) overall-mean. We can estimate \( \epsilon \) using an EM procedure (Harvey et al., 2014), but here for MPRA we fixed \( \hat{\epsilon} = 0.001 \) as a conservative estimate of the true base-calling error rate.

We have found previously for ASE that overdispersion decreases with greater depth of coverage (Figure S9 in Moyerbrailean et al. (2016a)). Therefore here, as compared to our previous implementation of QuASAR, we use different \( M_b \) parameters depending on the sequencing depth \( N_l \). We bin \( N_l \) into different quantiles (here deciles) and we estimate \( M_b \) for each bin separately using a grid search:

$$
\hat{M}_b = \arg \max_{M_b} \left( \prod_{l=1}^{L} \Pr (R_l|N_l, \hat{\epsilon}, \rho_l = r_l, M_b) \right)
$$

(0.3)

This should work well when the number of sites (i.e., SNPs tested) is relatively large so each bin \( b \) has > 200 observations to estimate \( M_b \). In our experience sequencing depth
is a major determinant for M, and because we estimate M under the null, we tend to be conservative (i.e., M is the worst case scenario for all the constructs that belong to the same group). As a consequence, the QuASAR-MPRA p-values remain well calibrated (or in the worst case scenario they will tend to be slightly conservative).

We estimate \( \hat{\rho}_l \) using (0.1) with \( M_b = \hat{M}_b \) from (0.3) and a standard gradient method (L-BFGS-B) to maximize the log-likelihood function:

\[
l(\rho_l; \hat{M}_b, \hat{\epsilon}) = \log \Pr \left( R_l | N_l, \psi_l = \psi(\rho_l, \hat{\epsilon}_s), \hat{M}_b \right) \quad (0.4)
\]

Finally, all parameters are used to calculate the LRT statistic, contrasting \( H_1 : \rho_l = \hat{\rho}_l \) to \( H_0 : \rho_l = r_l \) and the resulting p-value.

For comparison, we performed the original QuASAR analysis on the data as well, as described in Harvey et al. (2014).

**QuASAR meta-analysis**

Using the QuASAR approach, we can generate summary statistics of the allelic imbalance that can be used for downstream analyses. For example, to compare DNA to RNA, or between RNA of different cell-types, or to perform meta-analysis of multiple MPRA libraries. Instead of using an estimate of the allelic proportion \( \rho_l \), in the QuASAR approach we report the estimate of \( \beta_l = \log(\rho_l/(1 - \rho_l)) \) and its standard error \( \hat{\sigma}_l \) using the second derivative (i.e. Hessian) of the log-likelihood function in (0.4). We prefer the logistic transformed parameter \( \beta_l \) as it provides a more robust fit and the second derivative is better behaved than that of \( \rho_l \) on the edges.

To illustrate this for the Tewhey et al. (2016) data, we combined the summary statistics
for the two LCL individuals using standard fixed effects meta-analysis. The effect size $\beta_{l,n}$ of each replicate $n$ is weighted by $w_{n,l} = 1/\hat{\sigma}_{n,l}^2$, to calculate the overall effect size and standard error:

$$
\beta^*_l = \frac{1}{w^*_l} \sum_n \beta_{n,l} w_{n,l} \quad \sigma^*_l = \sqrt{1/w^*_l} \quad (0.5)
$$

where $w^*_l = \sum_n w_{n,l}$. We can then calculate the $Z$-score and $p$-value to test for an overall change between all the RNA replicates combined with respect to the original DNA proportion $\beta_0$:

$$
Z_l = \frac{\beta^*_l - \beta_0}{\sigma^*_l}, \quad \beta_0 = \log \frac{r_l}{1-r_l}, \quad p = 2\Phi(-|Z_l|) \quad (0.6)
$$

Across all the paper, $p$-values were corrected for multiple testing using the Benjamini-Hochberg’s (BH) method (Benjamini and Hochberg, 1995). To compare the different approaches we quantified the genomic inflation parameter, $\lambda$, for a set of $p$-values (Yang et al., 2011). For this we calculated the ratio of the median of the $p$-value distribution to the expected median, thus quantifying the extent of the bulk inflation and the excess false positive rate. We also use a rank sum paired test to assess statistical significance in the $p$-value inflation between QuASAR-MPRA and other methods with similar performance.

**Simulations**

To simulate MPRA data we randomly sampled from a beta-binomial distribution with parameters set to approximate the real data in Tewhey et al. (2016). The advantage of a simulation is that we have full knowledge of which SNPs are truly imbalanced and we can
empirically calculate statistical power (i.e, sensitivity) and FDR under specific assumptions. The true underlying distribution may not exactly be beta-binomial but simulations are still very useful to know how the test performs and compares to other tests. We started by simulating the DNA reads and proportions for each SNP using a beta-binomial. For this we used the DNA proportion from the Tewhey et al. (2016) data as the expected proportion \( \rho_l \) and we set the concentration parameter \( M \) to be 200, and the total number of DNA reads \( N = 10,000 \).

\[
\theta_l \sim \text{Beta}(\rho_l M, (1 - \rho_l) M)) \quad R_{i}^{\text{DNA}} \sim \text{Bin}(N, \theta_l)
\]

After we simulated the DNA counts \( R_{i}^{\text{DNA}} \), we recalculated the new DNA proportions \( r_i^* = R_{i}^*/N \). The exact value of the parameters used to generate the DNA counts are not very important and should have no effect for the simulation as the differences are captured on the RNA data once the DNA proportion is specified.

To simulate the RNA data we need to simulate two conditions: 1) SNPs without ASE and the same proportion as in DNA, and 2) SNPs with ASE and a different allelic proportion than those in DNA. To do this, we explored different parameter settings for the concentration parameter \( M \) (10, 60 and 100), effect size \( \Delta \beta \) (0.5, 1.0 and 2.0) and number of RNA replicates (2, 5 and 8). The number of reads \( N_i \) observed for each SNP was set up to match the average NA12878 RNA counts for each SNP (so multiple coverages are being simulated) and we divided these by a constant factor to simulate sequencing depths (1/2, 1/5 and 1/10) lower than those obtained by Tewhey et al. (2016). Each RNA
replicate was simulated with the same parameter setting.

\[ \rho^*_l = \logit(\logit^{-1}(\tau^*_l) + \Delta \beta) \]  \hspace{1cm} (0.8)

\[ \theta_l \sim \text{Beta}(\rho^*_l M, (1 - \rho^*_l)M) \]  \hspace{1cm} \[ R^{RNA}_l \sim \text{Bin}(N_l, \theta_l) \]  \hspace{1cm} (0.9)

To sample from Beta and Bin, we used the rbeta and rbinom in R respectively. The proportion of SNPs with RNA counts with \( \Delta \beta \neq 0 \) and simulated to have an allelic imbalance is 0.1% of the total SNPs. For each simulated data set we then ran the \( t \)-test and QuASAR and adjusted the \( p \)-values for multiple testing using the same BH procedure as in the real data. For each FDR control threshold we empirically calculated power (Sensitivity) and false discovery rate (eFDR). To ensure that we get robust sensitivity and FDR estimates we repeated the entire procedure 20 times and reported the average.

**Annotation Overlap**

Table 9 in Appendix A reports the annotations we have considered with their sources. More specifically, we considered two major sets of annotations: experimentally and computationally derived. The experimental annotations include allele-specific hypersensitivity (ASH) from (Moyerbrailean et al., 2016b), dsQTLs (Degner et al., 2012), and GTEx eQTLs (Consortium et al., 2015).

In terms of computational annotations, a variety of different methods have been used recently to predict the allelic effect of SNP on TF binding and chromatin accessibility. GKM-svm (Lee et al., 2015a) uses gapped k-mer frequencies to predict the activity of larger functional genomic sequence elements, including the impact of a variant on DNase I sensitivity. It utilizes support vector machinery based on the structural risk minimization principle.
from statistical learning theory and kernel function which calculates the similarity between any two sequences. CATO (Maurano et al., 2015) quantifies the effect of SNPs on the energy of TF binding, through overlapping SNP DHS profiles with TF motifs and applying a logistic model which takes into account site dependent features and phylogenetic conservation. DeepSEA (Zhou and Troyanskaya, 2015) uses TF binding, DHS, and histone-mark profiles with genomic sequence information as input for training a deep learning-based algorithm and predict the effects that sequence alterations have on the chromatin. DeepSEA has three major features: integrating sequence information from a wide sequence context, learning sequence code at multiple spatial scales with a hierarchical architecture, and multitask joint learning of diverse chromatin factors sharing predictive features. Finally we also used CentiSNPs, an annotation that we recently developed (Moyerbrailean et al., 2016b) that uses the Centipede framework (Pique-Regi et al., 2011) to integrate DNase-seq footprints with a recalibrated position weight matrix (PWM) model for the sequence to predict the functional impact of SNPs in footprints. In CentiSNPs, SNPs in footprints “footprint-SNPs” are further categorized using Centipede hierarchical prior for each allele as “effect-SNP” if the prior relative odds for binding are > 20 or as “Non-effect-SNPs” otherwise.

For the other computational annotations we set the following thresholds. To run GKM-svm (Lee et al., 2015a), we extracted sequences around MPRA variants (19bp total) and then ran the reference vs alternate allele sequences with either the GM12878 or HepG2 weights. We then used a threshold of $<-6$ or $>6$ for the variant scores. DeepSEA (Zhou and Troyanskaya, 2015) variant scores were identified using the website tool with a vcf file input (containing the MPRA variants). The functional significance predictions have a
threshold of < 0.05. We overlapped SNPs from MPRA counts data with each annotation type. To identify particular annotations that predict the ASE found in the MPRA, we built logistic models \( \log(p_l/(1 - p_l)) = \beta_0 + \beta_1 \times a_l \) using the QuASAR \( p \)-values \( (p < 0.001) \) as the observed binary outcome, and the genomic annotations \( a_l \) as the predictor. For this type of analysis we use the nominal \( p \)-value \( (p < 0.001) \), as we test for an enrichment with respect to what would be expected from the null uniform distribution \( (0.1\% \text{ of the tests}) \). This nominal \( p \)-value corresponds to a FDR threshold of 7.2\% for FDR (enrichments are not sensitive to variations of this threshold). A significant \( p \)-value from the annotation logistic model together with the QQ-plot are useful to evaluate which annotations work best in predicting changes in gene regulation.

**Results**

**Functional regulatory variants help identify and interpret causal GWAS hits**

To better test if the annotated CentiSNPs can help fine-mapping and give a mechanistic support for variants associated with complex traits, we integrated them into GWAS meta analyses for 18 traits (see **Summary of GWAS meta analysis traits examined**.) using the recently developed hierarchical model fgwas (Pickrell, 2014). Importantly, in this analysis we used as input the association \( p \)-values measured or imputed to all known common variants in the genome. Furthermore, for each trait we compare to a baseline model (Pickrell, 2014) that considers previously defined annotations (Maurano et al., 2012; Thurman et al., 2012) and confounders (e.g., distance to TSS, coding region, and others). For each trait, we identified factors whose binding sites were enriched for associated SNPs (Figure 3A & B, 18 and Table 6) over the baseline model (the enrichments reported by fgwas are
log-odds ratios from the model parameters).

Overall, we observed high enrichments for biologically relevant factors. For example, the enrichment for CentiSNPs in OCT-4 (POU5F1, a TF with a key role in embryonic development and stem cell pluripotency (Nichols et al., 1998)) regulatory sequences when considering genetic variants associated with human height is 6.6-fold higher (95%CI: 3.7-8.2) than in the baseline model. This is consistent with previous observations of genetic variants associated with height being enriched in embryonic stem cell DHS sites (Trynka et al., 2015). We also observed an enrichment for the developmental regulators TBX15 (3.9x), FOXD3 (3.9x), and NKX2-5 (4.7x) for genetic variants associated with height. From a study of low-density lipoprotein (LDL) levels in the blood, enriched factors include the liver-specific factor HNF4A (9.1x), as well as several regulators of immune function, including CREB1 (3.7x), IRF1 (6.2x), and IRF2 (7.1x).
Our high resolution annotations allowed us to dissect the most likely functional variant (posterior probability of association, PPA > 0.2) in 88 previously identified GWAS regions (Table 7, Figure 20). For all 88 but 2 of these SNPs we have at least a 2-fold increase on the posterior odds of picking the potentially causal genetic variant according to fgwas (8.5x median fold increase) when compared to the comprehensive baseline annotation.
used by (Pickrell, 2014). We then performed reporter gene assays for 21 SNPs to validate the predicted allelic effect on gene expression and the underlying molecular mechanism (Figure 4 A & B, Table 8, Methods). Among the regions tested we validated that 11 have enhancer/repressor activity and 10 have variants with allele-specific activity ($p < 0.05$, BH-FDR=10%). This corresponds to 48% validation rate which is much greater than the 5% that would be expected by chance (Binomial test $p = 2.01 \times 10^{-8}$). Overall the predicted effect on binding and the change in gene expression are well correlated (Spearman $\rho=0.612$, $p$-value=0.0032), and the three SNPs with opposite effects may represent binding sites for repressors. Spearman correlation is robust to outliers, removing potential outlier rs540909 results in $\rho=0.657$ ($p$-value = 0.002).

As an example, rs4519508, associated with a 2.1cm decrease in height (Lango Allen et al., 2010), is in a binding site for the cell-cycle regulator family E2F (Figure 3D). Our annotation increased the PPA from a baseline of 10.5% to 44.4%, and it is the highest associated SNP in the association block (Figure 19A). This E2F footprint is active in >300 tissues (most of them fetal) and we detected ASH at this SNP in lung fibroblasts, validating that the reference allele at rs4519508 confers stronger binding than the alternate. Interestingly, in the reporter assay we observed 1.5-fold increased expression in the presence of the alternate allele, suggesting that at this location, E2F is acting as a repressor. Finally, this SNP is located within the promoter of PPP3R1, a regulatory subunit of calcineurin important for cardiac and skeletal muscle phenotypes; and a SNP in the same region has been shown to be associated with endurance (He et al., 2010) in humans. The $p$-value of association for this GWAS locus ($p = 8.1 \times 10^{-6}$) does not reach genome-wide significance in the height meta-analysis data we used (Lango Allen et al., 2010); however, in
a recent more extensive meta-analysis for height (Wood et al., 2014) this locus achieves genome-wide significance $p = 8.4 \times 10^{-10}$, demonstrating that our annotation can be useful to rescue relevant loci.

![Figure 4: Reporter gene assay validation of allelic regulatory activity.](image)

Finally, a SNP associated with LDL levels, rs532436, is within a footprint for USF, an E-box motif (Figure 3C). Adding our annotation increased the PPA of the SNP from 39.7% to 94.7% (Figure 19B). We found that the alternate allele, associated with a 0.0785 mg/dL increase of LDL in the blood, is predicted to have a lower binding probability and results
in 1.8-fold lower expression, compared to the reference allele. This SNP is identified by GTEx (Consortium et al., 2015) as an eQTL for two proximal genes in whole blood: ABO ($p = 5 \times 10^{-5}$) and SLC2A6 (GLUT6, a class III glucose transport protein; $p = 8 \times 10^{-5}$). The SNP has an opposite effect on expression of the two genes, with the alternate allele showing lower expression for ABO and higher expression for SLC2A6.

These results show that our integrated analysis provides support for likely mechanisms linking regulatory sequence changes to complex organismal phenotypes. Furthermore, these mechanisms can be directly investigated through molecular studies, providing additional support that these sequence changes are truly functional.

**Applying QuASAR-MPRA to identify ASE**

We used the method proposed here, QuASAR-MPRA, to detect ASE in the MPRA data collected by (Tewhey et al., 2016). In MPRA, ASE is defined as the departure in the RNA reads from the DNA proportion (the input allelic ratio). Because strand orientation may affect the enhancer function of the sequences tested, each SNP was tested for ASE in the two strand orientations separately (forward/reverse). The two LCL biological replicates were combined using meta-analysis (see Methods). The number of SNPs with significant ASE (10% FDR) were 309 (forward) and 293 (reverse) in LCLs (Table 10 and Figure 5), 85 (forward) and 84 (reverse) in HepG2 (Table 11 and Figure 21). We then compared these results to those obtained using other methods previously used for MPRA/ASE analysis using the same input file with the same pre-processing filters (see Methods).
While some of the other methods seem to identify a larger number of SNPs with significant ASE, the distribution of $p$-values (Figure 5) shows that those methods have very skewed distributions. The majority of genetic variants tested are expected to have no impact and only those that were the truly causal eQTL SNP should have a significant $p$-value.

We do not know a priori which variants have ASE, but in Figure 5 we would expect that the majority of $p$-values would follow the expected uniform distribution if the approach correctly models the data under the null hypothesis. In other words, only a fraction of MPRA constructs are expected to have significant allelic effects. To better quantify the departure from the expected distribution of $p$-values for each testing method we used the genomic inflation method. In this method, a greater departure from a lambda value of 1 corresponds to greater inflation in the test results (see Appendix A for reverse oligo results). Based on the genomic inflation value $\lambda$, QuASAR-MPRA results in the lowest inflation, with $\lambda = 1.161$. A paired $t$-test with independent estimation of variance and Welch’s adjustment, as in (Tewhey et al., 2016), results in a moderately but significantly
larger $\lambda = 2.89 \ (p < 2.2 \times 10^{-16})$. The binomial test produces the greatest inflation, with $\lambda = 57.95$, followed by Fisher’s exact test, as in (Vockley et al., 2015) resulting in $\lambda = 38.68$. The methods with the lowest inflation, QuASAR-MPRA and the $t$-test, have a 69% match at 10% FDR.

These results are also similar if we use a different dataset (Ulirsch et al., 2016) (Figure 6). QuASAR-MPRA results in the lowest inflation, with $\lambda = 0.58$, while the binomial test produces the greatest inflation, with $\lambda = 33.31$ followed by Fisher’s exact test $\lambda = 16.86$. The paired $t$-test is relatively well calibrated $\lambda = 1.74$ but detects less hits than QuASAR-MPRA (only 64 constructs containing 53 variants at FDR 10%). Using QuASAR-MPRA we were able to identify 103 constructs containing 95 variants (FDR 10%) with significant ASE.

![Figure 6: Comparing ASE testing methods in Ulirsch et al. (2016). QQplot depicting the p-value distributions from testing for ASE using four different methods in K562 for all SNPs. $\lambda$ measures genomic inflation deviation from the uniform.](image)

Alternatively, we also considered the $p$-value distributions only for the SNPs not predicted to affect TF binding (non-effect SNPs), as these SNPs are more likely to be true
negatives 5. Note that our computationally predicted effects are not a perfect gold standard and in fact one major application of this type of data and its analysis is to precisely validate the accuracy of these computational annotations and predictions as we will show later. Nevertheless, we see (in Figure 5, 6 and 21) that the two methods with lowest lambda values show an even lower departure from the null, consistent with the computational method correctly predicting a large number of true positives.

**Applying QuASAR-MPRA to simulated data**

To further investigate our proposed new method we used simulated data where we know exactly the underlying true ASE signal to evaluate the detection accuracy. It is important to note that the simulation conditions may not exactly match those from the real data (see Methods) but they are very useful for getting more insights about the scenarios that may have larger impact on performance. Here we only compare the two methods that seem to be well-calibrated under the null hypothesis QuASAR-MPRA and the \( t \)-test. Under the null distribution for all our simulations both tests do not show a significant departure from the expected uniform distribution for the \( p \)-values.

We then compared results from QuASAR-MPRA and the \( t \)-test in scenarios when a fraction of the tests do have ASE (see Methods). In every condition QuASAR-MPRA has greater sensitivity to detect ASE than the \( t \)-test (Figure 7). The \( t \)-test seems to perform better when the over-dispersion is low \( (M=100) \), or when the effect size of ASE is high \( (\Delta \beta=2) \). QuASAR-MPRA also handles well low coverage data and a small number of replicates to achieve good statistical power (Figure 7). This is consistent with our original findings with QuASAR (Harvey et al., 2014) demonstrating that we can measure ASE in a small number of replicates if there is enough read coverage. The \( t \)-test appears to require a
large number of replicates in order to have power to detect ASE as compared to QuASAR-MPRA.

Figure 7: Exploring the performance across multiple simulated conditions. Plots depicting empirical power (sensitivity, Left y-axis) and empirical FDR (eFDR, Right y-axis) achieved at different BH-FDR control levels (x-axis) for ASE testing using QuASAR-MPRA (green) and a t-test (blue) across multiple simulated conditions (rows). Default conditions are M=60, Δβ=1, 5 replicates, and reads/5. Each row explores changing different simulation settings: A) over-dispersion high (M=10), medium (M=60) and low (M=100); B) effect-size high (Δβ=2), medium (Δβ=1) and low (Δβ=0.5); C) number of replicates (3, 5, or 8) D) overall sequencing depth compared to (Tewhey et al., 2016) (1/10, 1/5, or 1/2).

Validation of experimental and computational annotations for functional non-coding variants

High-throughput reporter assays can be used not only to fine-map causal variants in both GWAS and eQTL studies, but also to validate SNP functional annotations (Kwas-
cieski et al., 2014). Here we take advantage that the \( p \)-values derived from QuASAR are well calibrated under the null hypothesis to examine enrichments for low \( p \)-values in both experimentally and computationally derived annotations for allele-specific effects on TF binding. The experimentally derived annotations included LCL dsQTLs (Degner et al., 2012), allele-specific hypersensitivity (ASH) SNPs (Moyerbrailean et al., 2016b), and GTEx eQTLs (Consortium et al., 2015). In both LCLs (Figure 8) and HepG2 (Figure 22), ASH SNPs had the greatest departure from the null, followed by LCL dsQTLs.

![Figure 8: Validating experimental annotations in LCLs. QQ plot depicting the \( p \)-value distributions from testing for ASE using QuASAR, overlapping with experimental genomic annotations including allele-specific hypersensitivity (ASH) (Moyerbrailean et al., 2016b), DNase I sensitivity QTLs (dsQTLs) (Degner et al., 2012) and GTEx (Genotype-Tissue Expression) lead SNP in LCLs (Consortium et al., 2015). An annotation enrichment \( p \)-value is reported next to their labels, but only for those annotations that are significantly enriched for small QuASAR-MPRA \( p \)-values according to the logistic model (see Methods).](image)

We then asked which computational annotations seem to be the most complete and accurate predictors of the effect of a sequence variant on gene regulation as validated by MPRA. We considered effect-SNPs active in LCLs or HepG2 (Moyerbrailean et al., 2016b), non-effect SNPs (negative control) (Moyerbrailean et al., 2016b), predicted functional SNPs from CATO (Maurano et al., 2015), GKM-svm (Lee et al., 2015a) (a gapped kmer...
sequence-based computational method to predict the effect of regulatory variation), and DeepSEA (Zhou and Troyanskaya, 2015) (predicts genomic variant effects at the variant position using deep learning-based approach). Each of the functional annotations show marked differences in \( p \)-value distribution. As expected, SNPs in active TF footprints, but not predicted to affect binding, show no departure from the overall distribution. In both LCLs (Figure 9) and HepG2 (Figure 23), CATO and GKM-svm SNPs had the greatest departure from the null, closely followed by effect-SNPs.

However, effect-SNPs annotated a considerably larger number of SNPs for both cell-types and were also able to predict cell type-specific effects. LCL effect-SNPs in LCLs had a \( p \)-value distribution with a greater departure from the null than the HepG2 effect-SNPs (Figure 24) \( (p = 1.77 \times 10^{-15} \) for LCL effect-SNPs vs \( p = 0.14 \) for HepG2 effect-SNPs), whereas HepG2 effect-SNPs in HepG2 had a \( p \)-value distribution with a greater departure from the null than the LCL effect-SNPs \( (p = 1.81 \times 10^{-4} \) for HepG2 effect-SNPs vs \( p = 1.06 \times 10^{-7} \) for LCL effect-SNPs Figure 25). The differences found here in HepG2 however are minor, potentially due to fewer annotations (993 annotated LCL effect-SNPs vs 193 HepG2 effect-SNPs).
Figure 9: **Validating computational genomic annotations in LCLs.** QQ plot depicting the \( p \)-value distributions from testing for ASE using QuASAR, overlapping with computational genomic annotations in LCLs. Effect-SNP scores have a threshold of \(< -3 \) or \( > 3 \). CATO (Maurano et al., 2015) prediction scores have a threshold of \( > 0.1 \). GKM-svm (Lee et al., 2015a) gapped kmer sequence-based computational method to predict the effect of regulatory variation has a threshold of \(< -6 \) or \( > 6 \). DeepSEA (Zhou and Troyanskaya, 2015) predicts genomic variant effects at the variant position using deep learning-based algorithmic framework. The functional significance predictions have a threshold of \( < 0.05 \). An annotation enrichment \( p \)-value is reported next to their labels, but only for those annotations that are significantly enriched for small QuASAR-MPRA \( p \)-values according to the logistic model (see Methods).

Finally, to formally quantify which annotations are the best predictors of the ASE found in the MPRA, we used all experimental and computational annotations within a logistic model to predict which SNPs in the MPRA data have a nominally significant QuASAR \( p \)-value \((p < 0.001)\). The top predictors were GKM-svm SNPs \((p < 2 \times 10^{-16})\) and effect-SNPs \((p = 2.17 \times 10^{-15})\) in LCLs (Table 12). In HepG2, effect-SNPs were the greatest predictor \((p = 1.18 \times 10^{-10})\) (Table 13).

**Discussion**

High throughput reporter assays have proven extremely useful for the experimental validation of enhancer regions. The recent adaptation of MPRA to investigate ASE additionally allows for validation of regulatory variants in TF binding sites, which have been
shown to be functionally relevant to fine map eQTLs and GWAS signals. These large datasets, however, require analysis methods to handle the intrinsic overdispersion resulting from the original plasmid proportions, variability in the allelic imbalance, and base-calling errors.

The major advantage of QuASAR-MPRA compared to other well calibrated methods is that it requires a small number of replicates allowing for a more efficient study design. QuASAR-MPRA (along with the other methods used here) resulted in a computation time of under a minute and should scale linearly with the number of SNPs being tested. Our QuASAR-MPRA approach identifies causal regulatory variants from high-throughput reporter assays by taking into account overdispersion present in the data. This results in a well calibrated test, with minimal inflation, as determined by lambda values close to 1. In addition to being a robust method to identify ASE in high-throughput reporter assays, this method estimates effect sizes and standard errors for each SNP, which can be used in fixed effects meta-analysis to easily combine datasets. Additionally, we retain a larger number of discoveries 602 (FDR 10%) compared to the original MPRA study (441 at 10%FDR) in LCLs.

Finally, we show that the allele-specific regulatory functions identified with QuASAR-MPRA can be used to validate genomic annotations as predictors for allele-specific effects. Knowing which annotations are the best predictors can aid in identifying true causal SNPs. Here we find that LCL dsQTLs and effect-SNPs are significantly predictive of ASE in LCLs and HepG2 with CATO, while GKM-svm is significant in only LCLs. Using genomic annotations can additionally help us assign mechanism of action to these regulatory variants. If a variant impacts a TF binding site for example, this can lead to gene expression changes,
and therefore phenotypic effects. The less compelling results found in HepG2 may be due to HepG2 having fewer RNA replicates in the MPRA dataset than LCLs. Also there is less data available by ENCODE for the various genomic annotations, likely due to the fact that LCLs (particularly GM12878) are a tier 1 cell line and that other studies also used it for dsQTL analysis (Degner et al., 2012), where HepG2 is only a tier 2 cell type.

Here we have used QuASAR-MPRA on two MPRA datasets, however this method can potentially be used for other high-throughput reporter assays, such as the ones derived from the STARR-seq protocol (e.g., POP-STARR-seq) (Vockley et al., 2015) and CRE-seq protocols (Kwasnieski et al., 2012), and in the context of high-throughput mutagenesis experiments. As the quest for functional validation of regulatory variants becomes more and more wide-spread, these high throughput reporter assays, when combined with a robust statistical test, represent a unique resource to functionally characterize genetic variants at an unprecedented and expandable scale.
CHAPTER 3 High throughput characterization of genetic effects on DNA:protein binding and gene transcription

Introduction

Genome wide association studies (GWAS) have identified thousands of common genetic variants associated with complex traits, including normal traits and common diseases. Many GWAS hits are in non-coding regions, so the underlying mechanism leading to specific phenotypes is likely through disruption of gene regulatory sequence. Quantitative trait loci (QTLs) for molecular and cellular phenotypes (Dermitzakis, 2012), such as gene expression (eQTL) (Brem and Kruglyak, 2005; Stranger, 2007; Innocenti et al., 2011; Wen et al., 2015; Aguet et al., 2017), transcription factor binding (Kasowski et al., 2010), and DNase I sensitivity (dsQTL) (Degner et al., 2012) have been crucial in providing strong evidence and a better understanding of how genetic variants in regulatory sequences can affect gene expression levels (Albert and Kruglyak, 2015; Aguet et al., 2017; Gibbs et al., 2010; Melzer et al., 2008). In recent work, we were able to validate 48% of computationally predicted allelic effects on transcription factor binding through traditional reporter assays (Moyerbrailean et al., 2016b). However, traditional reporter assays are limited by the time and the cost of testing variants one at a time.

Massively parallel reporter assays (MPRA) have been developed for the simultaneous measurement of the regulatory function of thousands of constructs at once. For MPRA, a pool of synthesized DNA oligos containing a barcode at the 3’UTR of a reporter plasmid is transfected into cells, and transcripts are isolated for RNA-seq. Expression driven by the

---

synthesized enhancer region is estimated from the number of RNA reads normalized by the number of corresponding DNA reads (Melnikov et al., 2012; Kwasnieski et al., 2012; Patwardhan et al., 2012; Sharon et al., 2012; Kwasnieski et al., 2014). An alternative to MPRA is STARR-seq (self-transcribing active regulatory region sequencing) (Arnold et al., 2013), whose methods involve fragmenting the genome and cloning the fragments 3′ of the reporter gene. The approach is based on the concept that enhancers can function independently of their relative positions, so putative enhancers are placed downstream of a minimal promoter. Active enhancers transcribe themselves, with their strength quantified as the amount of RNA transcripts within the cell. Because they do not use separate barcodes, STARR-seq approaches have streamlined protocols that allow for higher throughput. Recently, high-throughput assays have been used to assess the enhancer function of genomic regions (Arnold et al., 2013; Wang et al., 2017), the allelic effects on gene expression for naturally occurring variation in 104 regulatory regions (Vockley et al., 2015), fine-map variants associated with gene expression in lymphoblastoid cell lines (LCLs) and HepG2 (Tewhey et al., 2016), and fine-map variants associated with red blood cell traits in GWAS (Ulirsch et al., 2016). These and other approaches with higher scalability and efficiency are necessary to validate and understand the validity of computational predictions and statistical associations for likely causal genetic variants.

In addition to using reporter assays to measure enhancer function on gene expression, there are several methods to directly measure binding affinity of DNA sequences for specific transcription factors. These methods include Spec-seq (Stormo et al., 2015), EMSA-seq (electrophoretic mobility shift assay-sequencing) (Wong et al., 2011), and BUNDLE-seq (Binding to Designed Library, Extracting, and sequencing) (Levo et al., 2015). In
these assays, synthesized regions are combined in-vitro with a purified transcription factor. The bound DNA-factor complexes are then isolated by polyacrylamide gel electrophoresis (PAGE), where sequencing of the derived libraries allows for quantification of the binding strength of regulatory regions. The benefit to these methods is that it is possible to assay any potential genetic variant of interest. In-vivo methods (such as DNase-seq, ChIP-seq), instead, are limited to existing variation within a given sample. Also, in-vivo methods cannot look at each transcription factor separately to identify the specific factor directly causing the change in binding, as the binding could be indirect with any number of co-factors. While BUNDLE-seq compared binding and reporter gene expression, and EMSA has been previously used to ascertain allelic effects, none of the high-throughput EMSA methods have been previously used to determine allelic effects on binding.

Methods

BiT-STARR-seq

Tables 15, 16, 17 report the annotations we have considered with their sources and Tables 29, 25 include the library composition. Each regulatory region was designed to have two oligos: one for each of the alleles. DNA inserts 230bp long, corresponding to 200bp of regulatory sequence, were synthesized by Agilent to contain the regulatory region and the SNP of interest within the first 150bp (Figure 27). We performed a first round of PCR to double strand the oligos and complete the sequencing primers, followed by a subsequent round of PCR to amplify the material. Plasmid pGL4.23 (Promega) was linearized using CloneAmp HiFi PCR Premix (Clontech), primers [STARR_F_SH and STARR_R_SH]. Inserts were cloned into the linear plasmid using standard Infusion (Clontech) cloning protocol.
Clones were transformed into XL10-Gold Ultracompetent Cells (Agilent) in a total of 7 reactions. DNA was extracted using Endofree maxiprep kit (Qiagen).

The DNA library was transfected into LCLs (GM18507) using standard nucleofection protocol, program DS150. We performed nine biological replicates of the transfection from 7 independent cell growth cultures. After transfection, cells were incubated at 37°C and 5% CO2 in RPMI1640 with 15%FBS and 1% Gentamycin for 24h. Cell pellets were then lysed using RLT lysis buffer (Qiagen), and cryopreserved at -80°C. RNA libraries. Total RNA was isolated from the thawed lysates using RNeasy Plus Mini Kit (Qiagen). RNA-seq libraries from the poly-Adenylated RNA were prepared using a custom protocol described in Appendix B: Library preparation. We prepared 7 replicates of the DNA library using a modified version of the PCR protocol as described in (Buenrostro et al., 2013) (Appendix B: Library preparation).

BiT-BUNDLE-seq

We developed BiT-BUNDLE-seq, by modifying the design of the BUNDLE-seq protocol (Levo et al., 2015). Specifically, input DNA sequences were extracted from the BiT-STARR-seq DNA plasmid library. We used N-terminal GST-tagged, recombinant human NFKB1 from EMD Millipore. Experiments were performed in triplicates for each NFKB1 concentration. Libraries extracted from the bound and unbound DNA bands after PAGE were quantified and loaded on the Illumina NextSeq 500 for sequencing.

Data Processing

Reads were mapped using the HISAT2 aligner (Kim et al., 2015), using the 1Kgenomes snp index so as to avoid reference bias. Realigning the reads to GRCh38 should not affect the conclusions as any problematic region of the genome is excluded from any analysis.
(Appendix B: Oligo selection and design). We then ran UMItools (Smith et al., 2017) using standard flags to remove duplicates. To identify SNPs with allele-specific effects, we applied QuASAR-MPRA (?), where for each SNP the reference and alternate allele counts were compared to the DNA proportion. QuASAR-MPRA results from each replicate were then combined using the fixed effects method, and corrected for multiple tests (Benjamini and Hochberg, 1995).

Each replicate for the bound and unbound libraries from BiT-BUNDLE-seq were run through QuASAR-MPRA using the calculated reference proportion (combined unbound and bound DNA). These were then compared using ΔAST (Moyerbrailean et al., 2016a) to identify ASB in the bound fraction that is differential relative to the unbound fraction. The replicates were combined using Stouffer’s method (STOUFFER et al., 1949) to identify ASB for each NFKB1 concentration, and combined again to identify the total ASB. Libraries were additionally analyzed with DESeq2 (Love et al., 2014) to discriminate between bound and unbound constructs.

Results

We have developed a new streamlined method called BiT-STARR-seq (Biallelic Targeted STARR-seq) to test for allele specific effects in regulatory regions (Figure 10A, 27). We selected different categories of regulatory variants for this study including eQTLs (Wen et al., 2015; Innocenti et al., 2011), CentiSNPs (Moyerbrailean et al., 2016b), ASB SNPs (Moyerbrailean et al., 2016b), variants associated with complex traits in GWAS (Pickrell, 2014), and negative ASB controls (Moyerbrailean et al., 2016b) for a total of 50,609 SNPs (see Methods). We designed two oligos targeting each of the alleles for a SNP and
containing the regulatory region (200bp) centered on the SNP (Figure 10A, 27, see Methods). We also included the use of unique molecular identifiers (UMIs), added during cDNA synthesis. With these random UMIs we are in effect tagging identifiable replicates of the self-transcribing construct, which improves the analysis of the data by accounting for PCR duplicates. Our protocol also has the advantage of being highly streamlined. Unlike STARR-seq, our method does not require preparation of DNA regions for use in the assay, such as whole genome fragmentation (Arnold et al., 2013), or targeting regions (Vanhille et al., 2015), while, similar to STARR-seq, it requires only a single cloning and transformation step. Because the UMIs are inserted after transfection, there are no additional bottleneck issues (due to library complexity) in the cloning and transformation steps.

We generated 7 replicates of the DNA library, which were highly and significantly correlated (Figure 28, Spearman's $\rho = (0.97, 0.98)$, $p$-value<0.01). The DNA library was then transfected in LCLs (9 biological replicates) and we were able to recover a total of 43,500 testable SNPs (see methods for RNA counts filter). Read counts for the 9 biological replicates were highly correlated (Figure 29, Spearman's $\rho = (0.35, 0.72)$, $p$-value<0.01). To identify SNPs with allele-specific effects, we applied QuASAR-MPRA (?). For each SNP, the reference and alternate allele counts were compared to the DNA proportion in the plasmid library. We identified a total of 2,720 SNPs with ASE from the combined replicates (FDR 10%) (Table 14). To investigate the importance of UMIs in this experimental approach, we re-analyzed our data without removing duplicates. For the combined replicates, inflation (genomic inflation test, see Methods) is greatly increased (from 1.10 to 1.73). If only a single RNA library replicate is considered the number of detected ASE is about 4x reduced (Figure 10B).
SNPs with ASE are significantly enriched for variants predicted to impact transcription factor binding (CentiSNPs) (Moyerbrailean et al., 2016b) (Figure 10C, 30, Table 15, Table 18, Fisher’s exact test OR=2.49, p-value=4.55 × 10^{-06}). Additionally, SNPs with ASE are enriched for low p-values in an eQTL mapping study performed in immune cells (Nédélec et al., 2016) (Figure 10D, Table 17), thus confirming that our synthetic oligos can reproduce allele-specific regulatory effects observed in a native (non-episomal) cellular context.
Figure 10: **BiT-STARR-seq and BiT-BUNDLE-seq identify regulatory variants in non-coding regions.** A) Experimental outline. Oligos targeting the regulatory regions of interest (and either reference or alternate alleles) are designed to contain, on their ends, 15bp matching the sequencing primers used for Illumina NGS. The DNA library is used both in the BiT-STARR-seq and BiT-BUNDLE-seq experiments. UMIs are added during cDNA synthesis for the BiT-STARR-seq RNA-seq library and prior to PAGE in the BiT-BUNDLE-seq protocol. B) QQplot depicting the $p$-value distributions from QuASAR-MPRA for a single experimental replicate processed without removing duplicates (purple) or after removing duplicates using the UMIs (pink). C) QQplot depicting the $p$-value distributions from the ASE test performed using QuASAR-MPRA on all replicates after removing duplicates. CentiSNPs are in (green) (Moyerbrailean et al., 2016b) while SNPs in the negative control group are in (grey). D) QQplot depicting the $p$-value distributions for eQTLs from (Nédélec et al., 2016). SNPs with significant ASE (FDR 10%) are in (blue) or not significant ASE are in (grey).
Motif, regulatory region size and chromatin context effects

The CentiSNP annotation is informative of the specific transcription factor motif being disrupted by a SNP. By leveraging this information, we were able to analyze allelic effects for specific transcription factor motifs (Figure 11A, Table 19). Additionally, by combining the ASE results with the direction of the motif, we can characterize whether the motif is active in both directions or only in one direction. This would suggest that some TF binding motifs tend to function specifically in one direction. We found that when both alleles are covered in both directions, the allelic effects on gene expression tend to agree in direction and magnitude. If we categorize these directional allelic effects per motif, we do not observe major differences with the notable exception of CTCF (Figure 11B). Specifically, we find that SNPs in footprints for CTCF are significantly enriched (Fisher’s exact test OR=1.57, Bonferroni p=0.02) for ASE when the direction of transcription of the reporter gene is opposite to that of the motif strand (Figure 11C).

While oligos were designed to have the variant centered in the middle of the synthesized region, this does not necessarily mean that the SNP is centered in a DNase window (Figure 27). While position within the window does not affect the ASE signal, the main effect seems to depend on the presence / absence of the tested site within the DNase window (Figure 31) and in a lesser degree it depends of the peak size (Figure 32). SNPs were originally selected based on the CentiSNP annotation, but when we considered chromatin states (Broad ChromHMM state) we detect enrichment for promoter state among SNPs with ASE (Figure 11D).
Allele specific binding for NFKB1

In order to understand the effect of a regulatory variant on complex traits it is necessary to first dissect the molecular function that is impacted by the nucleotide change. The CentiSNP prediction provides specific hypotheses for allelic effects on transcription factor binding that can be directly tested experimentally. Further matching ASB to ASE identified in BiT-STARR-seq would provide a complete molecular mechanism, from predicted binding...
effects, to experimentally validated binding effects, to validated effects on expression. Due
to the enrichment of CentiSNPs among SNPs with ASE in BiT-STAR-seq, we performed
BiT-BUNDLE-seq to validate their effect on transcription factor binding. This is a new and
efficient extension of high throughput reporter assays, since it uses the same input DNA
library. We performed BiT-BUNDLE-seq with purified NFKB1 (at three different concentra-
tions), which is an important regulator of the immune response in LCLs and other immune
cells (Li and Verma, 2002; Beinke and Ley, 2004; Smale, 2010). Previous studies have
successfully identified ASB from ChIP-seq for all NF-kB subunits in LCLs (Zhao et al., 2014;
Heinz et al., 2010; Jin et al., 2013; Kasowski et al., 2010; Lim et al., 2007; Martone et al.,
2003) and NFKB1 footprints are induced in response to infection (Pacis et al., 2015). Ad-
ditionally, NF-kB complex was found to be 50 fold enriched for reQTLs from response to
Listeria and Salmonella (Nédélec et al., 2016).

We first analyzed NFKB1 binding between the bound and unbound libraries and iden-
tified 9,361 significantly (logFC>1 and FDR<1%) over-represented regions in the bound
library (Table 20, 21, 22, 23). As expected, these regions were enriched (OR=11.70 to
13.75, \( p\)-value=7.95 \( \times \) 10\(^{-27}\) to 1.23 \( \times \) 10\(^{-15}\)) for NF-kB complex footprints (Figure 12A, 33),
with a higher portion of these regions in the mid concentration of NFKB1 as compared to
the low or high concentrations (Figure 12B). We hypothesize that this is because the low
concentration doesn't capture all of the NFKB1 binding, whereas the high concentration
likely results in non-specific binding. We then used \( \triangle \)AST (Moyerbrailean et al., 2016a)
to identify ASB in the bound library (as compared to the unbound library), and combined
the replicates using Stouffer's method (see Methods, Figure 12C).
Figure 12: Allele-specific binding for NFKB1. A) Density plot of the logFC (from DESeq2) between bound and unbound DNA fractions from the BiT-BUNDLE-seq experiment. In red are the regions containing a SNP in a NF-kB complex footprint, in blue the regions containing a SNP in footprints for other transcription factors. B) Barplot representing the number of independent enhancer regions in bound (dark color, DESeq2 logFC > 1 and FDR < 1%) and unbound (light color) DNA. NFKB1 concentration and presence of a NF-kB complex footprint are indicated in the two columns on the left of the panel. C) QQplot depicting the p-value distributions from testing for ASB signal specific to the bound DNA fraction using ∆AST (black) and SNPs in the negative control group are in (grey). D) QQplot depicting the ASE p-value distribution from QuASAR-MPRA for SNPs with significant (FDR < 10%) ASB (green), SNPs with significant (FDR < 10%) ASB and are also in CREB1 or AML1 footprints (maroon), or not significant ASB (grey) in the BiT-BUNDLE-seq experiment.

We successfully identified 386 SNPs at low concentration, 797 SNPs at mid concentration, and 894 SNPs at high concentration with significant ASB at 10% FDR (Figure 12D, 34), for a total of 2,684 SNPs when aggregating all experiments (Table 16, 24). These spanned our designed regulatory categories, with the greatest number covering CentiSNPs (Table 25). When we considered these ASB SNPs in combination with the ASE results from the BiT-STARR-seq (Figure 35), we found that SNPs with ASE are enriched for NFKB1 ASB (Figure 12D, Fisher’s exact test (OR=2.04, p-value=1.51 × 10^{-16})). For ASB variants
not showing ASE, we found that there is enrichment for these being in either the CREB1 or AML1 motifs (Figure 12D, 36, Fisher’s test see Table 26), which are factors known to antagonize NF-kB complex binding (Ollivier et al., 1996; Parry and Mackman, 1997; Nakagawa et al., 2009, 2011). This confirms our hypothesis that disruption of NFKB1 binding is one of the mechanisms underlying allele-specific expression in our dataset.

**Overlap with signals from GWAS**

We used ASB and ASE in combination with transcription factor binding motifs to assign mechanistic function to putatively causal SNPs linked to complex traits. We found 2,054 CentiSNPs with ASB ($p$-value<0.05) and 1,769 CentiSNPs with ASE ($p$-value<0.05) associated to a complex trait in the GWAS catalog (Table 27, 28) or from fgwas (Moyerbrailean et al., 2016b). Considering all SNPs tested, there are 173 SNPs that have both ASB and ASE (FDR 10%), and 164 of them (95%) are also associated with a complex trait. We were able to dissect the molecular mechanism for rs3810936, a variant associated with risk for Crohn’s disease in multiple populations (Yamazaki et al., 2005; Franke et al., 2010; Lee et al., 2015b; Baskaran et al., 2014) (Figure 13A,B). This variant is a CentiSNP for the factor Hmx3 (Nkx5-1) and we find ASB for NFKB1 ($p$-value=0.006) in the BiT-BUNDLE-seq assay and ASE ($p$-value=0.034) in both directions in the BiT-STARR-seq. This SNP is a synonymous variant in gene *TNFSF15* (also known as *TL1A*), which encodes for a cytokine that belongs to the tumor necrosis factor (TNF) ligand family.
Increased TL1A expression has been reported in inflamed Crohn’s disease tissue compared with non-inflamed areas, and in ulcerative colitis patient serum (Bamias et al., 2010, 2003; Prehn et al., 2004). TL1A gives costimulatory signals to activated lymphocytes through binding to death-domain receptor 3 (DR3) (Migone et al., 2002) which induces the secretion of interferon gamma (IFNG) (Papadakis et al., 2005; Prehn et al., 2004). This gene modulates Th-1 and Th-17 (Bamias et al., 2003; Takedatsu et al., 2008), creating an immunological state that leads to the mucosal inflammation of Crohn’s disease. Stimulation of the TL1A pathway, in monocytes and T cells from patients carrying the disease-associated TL1A SNPs, showed higher levels of TL1A expression, therefore aberrant TL1A expression may be a factor driving IBD development (Michelsen et al., 2009; Kakuta et al., 2009). This gene additionally has been found to be downregulated in response to dexamethasone (Moyerbrailean et al., 2016a), a corticosteroid used to treat many inflammatory and autoimmune conditions. While this variant is not found in ChIP-seq from ENCODE, ENCODE studies used RELA (p65) for NF-κB subunit, where our study used NFKB1 (p50). We therefore identify a novel variant that disrupts binding of NFKB1, where the alternate
allele (C) has increased binding. This leads to an increase in gene expression for the alternate allele, which is also the risk allele for Crohn’s disease (OR=1.21, \(p\)-value\(=1 \times 10^{-15}\)).

**Discussion**

The recent adaptation of MPRA to investigate ASE allows for validation of regulatory variants in transcription factor binding sites, which have been shown to be functionally relevant to fine map eQTLs (Tewhey et al., 2016) and GWAS signals (Ulirsch et al., 2016). However, the use of functional genomics to select relevant regions prior to experimental validation can reduce the number of sites it is necessary to validate. We developed a high throughput reporter assay that synthesizes these selected regions (similar to MPRA), clones them in 3′ of the reporter gene (similar to STARR-seq), and includes the addition of a UMI during cDNA synthesis (new to our protocol). This is the most streamlined protocol to date, and allows for removal of PCR duplicates which reduces noise in the data for greater power to detect ASE.

Our results show that using existing annotations to prioritize regulatory variants for high-throughput reporter assays is an effective strategy. The CentiSNP annotation in particular contains information that can be used to analyze ASB/ASE for individual transcription factor motifs and investigate potential molecular mechanisms of action. We found that direction is an important factor in the case of CTCF, most likely due to how CTCF functions as an insulator between the enhancer and the promoter when they are in antiparallel directions. Previous studies have shown that CTCF, a well characterized insulator, has binding sites at the anchors of chromatin loops. These are arranged in forward-reverse orientations (Guo et al., 2015; Alt et al., 2013; Guo et al., 2012; Monahan et al., 2012;
Rao et al., 2014; Vietri et al., 2015), where the relative positions and orientations of the binding sites are important for the mechanism of action (Guo et al., 2015). In our case, the interaction could be mediated either by the basal transcriptional machinery at the TSS or also an additional weak CTCF binding site (M01259) that is present in the promoter and could help to establish a DNA loop. However, there may be alternative explanations for this result as reporter assays may not reflect the native regulatory landscape in human cells (Muerdter et al., 2017; Huervano et al., 2013).

Generally, caution should be used in interpreting reporter assay gene expression differences across cell types, because transfection may perturb the cell state. However, it is important to highlight that any trans-acting effects (e.g., promoter strength, type 1 interferon response activation) should affect both alleles similarly and therefore should not induce false positives in the allele-specific signal.

We used our library of oligos also in a BiT-BUNDLE-seq assay for identification of ASB for NFKB1. This is a novel approach to combine ASB and ASE identification in high throughput assays using the same sequences. Our results show that this integration is a useful approach to validate the molecular mechanism for specific transcription factors. Allelic effects on transcription factor binding and gene expression are not always concordant. Some of this discordance is due to lack of power to detect ASB/ASE overlap, as well as other technical considerations. For example, in BiT-BUNDLE-seq only one single TF (NFKB1) is available for binding, whereas in BiT-STARR-seq, other co-factors are present in the cell to affect binding. Additionally, there can be discordance in direction of effect, where, for example, an allele can lead to increased binding of a factor with repressing activity on gene expression (e.g. variants in CREB1/AML1 binding sites). These regulatory
events are likely to be captured in the BiT-STARR-seq assay, which is performed in LCLs where CREB1, AML1 and NFKB1 are active. These results highlight that multiple type of assays are necessary to capture the detailed molecular mechanism of gene regulation. Additionally, integration with GWAS can identify and further characterize the molecular mechanisms linking causal genetic variants with complex traits.
CHAPTER 4 Rapid profiling of regulatory regions and variants to study GxE in complex traits

Introduction

A large fraction of loci important in determining human traits and disease conditions are located in non-coding regions of the genome. These regions likely contain specific regulatory sequences that control gene transcription and can also interact with changes in the cellular environment (e.g. drug treatment). Recent large scale efforts in functional genomics have facilitated the profiling of regulatory sequences across many cell-types and tissues (Consortium, 2012; Consortium et al., 2015), yet we are still very far from mapping the sequences that control the transcriptional response to many external stimuli.

Importantly, much of the missing heritability in GWAS may be a consequence of small effect sizes dampened by unaccounted environmental interactions. It can be difficult to study the effect of the environment on complex traits however, as the (human) environment is very complicated. For example, in GWAS studies it would be very hard to test for some environmental influences while controlling for all others. One viable alternative to identify GxE is to use the cellular environment as a proxy. Using an in-vitro system with relevant cell-types, we can better control for environment and measure molecular phenotypes. Many studies have used this system to study the differences in gene expression due to treatments. One common major cellular pathway studied is the glucocorticoid response (Mukherjee et al., 2017; Chang et al., 2015; Poon et al., 2012; Richards et al., 2017). It's beneficial role in human health dates all the way back to 1950 where it was used to treat

---

asthma (CARRYER et al., 1950). Variation in gene expression has been found to play a role in complex traits in additional studies as well (Li et al., 2015; Nicolae et al., 2010; Raj et al., 2014; Grundberg et al., 2012). Supporting this hypothesis, we recently demonstrated that genes with gene-environment (GxE) interactions at the molecular level are highly enriched in GWAS Moyerbrailean et al. (2016a). More precisely, 49% of genes with condition-specific allele specific expression in a study screening 250 cellular environments were also found associated with GWAS traits. Beyond looking at gene expression, it is well known that many enhancers are controlled by external stimuli (Hurtado et al., 2011; Biddie et al., 2011; Hah et al., 2013; Shlyueva et al., 2014). However, few studies have investigated the effect of external stimuli on regulatory regions using high-throughput reporter assays (Shlyueva et al., 2014; Vockley et al., 2016), and none have investigated the allele specific effects.

Applying GEMMA (genome wide efficient mixed model association) using conditional allele specific expression (cASE) (Moyerbrailean et al., 2016a) or eQTLs (Wen et al., 2015; Consortium et al., 2015) we were able to calculate enrichment (heritability) for 18 complex traits from GWAS. The highest values for cASE were observed for blood total cholesterol level (TC) (9.7-fold), triglycerides (TG) (7.9-fold), and for mean corpuscular hemoglobin levels (MCH) (8.6-fold). We additionally profiled five of the environmental conditions from Moyerbrailean et al. (2016a) with very large gene expression changes for transcription factor binding activity. This was accomplished at a genome-wide scale by ATAC-seq, which utilizes the Tn5 transposase to fragment and tag accessible DNA. We further modeled the Tn5 cleavage pattern “footprint” of transcription factors with known motifs with Centipede to identify bound binding sites. From our analyses we were able to characterize
9,263 (dexamethasone), 2,615 (copper), and 2,115 (selenium) regions that have differentially accessible regions (FDR<10%) in response to treatment. Using BiT-STARR-seq (a high throughput allele-specific reporter assay (Kalita et al., 2018b)) with treatment we then further validated these gene by environment (GxE) effects for specific variants. We were able to identify thousands of ASE (FDR<10%) with retinoic acid (1,939), dexamethasone (938), selenium (3,790), and caffeine (3,089).

Our results demonstrate that ATAC-seq together with an improved footprint model and BiT-STARR-seq are excellent tools for rapid profiling of transcription binding factor activity to study cellular regulatory response to the environment and molecular mechanisms underlying GxE.

Methods

Analysis of heritability enrichment using GEMMA

To run GEMMA (Zhou and Stephens, 2012, 2014) we partitioned SNPs genome wide to create a category file. Each SNP was assigned to one of the following categories for the cASE analysis: cASE (genic regions with conditional allele specific expression) or iASE (genic regions with induced allele specific expression), ASE (genic regions with allele specific expression), Other Genic (genic regions that do not fall into any of the categories above) and Intergenic (greater than 100kb from any gene). Each SNP was assigned to one of the following categories for the eQTL analysis: eQTL (genic regions with eQTL), Other Genic (genic regions that do not contain an eQTL) and Intergenic (greater than 100kb from any gene). We then used GEMMA to compute the SNP correlations among different categories from a reference panel (502 individuals of European ancestry from the 1000
genome project). This was followed by summing the $Z^2$ statistics from the GWAS meta-analysis within the categories. Finally, we computed the proportion of variance, and fold enrichment of heritability explained by each category. A table of the results can be found in Table 3.

**ATAC-seq data collection**

The lymphoblastoid cell line (LCL) GM18508 was purchased from Coriell Cell Repository. LCLs were cultured in serum containing charcoal-stripped FBS and treated for 6 hours with 1 µM selenium, dexamethasone or copper as described in Moyerbrailean et al. (2016a). Cells were also cultured in parallel with the vehicle control (water or ethanol (1 ml in 10,000 ml), to represent the solvent used to prepare the treatment. We then followed the protocol by Buenrostro et al. (2013) to lyse 25,000-100,000 cells and prepare ATAC-seq libraries, with the exception that we used the Illumina Nextera Index Kit (Cat #15055290) in the PCR enrichment step. Individual library fragment distributions were assessed on the Agilent Bioanalyzer and pooling proportions were determined using the qPCR Kapa library quantification kit (KAPA Biosystems). Library pools were run on the Illumina NextSeq 500 sequencer in the Luca/Pique-Regi laboratory. Barcoded libraries from three ATAC-seq samples, performed with 25,000, 50,000 and 75,000 cells, were pooled and sequenced on multiple sequencing runs for 100M 38bp PE reads.

**ATAC-seq preprocessing and differential chromatin accessibility**

Reads were aligned to the reference human genome hg19 using bwa mem (Li and Durbin (2009) http://bio-bwa.sourceforge.net). Reads with quality <10 and without proper pairs were removed using samtools (http://github.com/samtools/).

To identify differentially accessible regions, we used DESeq2 (Love et al., 2014) (R ver-
version 1.8.1). We separated the genome into 300bp regions and coverageBed was used to count reads in these regions. Tiles were only included if they contained the top 5% of read depths excluding those which were greater than 99.995% of read depths. These counts were then utilized in DESeq2 using the following model:

\[
Y_{jn} = \sum_t \beta_{jt} M_{tn}
\]

where \(Y_{jn}\) represents the internal DEseq mean accessibility parameter for region \(j\) and experiment \(n\), \(M_n\) is the treatment indicator (control or treatment), and \(\beta_{jt}\) parameter is the treatment effect. Regions that changed following treatments were defined with BH FDR < 10%.

**Transcription factor binding footprints, and differential binding analysis**

To detect which transcription factors motifs have footprints in each condition we adapted Centipede in two ways: i) we use the fragment length information contained in the ATAC-seq in the footprint model, and ii) we also developed a new test to detect global shifts in footprints for each TF motif to reveal which factors may be activated in each condition.

As in Centipede we need to start from candidate binding sites a given motif model. For each transcription factor we scan the entire human genome (hg19) for matches to its DNA recognition motif using position weight matrix (PWM) models from TRANSFAC and JASPAR as previously described (Pique-Regi et al., 2011). Then for each candidate location \(l\) we collect all the ATAC-seq fragments which are partitioned into four binds depending on the fragment length: 1) [39-99], 2) [100-139], 3) [140-179], 4) [180-250]. For each
fragment, the two Tn5 insertion sites were calculated as the position 4bp after the 5′-end in the 5′ to 3′ direction. Then for each candidate motif, a matrix $X$ was constructed to count Tn5 insertion events: each row represented a sequence match to motif in the genome (motif instance), and each column a specific cleavage site at a relative bp and orientation with respect to the motif instance. We built a matrix $\{X_i\}_{i=1}^4$ for each fragment length bin, each using a window half-size $S=150\text{bp}$ resulting in $(2 \times S + W) \times 2$ columns, where $W$ is the length of the motif in bp.

First, we fit the Centipede model in a subset of instances to determine which motifs have active binding (i.e. show footprints) with a Bonferroni’s corrected p-value of $p < 0.05$. The statistical significance is assessed by calculating a $Z$-score corresponding to the PWM effect in the prior probability in the Centipede’s logistic hierarchical prior. Then we used Centipede and motif instances with posterior probabilities higher than 0.99 to denote locations where the transcription factors are bound. To determine which TF were active in the first step, we calculate a $Z$-score corresponding to the PWM effect in the prior probability in Centipede’s logistic model and we determined as active those that had a Bonferroni corrected $p<0.05$. The $Z$-score corresponds to the $\beta$ parameter in:

$$
\log \left( \frac{\pi_l}{1-\pi_l} \right) = \alpha + \beta \text{PWMscore}_i
$$

(0.11)

where $\pi_l$ represent the prior probability of binding in Centipede’s model in motif location $l$. In the second step, we first trained Centipede assuming that the footprint was bound in two conditions. Then, we fixed the model parameters and generated a likelihood ratio and posterior probability $\pi_{tl}$ for each condition $t$ separately and for each site $l$. To detect
if the footprint was more active in one of the two conditions, we fit a logistic model that included an intercept for each condition (α and δ), the PWM effect β, and PWM times the treatment effect γ:

\[
\log \left( \frac{\pi_l}{1 - \pi_l} \right) = \alpha x (1 - I_t) + \beta x \text{PWMscore}_l + \delta x I_t + \gamma x (I_t x \text{PWMscore}_l)
\]  

(0.12)

where \( I_t \) is an indicator variable that takes the value 1 if \( t = \) “treatment” and 0 if \( t = \) “control”. We then calculated a Z-score for the interaction effect \( \gamma \), corresponding to the evidence for condition specific binding.

**BiT-STARR-seq oligo preparation**

The Oligo library was taken from a plasmid pool of previous oligos used in Kalita *et al.* (2018b). Extraction of the insert was carried out using Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) and primers [F_transposase and R_transposase] (Table 2) with cycling conditions: 98°C for 30s, followed by 4 cycles of 98°C for 10s, 50°C for 30s, 72°C for 60s, followed by 6 cycles of 98°C for 10s, 65°C for 30s, 72°C for 60s, followed by 72°C for 5 min. 8 samples of the template plasmid pool each with approximately 250ng of DNA were used. The PCR product was run on a 2% agarose gel, extracted and purified with the NucleoSpin Gel and PCR Clean-Up Kit (Clontech).

Table 2: Primers used in BiT-STARR-seq.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>STARRseqP_F_SH</td>
<td>TCTCCGAGCCACGAGACGTCGACGAATTCGGCC</td>
</tr>
<tr>
<td>STARRseqP_R_SH</td>
<td>ATCTGACGCTGCCGACGACCGGTGCATGCTCTA</td>
</tr>
<tr>
<td>F_transposase</td>
<td>TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG</td>
</tr>
<tr>
<td>R_transposase</td>
<td>GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG</td>
</tr>
<tr>
<td>F_trans_short</td>
<td>TCGTCGGCAGCGTCAGAT</td>
</tr>
<tr>
<td>_I2.1</td>
<td>CAAGCAGAAGACGGCATACG</td>
</tr>
<tr>
<td>Nextera_i7_10N</td>
<td>CAAGCAGAAGACGGCATACGAGATRDHBVDHBVDGTCTCGTGGGCTCGG</td>
</tr>
</tbody>
</table>
Cloning Regulatory regions into the STARR-seq Human plasmid

STARR-seq Human plasmid (Arnold et al., 2013) (addgene #71509) was linearized using CloneAmp HiFi PCR Premix (Clontech), primers [STARRseqP_F_SH and STARRseqP_R_SH], and 35 cycles of 98°C for 10s, 60°C for 15s, and 72°C for 5s. The PCR product was purified on a 1% agarose gel as described above. Inserts were cloned into the linear plasmid using standard NEBuilder HiFi DNA Assembly Master Mix cloning protocol. Clones were transformed into MegaX DH10B electrocompetent cells (Thermo Fisher Scientific) in a total of 2 reactions. These reactions were pooled and grown overnight in 400ml LB at 37°C in a shaking incubator. DNA was extracted using Endofree maxiprep kit (QIAGen).

Transfection of LCL library

Previous studies (Muerdter et al., 2017; Huerfano et al., 2013) have found that transfection, especially from nucleofection, can lead to activation of IFN-1 response, which may complicate comparison of enhancer activities between different cell types. In our study design, allelic effects are measured and contrasted within the same cell type, thus any trans-effect is inherently controlled. Furthermore, in LCLs the immune response is already activated because of EBV transformation. DNA library was transfected into LCLs (GM18507) using standard nucleofection protocol, program DS150, 3µg of DNA and 7.5 × 10^6 cells. For each of 3 biological replicates (from independent cell culture growths), we used 12 cuvettes, each set of 4 were pooled and plated. Each set was then treated with either control or treatment (1 control and 2 treatments total per experiment). Treatments concentrations used were as in Moyerbrailean et al. (2016a). Cells were incubated at 37°C and 5% CO2 in RPMI1640 with 15%FBS and 1% Gentamycin for 24h. Cell pellets were then lysed using
RLT lysis buffer (QIAgen), and cryopreserved at -80°C.

**Library preparation**

RNA-libraries. Thawed lysates were split in three aliquotes and total RNA was isolated using RNeasy Plus Mini Kit (QIAgen). Poly-Adenylated RNA was selected using Dynabeads mRNA Direct Kit (Ambion) using the protocol for total RNA input. RNA was reverse transcribed to cDNA using Superscript III First-Strand Synthesis Supermix kit (ThermoFisher) with primer [Nextera_i7_10N] and following the manufacturer’s protocol with the addition of RNase treatment for 30 min. cDNA technical replicates were pooled and SPRI Select beads (Life Tech) were used for purification and size selection at a ratio of 0.9X. PCR Library Enrichment was performed using a nested PCR protocol. For the first round of PCR we used Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) and primers [F_trans_short and Illumina2.1] with cycling conditions: 98°C for 30s, followed by 15 cycles of 98°C for 10s, 72°C for 15s, followed by 72°C for 5 min. PCR product was purified on a 2% agarose gel as described above. The nested PCR used Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) and primers [fixed N5xx adapter (Illumina) (unique per each library replicate) and Illumina2.1] with cycling conditions: 98°C for 30s, followed by 5 cycles of 98°C for 10s, 72°C for 15s, followed by 72°C for 5 min. In a side quantitative real-time PCR reaction, 5 µL of PCR product, 10X SYBR Green I, and the same primers and master mix were run in conditions: 98°C for 30s, 30 cycles of 98°C for 10s, 63°C for 30s, and 72°C for 60s. To determine the number of PCR cycles needed to reach saturation, we plotted linear Rn versus cycle and determined the cycle number that corresponds to 25% of maximum fluorescent intensity on the side reaction (Buenrostro et al., 2013). The PCR product was purified on a 2% agarose gel as described above.
DNA-libraries. We prepared 5 replicates of the DNA library using the PCR protocol as described in (Buenrostro et al., 2013) except using NEB Next 2x Master mix, primers [fixed N5xx adapter (Illumina) (unique per each library replicate) and Nextera_i7_10N] and 30ng of input plasmid DNA. Cycling conditions were: initial denaturation at 72°C for 5 min, followed by 5 cycles of 98°C for 10s, 63°C for 30s, and 72°C for 60s. PCR product was purified on a 2% agarose gel using Qiaquick gel extraction kit (Qiagen).

**Library Sequencing**

Pooled RNA and DNA libraries were sequenced on the Illumina Nextseq500 to generate 125 cycles for read 1, 30 cycles for read 2, 8 cycles for the fixed multiplexing index 2 and 10 cycles for index 1 (variable barcode).

**Data Processing**

Reads were mapped using the Hisat2 aligner (Kim et al., 2015), using the 1Kgenomes snp index so as to avoid reference bias. First we removed variants whose UMI was not possible to be present, given the UMI pattern selected. We then ran UMItools (Smith et al., 2017) using standard flags, as well as a q20 filter. We then ran the deduplicated files through mpileup using a bed file of our full SNP list, the -t DP4, -g, and -d 1000000. Counts filters were as in Kalita et al. (2018b). To identify SNPs with allele-specific effects, we applied QuASAR-MPRA (Kalita et al., 2018a), where for each SNP the reference and alternate allele counts were compared to the DNA proportion. QuASAR-MPRA results from each replicate were then combined for each treatment (or control) using the fixed effects method, and corrected for multiple tests using BH procedure (Benjamini and Hochberg, 1995).
Results

Gene-by-Environment interactions and Complex Traits

From previous work in our lab, we characterized genetic effects on the transcriptional response to 50 treatments in 5 cell types and found that genes with a transcriptional response to environmental perturbations showed 7-fold higher odds of being found in GWAS. To investigate the role of GxE in complex traits directly, we analyzed per SNP heritability for 18 complex traits using Genome-wide Efficient Mixed Model Association (GEMMA) (Zhou and Stephens, 2012, 2014). Similar to the LD-score regression method that partitions heritability estimates across SNPs functional categories (Gusev et al., 2014), we contrasted SNPs in genes with condition specific ASE, genes with ASE, genes without ASE, and inter-genic regions (Table 3). eQTL studies in particular have been successful and widely applied in identifying genomic regions associated with gene expression in various tissues and conditions (Dimas et al., 2009; Maranville et al., 2011; Melé et al., 2015; Nica et al., 2011). We therefore additionally contrast eGenes, genes without eQTLs, and intergenic regions, using both eQTLs fine-mapped in (Wen et al., 2015) and GTEx eQTLs (Consortium et al., 2015). The per SNP heritability for each of these categories is then compared to the genome average. A higher value of per SNP heritability for one of these categories indicates a higher number of causal SNPs, higher effect sizes or both in that category.

We found that the per SNP heritability estimate for SNPs in genes with ASE is 11.1 times higher than the genome average for high-density lipoprotein (HDL). For 13 of the 18 traits analyzed, per SNP heritability estimates for SNPs in genes with cASE, iASE, or ASE were
significantly higher than the genome average. For 7 of these, the cASE and iASE category estimates were even higher than any other partition (Figure 14A), thus indicating that GxE for these traits are particularly relevant. The highest values for cASE and iASE were observed for blood total cholesterol level (TC) (9.7-fold), triglycerides (TG) (7.9-fold), and for mean corpuscular hemoglobin levels (MCH) (8.6-fold). There is high enrichment for GTEx eGenes in bone mineral density (femur) in stomach tissue (enrichment=15, Figure 14C). GTEX eGenes do however show similar enrichment patterns across tissues, likely due to many eQTLs being shared between tissues. LCL eQTLs fine-mapped in Wen et al. (2015) show high enrichment for bone mineral density (spine) (LSBMD) (enrichment=7, Fig 14B). Overall, these results suggest an important role for GxE in a large number of traits.
Table 3: **GEMMA per SNP heritability estimates relative to the genomic average.** Shown for each GWAS trait tested (see Table 5) are the estimates for cASE/iASE (SNPs in genic regions with cASE or iASE), ASE (SNPs in genic regions with ASE), No ASE (SNPs in genic regions without ASE) and Intergenic (SNPs farther than 100kb from any gene). PVE: proportion of variance in phenotypes explained.

<table>
<thead>
<tr>
<th>GWAS Trait</th>
<th>Category</th>
<th>PVE</th>
<th>Std Err</th>
<th>$\sigma^2$</th>
<th>Std Err</th>
<th>Enrichment</th>
<th>Std Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>cASE/iASE</td>
<td>0.0003</td>
<td>0.0001</td>
<td>5.96E-09</td>
<td>2.17E-09</td>
<td>0.9697</td>
<td>0.3819</td>
</tr>
<tr>
<td>BMI</td>
<td>ASE</td>
<td>0.0026</td>
<td>0.0010</td>
<td>1.68E-08</td>
<td>6.11E-09</td>
<td>2.7306</td>
<td>0.9696</td>
</tr>
<tr>
<td>BMI</td>
<td>Intergenic</td>
<td>0.0021</td>
<td>0.0006</td>
<td>5.69E-09</td>
<td>1.50E-09</td>
<td>0.9260</td>
<td>0.2757</td>
</tr>
<tr>
<td>BMI</td>
<td>No ASE</td>
<td>0.0267</td>
<td>0.0036</td>
<td>5.82E-09</td>
<td>7.81E-09</td>
<td>0.9469</td>
<td>0.0407</td>
</tr>
<tr>
<td>CD</td>
<td>cASE/iASE</td>
<td>0.0005</td>
<td>0.0007</td>
<td>8.60E-09</td>
<td>1.43E-08</td>
<td>0.0957</td>
<td>0.1619</td>
</tr>
<tr>
<td>CD</td>
<td>ASE</td>
<td>0.0229</td>
<td>0.0032</td>
<td>1.45E-07</td>
<td>2.03E-08</td>
<td>1.6199</td>
<td>0.2546</td>
</tr>
<tr>
<td>CD</td>
<td>Intergenic</td>
<td>0.0004</td>
<td>0.0067</td>
<td>9.88E-10</td>
<td>1.81E-08</td>
<td>0.0110</td>
<td>0.2015</td>
</tr>
<tr>
<td>CD</td>
<td>No ASE</td>
<td>0.4402</td>
<td>0.0314</td>
<td>9.60E-08</td>
<td>6.84E-09</td>
<td>1.0687</td>
<td>0.0169</td>
</tr>
<tr>
<td>FG</td>
<td>cASE/iASE</td>
<td>0.0000</td>
<td>0.0002</td>
<td>5.83E-10</td>
<td>3.86E-09</td>
<td>0.0847</td>
<td>0.5609</td>
</tr>
<tr>
<td>FG</td>
<td>ASE</td>
<td>0.0002</td>
<td>0.0006</td>
<td>1.56E-09</td>
<td>4.08E-09</td>
<td>0.2267</td>
<td>0.5973</td>
</tr>
<tr>
<td>FG</td>
<td>Intergenic</td>
<td>-0.0014</td>
<td>0.0009</td>
<td>-3.78E-09</td>
<td>2.56E-09</td>
<td>-0.5488</td>
<td>0.3597</td>
</tr>
<tr>
<td>FG</td>
<td>No ASE</td>
<td>0.0365</td>
<td>0.0070</td>
<td>7.99E-09</td>
<td>1.52E-09</td>
<td>1.1616</td>
<td>0.0364</td>
</tr>
<tr>
<td>FNBMD</td>
<td>cASE/iASE</td>
<td>0.0013</td>
<td>0.0004</td>
<td>2.42E-08</td>
<td>6.73E-09</td>
<td>5.1713</td>
<td>2.6679</td>
</tr>
<tr>
<td>FNBMD</td>
<td>ASE</td>
<td>0.0028</td>
<td>0.0011</td>
<td>1.76E-08</td>
<td>6.97E-09</td>
<td>3.7533</td>
<td>2.3087</td>
</tr>
<tr>
<td>FNBMD</td>
<td>Intergenic</td>
<td>-0.0092</td>
<td>0.0015</td>
<td>-2.50E-08</td>
<td>4.02E-09</td>
<td>-5.3267</td>
<td>2.3578</td>
</tr>
<tr>
<td>FNBMD</td>
<td>No ASE</td>
<td>0.0294</td>
<td>0.0109</td>
<td>6.41E-09</td>
<td>2.37E-09</td>
<td>1.3671</td>
<td>0.1246</td>
</tr>
<tr>
<td>HB</td>
<td>cASE/iASE</td>
<td>0.0025</td>
<td>0.0003</td>
<td>4.86E-08</td>
<td>4.93E-09</td>
<td>6.7590</td>
<td>0.0150</td>
</tr>
<tr>
<td>HB</td>
<td>ASE</td>
<td>0.0001</td>
<td>0.0006</td>
<td>6.81E-10</td>
<td>3.83E-09</td>
<td>0.0946</td>
<td>0.5336</td>
</tr>
<tr>
<td>HB</td>
<td>Intergenic</td>
<td>-0.0053</td>
<td>0.0008</td>
<td>-1.43E-08</td>
<td>2.19E-09</td>
<td>-1.9883</td>
<td>0.3669</td>
</tr>
<tr>
<td>HB</td>
<td>No ASE</td>
<td>0.0398</td>
<td>0.0062</td>
<td>8.67E-09</td>
<td>1.34E-09</td>
<td>1.2058</td>
<td>0.0297</td>
</tr>
<tr>
<td>HDL</td>
<td>cASE/iASE</td>
<td>0.0013</td>
<td>0.0002</td>
<td>2.57E-08</td>
<td>4.24E-09</td>
<td>2.6991</td>
<td>0.5198</td>
</tr>
<tr>
<td>HDL</td>
<td>ASE</td>
<td>0.0166</td>
<td>0.0029</td>
<td>1.06E-07</td>
<td>1.81E-08</td>
<td>11.1082</td>
<td>1.8199</td>
</tr>
<tr>
<td>HDL</td>
<td>Intergenic</td>
<td>-0.0059</td>
<td>0.0006</td>
<td>-1.61E-08</td>
<td>1.59E-09</td>
<td>-1.6912</td>
<td>0.2039</td>
</tr>
<tr>
<td>HDL</td>
<td>No ASE</td>
<td>0.0371</td>
<td>0.0067</td>
<td>8.09E-09</td>
<td>1.46E-09</td>
<td>0.8505</td>
<td>0.0593</td>
</tr>
<tr>
<td>HEIGHT</td>
<td>cASE/iASE</td>
<td>0.0017</td>
<td>0.0001</td>
<td>3.18E-08</td>
<td>2.84E-09</td>
<td>1.3101</td>
<td>0.1387</td>
</tr>
<tr>
<td>HEIGHT</td>
<td>ASE</td>
<td>0.0109</td>
<td>0.0008</td>
<td>6.91E-08</td>
<td>4.91E-09</td>
<td>2.8493</td>
<td>0.2311</td>
</tr>
<tr>
<td>HEIGHT</td>
<td>Intergenic</td>
<td>-0.0080</td>
<td>0.0007</td>
<td>-2.17E-08</td>
<td>1.80E-09</td>
<td>-0.8942</td>
<td>0.0718</td>
</tr>
</tbody>
</table>

Continued on next page...
Table 3 – continued from previous page

<table>
<thead>
<tr>
<th>GWAS Trait</th>
<th>Category</th>
<th>PVE</th>
<th>Std Err</th>
<th>$\sigma^2$</th>
<th>Std Err</th>
<th>Enrichment</th>
<th>Std Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEIGHT</td>
<td>No ASE</td>
<td>0.1205</td>
<td>0.0065</td>
<td>2.63E-08</td>
<td>1.41E-09</td>
<td>1.0854</td>
<td>0.0091</td>
</tr>
<tr>
<td>LDL</td>
<td>cASE/iASE</td>
<td>0.0021</td>
<td>0.0004</td>
<td>3.96E-08</td>
<td>6.75E-09</td>
<td>5.8858</td>
<td>2.0932</td>
</tr>
<tr>
<td>LDL</td>
<td>ASE</td>
<td>0.0044</td>
<td>0.0009</td>
<td>2.81E-08</td>
<td>5.89E-09</td>
<td>4.1724</td>
<td>1.6154</td>
</tr>
<tr>
<td>LDL</td>
<td>Intergenic</td>
<td>-0.0043</td>
<td>0.0008</td>
<td>-1.18E-08</td>
<td>2.24E-09</td>
<td>-1.7451</td>
<td>0.3600</td>
</tr>
<tr>
<td>LDL</td>
<td>No ASE</td>
<td>0.0327</td>
<td>0.0116</td>
<td>7.11E-09</td>
<td>2.53E-09</td>
<td>1.0563</td>
<td>0.0593</td>
</tr>
<tr>
<td>LSBMD</td>
<td>cASE/iASE</td>
<td>0.0030</td>
<td>0.0005</td>
<td>5.79E-08</td>
<td>1.03E-08</td>
<td>13.8333</td>
<td>6.5121</td>
</tr>
<tr>
<td>LSBMD</td>
<td>ASE</td>
<td>0.0029</td>
<td>0.0012</td>
<td>1.86E-08</td>
<td>7.65E-09</td>
<td>4.4363</td>
<td>2.8557</td>
</tr>
<tr>
<td>LSBMD</td>
<td>Intergenic</td>
<td>-0.0074</td>
<td>0.0016</td>
<td>-2.01E-08</td>
<td>4.39E-09</td>
<td>-4.7997</td>
<td>2.3134</td>
</tr>
<tr>
<td>LSBMD</td>
<td>No ASE</td>
<td>0.0231</td>
<td>0.0110</td>
<td>5.04E-09</td>
<td>2.40E-09</td>
<td>1.2022</td>
<td>0.1098</td>
</tr>
<tr>
<td>MCH</td>
<td>cASE/iASE</td>
<td>0.0091</td>
<td>0.0007</td>
<td>1.73E-07</td>
<td>1.34E-08</td>
<td>8.6428</td>
<td>0.3575</td>
</tr>
<tr>
<td>MCH</td>
<td>ASE</td>
<td>0.0007</td>
<td>0.0009</td>
<td>4.71E-09</td>
<td>5.99E-09</td>
<td>0.2355</td>
<td>0.3093</td>
</tr>
<tr>
<td>MCH</td>
<td>Intergenic</td>
<td>-0.0033</td>
<td>0.0013</td>
<td>-8.88E-09</td>
<td>3.47E-09</td>
<td>-0.4442</td>
<td>0.1560</td>
</tr>
<tr>
<td>MCH</td>
<td>No ASE</td>
<td>0.0968</td>
<td>0.0123</td>
<td>2.11E-08</td>
<td>2.68E-09</td>
<td>1.0552</td>
<td>0.0164</td>
</tr>
<tr>
<td>MCHC</td>
<td>cASE/iASE</td>
<td>0.0005</td>
<td>0.0002</td>
<td>9.88E-09</td>
<td>4.47E-09</td>
<td>-1.2357</td>
<td>0.6014</td>
</tr>
<tr>
<td>MCHC</td>
<td>ASE</td>
<td>-0.0003</td>
<td>0.0006</td>
<td>-1.91E-09</td>
<td>3.85E-09</td>
<td>0.2395</td>
<td>0.4845</td>
</tr>
<tr>
<td>MCHC</td>
<td>Intergenic</td>
<td>-0.0039</td>
<td>0.0008</td>
<td>-1.06E-08</td>
<td>2.17E-09</td>
<td>1.3290</td>
<td>0.3229</td>
</tr>
<tr>
<td>MCHC</td>
<td>No ASE</td>
<td>-0.0376</td>
<td>0.0047</td>
<td>-8.19E-09</td>
<td>1.03E-09</td>
<td>1.0251</td>
<td>0.0302</td>
</tr>
<tr>
<td>MCV</td>
<td>cASE/iASE</td>
<td>0.0063</td>
<td>0.0005</td>
<td>1.21E-07</td>
<td>9.74E-09</td>
<td>5.8436</td>
<td>0.3254</td>
</tr>
<tr>
<td>MCV</td>
<td>ASE</td>
<td>0.0019</td>
<td>0.0007</td>
<td>1.22E-08</td>
<td>4.50E-09</td>
<td>0.5892</td>
<td>0.2407</td>
</tr>
<tr>
<td>MCV</td>
<td>Intergenic</td>
<td>-0.0036</td>
<td>0.0011</td>
<td>-9.80E-09</td>
<td>2.88E-09</td>
<td>-0.4737</td>
<td>0.1246</td>
</tr>
<tr>
<td>MCV</td>
<td>No ASE</td>
<td>0.1023</td>
<td>0.0102</td>
<td>2.23E-08</td>
<td>2.23E-09</td>
<td>1.0774</td>
<td>0.0138</td>
</tr>
<tr>
<td>MPV</td>
<td>cASE/iASE</td>
<td>0.0032</td>
<td>0.0005</td>
<td>6.19E-08</td>
<td>9.40E-09</td>
<td>2.9909</td>
<td>0.6316</td>
</tr>
<tr>
<td>MPV</td>
<td>ASE</td>
<td>0.0251</td>
<td>0.0039</td>
<td>1.60E-07</td>
<td>2.46E-08</td>
<td>7.7145</td>
<td>1.4648</td>
</tr>
<tr>
<td>MPV</td>
<td>Intergenic</td>
<td>-0.0114</td>
<td>0.0017</td>
<td>-3.09E-08</td>
<td>4.66E-09</td>
<td>-1.4916</td>
<td>0.2313</td>
</tr>
<tr>
<td>MPV</td>
<td>No ASE</td>
<td>0.0899</td>
<td>0.0153</td>
<td>1.96E-08</td>
<td>3.34E-09</td>
<td>0.9475</td>
<td>0.0524</td>
</tr>
<tr>
<td>PCV</td>
<td>cASE/iASE</td>
<td>0.0019</td>
<td>0.0003</td>
<td>3.60E-08</td>
<td>5.12E-09</td>
<td>6.1117</td>
<td>0.9515</td>
</tr>
<tr>
<td>PCV</td>
<td>ASE</td>
<td>-0.0015</td>
<td>0.0006</td>
<td>-9.42E-09</td>
<td>3.90E-09</td>
<td>-1.5979</td>
<td>0.7217</td>
</tr>
<tr>
<td>PCV</td>
<td>Intergenic</td>
<td>-0.0072</td>
<td>0.0009</td>
<td>-1.94E-08</td>
<td>2.39E-09</td>
<td>-3.2921</td>
<td>0.6572</td>
</tr>
</tbody>
</table>

Continued on next page...
<table>
<thead>
<tr>
<th>GWAS Trait</th>
<th>Category</th>
<th>PVE</th>
<th>Std Err</th>
<th>$\sigma^2$</th>
<th>Std Err</th>
<th>Enrichment</th>
<th>Std Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCV</td>
<td>No ASE</td>
<td>0.0372</td>
<td>0.0057</td>
<td>8.11E-09</td>
<td>1.23E-09</td>
<td>1.3761</td>
<td>0.0595</td>
</tr>
<tr>
<td>PLT</td>
<td>cASE/iASE</td>
<td>0.0036</td>
<td>0.0003</td>
<td>6.83E-08</td>
<td>6.25E-09</td>
<td>3.5299</td>
<td>0.3717</td>
</tr>
<tr>
<td>PLT</td>
<td>ASE</td>
<td>0.0119</td>
<td>0.0013</td>
<td>7.54E-08</td>
<td>8.28E-09</td>
<td>3.8947</td>
<td>0.4394</td>
</tr>
<tr>
<td>PLT</td>
<td>Intergenic</td>
<td>-0.0089</td>
<td>0.0007</td>
<td>-2.40E-08</td>
<td>1.99E-09</td>
<td>-1.2390</td>
<td>0.1022</td>
</tr>
<tr>
<td>PLT</td>
<td>No ASE</td>
<td>0.0934</td>
<td>0.0064</td>
<td>2.04E-08</td>
<td>1.39E-09</td>
<td>1.0520</td>
<td>0.0183</td>
</tr>
<tr>
<td>RBC</td>
<td>cASE/iASE</td>
<td>0.0015</td>
<td>0.0003</td>
<td>2.87E-08</td>
<td>5.11E-09</td>
<td>1.8908</td>
<td>0.3775</td>
</tr>
<tr>
<td>RBC</td>
<td>ASE</td>
<td>0.0004</td>
<td>0.0007</td>
<td>2.29E-09</td>
<td>4.56E-09</td>
<td>0.1509</td>
<td>0.3031</td>
</tr>
<tr>
<td>RBC</td>
<td>Intergenic</td>
<td>-0.0033</td>
<td>0.0010</td>
<td>-9.03E-09</td>
<td>2.78E-09</td>
<td>-0.5939</td>
<td>0.1727</td>
</tr>
<tr>
<td>RBC</td>
<td>No ASE</td>
<td>0.0800</td>
<td>0.0084</td>
<td>1.74E-08</td>
<td>1.84E-09</td>
<td>1.1472</td>
<td>0.0188</td>
</tr>
<tr>
<td>TC</td>
<td>cASE/iASE</td>
<td>0.0048</td>
<td>0.0005</td>
<td>9.12E-08</td>
<td>9.10E-09</td>
<td>9.6485</td>
<td>1.8248</td>
</tr>
<tr>
<td>TC</td>
<td>ASE</td>
<td>0.0043</td>
<td>0.0009</td>
<td>2.73E-08</td>
<td>5.69E-09</td>
<td>2.8829</td>
<td>0.8570</td>
</tr>
<tr>
<td>TC</td>
<td>Intergenic</td>
<td>-0.0063</td>
<td>0.0007</td>
<td>-1.69E-08</td>
<td>2.02E-09</td>
<td>-1.7925</td>
<td>0.2413</td>
</tr>
<tr>
<td>TC</td>
<td>No ASE</td>
<td>0.0460</td>
<td>0.0104</td>
<td>1.00E-08</td>
<td>2.27E-09</td>
<td>1.0614</td>
<td>0.0358</td>
</tr>
<tr>
<td>TG</td>
<td>cASE/iASE</td>
<td>0.0037</td>
<td>0.0005</td>
<td>7.06E-08</td>
<td>9.64E-09</td>
<td>7.9232</td>
<td>0.0172</td>
</tr>
<tr>
<td>TG</td>
<td>ASE</td>
<td>0.0152</td>
<td>0.0034</td>
<td>9.67E-08</td>
<td>2.18E-08</td>
<td>10.8507</td>
<td>2.8737</td>
</tr>
<tr>
<td>TG</td>
<td>Intergenic</td>
<td>-0.0062</td>
<td>0.0008</td>
<td>-1.68E-08</td>
<td>2.04E-09</td>
<td>-1.8806</td>
<td>0.3053</td>
</tr>
<tr>
<td>TG</td>
<td>No ASE</td>
<td>0.0333</td>
<td>0.0098</td>
<td>7.26E-09</td>
<td>2.13E-09</td>
<td>0.8149</td>
<td>0.0807</td>
</tr>
</tbody>
</table>
Environmental exposures change chromatin accessibility

While gene expression changes in response to the environment have been widely studied, there has been little exploration into changes at the chromatin level. Based on transcriptional responses to 50 environmental perturbations (Moyerbrailean et al., 2016a) in 5 cell types (250 conditions), we chose 3 treatments (dexamethasone, copper, and selenium) that the cells had a strong response to. We then used ATAC-seq to map condition specific regulatory regions and variants with allele specific hypersensitivity in LCLs with these treatments and their respective controls (water and ethanol). We identified differential ac-
cessible regions (DeSeq) for each treatment, with the majority found with dexamethasone (9,263 at FDR 10%), followed by copper (2,615), and selenium (2,115) (Figure 15A).

![Graph showing observed vs expected p-values for ATAC-seq data]

Figure 15: Differential hypersensitive ATAC-seq. Log transformed p-values of ATAC-seq genomic tiles measured for differential deviation of dexamethasone, copper, and selenium from the control. B) ATAC-seq footprint for CRE in copper treated vs control LCLs. Venn diagram shows the number of bound sites in treated vs control. C) ATAC-seq footprint for CRE in dexamethasone treated vs control LCLs. Venn diagram shows the number of bound sites in treated vs control.

ATAC-seq data along with sequence-based transcription factor motif models can be further analyzed to identify the transcription factors involved in response to treatment, using Centipede (Pique-Regi et al., 2011) to perform footprinting analysis. By identifying the relevant TFs, we can identify the mechanisms behind many common traits. Using Centipede’s hierarchical prior, we then did a likelihood ratio test for condition-specific TF activity. For this, we fit the two conditions (treatment and control) simultaneously.
and extended Centipede hierarchical prior to detect factors with across all genome has a change in activity. We identified about 100 TF footprints each condition, with a subset then being condition specific (Table 4). TFs with the CRE motif, for example, were found to have depleted binding sites after treating with copper ($p < 0.001$), while in Dex there is no change ($p > 0.13$, Figure 15). We can see a decrease of binding sites, but also bound sites would show a weaker footprint in these aggregate plots which is indicative of less binding.

Table 4: Number of identified factors/differential factors from ATAC-seq in LCLs.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Active Factors</th>
<th>Diff. Active ($p&lt;0.005$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dexamethasone</td>
<td>171</td>
<td>26</td>
</tr>
<tr>
<td>Copper</td>
<td>91</td>
<td>19</td>
</tr>
<tr>
<td>Selenium</td>
<td>147</td>
<td>18</td>
</tr>
<tr>
<td>Ctrl 1 (vs. Ctrl 2)</td>
<td>124</td>
<td>0</td>
</tr>
</tbody>
</table>

Identifying genetic variants showing allele specific effects to environmental perturbation

To identify GxE that may explain inter-individual differences in disease risk conditional on specific environmental exposures, we used a modified version of BiT-STARR-seq. For this we used the same oligo library described in Kalita et al. (2018b), cloned now into the STARR-seq human plasmid (Arnold et al., 2013) (now 48,958 constructs) and tested in LCLs in triplicate. We selected the treatments dexamethasone, retinoic acid, selenium, and caffeine as LCLs showed the greatest response to these in previous work (Moyerbrailean et al., 2016a). We first assessed the extent of ASE in different cellular environments, using QuASAR-MPRA (Kalita et al., 2018a) to calculate the beta estimate for the allelic ratio and its standard error in each replicate. We then meta-analyzed by treatment. We successfully identified ASE in each condition, with the largest number of significant ASE (FDR 10%) in selenium (3,790) (Figure 16).
To then test the hypothesis that differential accessibility is the mechanism driving the conditional ASE, we looked for enrichment of the DARs identified through ATAC-seq with the ASE results (matching by treatment). We found a much farther departure from the null for the ASE $p$-values from BiT-STARR-seq in DARs as compared to the overall distribution (Figure 16B). When we narrow this down to footprints, there is enrichment for ASE (from BiT-STARR-seq) in TF footprints (from the ATAC-seq data) matched by treatment, in both dexamethasone and selenium (Figure 17). For example, we find enrichment for ASE in CNOT3 footprints (which is involved in early b-cell dev. (Inoue et al., 2015)) when LCLs are treated with dexamethasone. While we lack a good footprint model for GR, we can look for downstream targets. In an RNA-seq study done in our lab, we find that CNOT3 is downregulated in response to dexamethasone in LCLs (Moyerbrailean et al., 2016a).
therefore successfully capture downstream targets of GR activation. CNOT3 has also been linked to repression of other nuclear receptors (ie/Erα) (Winkler et al., 2006).

Figure 17: Enrichment for ASE in TF footprints matched by treatment. Barplot showing the odds ratio (y-axis) from Fisher’s test for ASE in the TF footprint from ATAC-seq (x-axis) in dexamethasone (left) and selenium (right). * indicates enriched, ** indicates significantly enriched after Bonferroni correction. Error bars indicate the CI from the Fisher’s test.

Discussion

In this study, we analyzed heritability, chromatin accessibility, and regulatory variants with environmental exposures to in-vitro treatments. We found strong enrichment for variants affecting GxE in many complex traits using ASE in RNA-seq and eQTLs from two different studies. These data confirm the interaction of genetics and the environment for GWAS and will aid future experiments by highlighting environmental variables that should be considered in any complex trait studies. By analyzing regulatory regions and
the variants within, in response to environmental stimuli, we can begin to understand the impact of compounds to which we are daily exposed.

While the treatments investigated in this study cause significant changes in gene expression, chromatin accessibility, and allelic expression, they act through varied pathways. Dexamethasone and retinoic acid are steroid hormones whose action as an anti-inflammatory factor at the molecular and transcriptional level has been well-established. Caffeine has been identified as inhibiting the breakdown of cAMP and selenium has an important role as a cofactor. Copper is required for oxygen metabolism and has been linked to decreased effector activities of B cells and neutrophils (Prohaska and Failla, 1993; Djoko et al., 2015). However, the effect of these environments on heritability and the regulatory response has been less characterized. We have annotated footprints and treatment response factors in LCLs. We were able to confirm that these regulatory regions affect expression in an allele specific manner when analyzing ATAC-seq DARs with BiT-STARR-seq. Additionally, we confirm with dexamethasone that our approach identifies biologically relevant response factors, as we found enrichment of a known downstream target (CNOT3) of the glucocorticoid pathway for ASE. By using BiT-STARR-seq to investigate allelic effects, we are no longer limited to looking at naturally occurring variation, as seen in re-QTL, conditional ASE, or conditional allele specific hypersensitivity studies. This therefore represents a cost-effective approach to rapidly screen the phenotypic effect of genetic and environmental factors in a variety of contexts. In combination with the computational tools our lab has developed, this approach is a unique resource to functionally characterize genetic variants in specific cellular environment at an unprecedented and expandable scale.
CHAPTER 5 CONCLUSION

Genome-wide association studies (GWAS) have identified a large number of genetic variants associated with disease as well as normal phenotypic variation for complex traits such as height (Mailman et al., 2007). However, challenges remain in determining the functional relevance of human DNA sequence variants (Ward and Kellis, 2012). GWAS identify large regions of association and in general, cannot directly pinpoint the true causative variant. Second, even after fine mapping, most variants are located in non-coding regions making it difficult to infer mechanisms linking individual genetic variants with the disease trait. In addition, we do not know under which environmental conditions the sequence variants have a functional impact, and whether they become one of many factors involved in complex phenotypes at the organismal level. Throughout this dissertation, I sought to address these problems and to develop methods to better study allele specific effects and their link to complex phenotypes.

In Chapter 2, I used computational methods to predict causal GWAS variants, validated a subset for ASE using a traditional reporter assay, and developed a method to identify ASE in high throughput validations. Many high throughput reporter assays validated only a small number of regulatory regions due to experimentally testing many regions with only a small amount of prior expectation of activity. By computationally narrowing down the regions likely to have an effect on expression, we were able to identify about 50% of our predicted regions as being enhancer/repressor. Additionally, the rapid development of MPRA and STARR-seq to screen regulatory regions meant that no one had taken the time to develop a well-calibrated and powerful method to identify ASE in these studies. QuASAR-MPRA was developed to tackle this problem, and we found it to be low in inflation
(bias) and in every simulated condition (variance in the data, effect size of ASE, number of replicates, and sequencing depth) it out performed the $t$-test (used previously in MPRA). This method will continue to be useful as more researchers begin using high throughput assays to identify allelic effects.

In Chapter 3 I develop a new modification of STARR-seq in order to streamline the assay and improve power to detect ASE through the addition of an UMI. By spending the effort to optimize the assay conditions, we make the assay easier for researchers to quickly and efficiently validate regulatory variants of interest. One advantage to our approach is in the synthesis of the oligos. This allows for testing any variant of interest, without relying on natural variation so allowing for testing of rare variants which is not typically possible in QTL studies. Another advantage is in adding the UMI after transfection of the oligos into the cells. This avoids complexity bottlenecks in cloning, transformation, and transfection efficiencies that often occur in MPRA. When these bottlenecks occur, researchers generally greatly increase for example, the number of cells transfected, making the assay no longer feasible for many researchers. The UMI also allows us to account for PCR duplicates in the library preparation stage and so have increased power to detect ASE. Additionally, by integrating BiT-STARR-seq with a high throughput allele-specific EMSA, we are able to identify the mechanism behind many ASE variants.

Studying GxE in human studies is extremely difficult, so our approach of using an *in-vitro* method and modeling molecular phenotypes is a useful alternative. Using GEMMA, we were able to identify environments that were enriched for complex traits. This shows an increased relevance for complex traits when using conditional allele-specific information, which confirms the significance of GxE in complex traits. With ATAC-seq data we were able
to identify differentially accessible regions, TF footprints, and differential TF footprints. Integrating this data with BiT-STARR-seq, we were able to identify enrichment for these differential chromatin accessibility regions with ASE.

Using the information generated through my research can lead to understanding the genetic and molecular basis of inter-individual differences in complex traits. For instance, genetic variants that cause altered gene expression can lead to development of these traits at the organismal level. Consequently, it is important to not only identify true gene-regulatory regions but also to test if specific genetic variants within these regions affect gene expression. In addition to tightly regulating baseline gene expression levels, these genetic and epigenetic controls determine the transcriptional response to external stimuli. Together, in my research I’ve used integration of computational predictions with experimental validation to identify allelic effects. This design is a useful approach to validate the molecular mechanism for specific transcription factors, and link these to the context of human health. We hope this work may lead to actionable therapies in the future.
APPENDIX A

We integrated our Centipede footprint annotations into the combined models learned in Pickrell (2014) for GWAS meta-studies corresponding to 18 traits (Table 5), using the fgwas command line program. We assessed enrichment or depletion for footprint annotations using the \( \log_2(\text{enrichment}) \) values, excluding any motifs whose 95% confidence interval (CI) spanned zero. For each TF motif whose binding sites are either significantly enriched or depleted for trait-associated SNPs (Figure 18), we examined the SNPs whose posterior probability of association (PPA) with a trait had been increased by the addition of our annotation. Overall we found 88 unique SNPs whose associations were strengthened by our footprint annotations (Table 7).

Validating putative causal SNPs by reporter gene assays

From the 88 SNPs identified by the fgwas analysis, we considered GWAS-relevant effect-SNPs located in active footprints in LCLs (the cell line used for transfection) and ranked them on the Spearman correlation coefficient in Moyerbrailean et al. (2016b). We initially selected the top 25 SNPs with a positive correlation, but the assays for 4 of them failed for several technical reasons (e.g., cloning step failed). To validate the predicted allelic effects on gene expression for the remaining 21 SNPs, we first constructed inserts containing the reference or alternate allele for each SNP of interest. Each region was amplified from genomic DNA extracted from LCLs (Coriell). Primers were designed using the Infusion Clontech online primer design tool for inserts containing the SNP of interest ±100bp. Primers were ordered from IDT technologies. Inserts were amplified by PCR and
pGL4.23 plasmid was linearized (inverse PCR) using Clontech Hi-fi PCR premix and following the manufacturer’s instructions. PCR products and linearized plasmid were resolved on agarose gel, excised and purified using Nucleospin gel extraction and PCR cleanup kit (Clontech). Inserts were cloned into linearized pGL4.23 using the Infusion Cloning HD kit (Clontech). Transformation was done using Stellar Competent cells (Clontech) and DNA was extracted from selected colonies using the PureYield kit (Promega). The allelic status and the absence of artifactual mutations of each clone was validated by Sanger sequencing performed by Genewiz. Transfections were performed into GM18507 using the standard protocol for the Nucleofector electroporation (Lonza). After 10 hours we measured Firefly and Renilla (transfection control) luciferase activity using the Dual-Glo Luciferase Assay Kit (Promega) on the GloMax instrument (Promega). Luciferase activity was measured for up to 20 replicate experiments. We then used a t-test to identify significant differences in the expression of the reporter gene, calculated by the ratio of the firefly to the renilla activity, normalized to the ratio of the activity in the untransfected cells. We contrasted the activity of each construct to the pGL4.23 vector, to assess enhancer/repressor activity of each region. To evaluate allele-specific effects, we contrasted the activity of the reference allele to the alternate allele for each region. These results are summarized in Table 8 and in Figure 7.
Table 5: **Summary of GWAS meta analysis traits examined.** Shown for each trait is the trait abbreviation and the citation for the meta analysis study.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Trait</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>HB</td>
<td>Hemoglobin levels</td>
<td>van der Harst, P., et al. (2012). Nature 492, 369-375</td>
</tr>
<tr>
<td>HDL</td>
<td>HDL cholesterol levels</td>
<td>Teslovich, T.M., et al. (2010). Nature 466, 707-713</td>
</tr>
<tr>
<td>Height</td>
<td>Height</td>
<td>Lango Allen, H., et al. (2010). Nature 467, 832-838</td>
</tr>
<tr>
<td>MCH</td>
<td>Mean red blood cell hemoglobin</td>
<td>van der Harst, P., et al. (2012). Nature 492, 369-375</td>
</tr>
<tr>
<td>MCHC</td>
<td>Mean corpuscular hemoglobin</td>
<td>van der Harst, P., et al. (2012). Nature 492, 369-375</td>
</tr>
<tr>
<td>MCV</td>
<td>Mean red blood cell volume</td>
<td>van der Harst, P., et al. (2012). Nature 492, 369-375</td>
</tr>
<tr>
<td>PCV</td>
<td>Packed red blood cell volume</td>
<td>van der Harst, P., et al. (2012). Nature 492, 369-375</td>
</tr>
<tr>
<td>TC</td>
<td>Total cholesterol levels</td>
<td>Teslovich, T.M., et al. (2010). Nature 466, 707-713</td>
</tr>
</tbody>
</table>

Table 6: **Factor binding sites enriched for GWAS SNPs.** For each trait, factors whose binding sites are enriched for SNPs associated with the trait are listed. Shown also are the lower and upper limits of the 95% confidence interval.

Available at [http://genome.grid.wayne.edu/centisnps/supp/](http://genome.grid.wayne.edu/centisnps/supp/)

Table 7: **SNPs associated with GWAS traits that fall in Centipede-predicted TF binding sites.** PPA, Posterior probability of association estimated by fgwas for each SNP. "Before" indicates the PPA from the base model, "after" indicates the PPA after adding footprint annotations to the model. The p-values listed are derived from the z-scores that are used as input for fgwas.

Available at [https://doi.org/10.1371/journal.pgen.1005875.s020](https://doi.org/10.1371/journal.pgen.1005875.s020)

Table 8: **Reporter gene assay results.** For each of the SNPs tested, listed are the results for the reference allele (top) and the alternate allele (bottom). Shown is the average and standard error (across replicates) of the firefly luciferase activity normalized to the renilla luciferase activity, for each construct (Norm Expr) and for the pGL4.23 vector (Empty Vector). The last two columns are the t-test p-values comparing the activity of the reference allele to the alternate allele (vs ref), and of each allele to the pGL4.23 vector (vs empty). Underlined alleles indicate the allele predicted to have stronger binding.

Available at [https://doi.org/10.1371/journal.pgen.1005875.s021](https://doi.org/10.1371/journal.pgen.1005875.s021)
Figure 18: Enrichment of transcription factors motifs from fgwas. Shown are the log2(enrichment) values with 95% confidence intervals for each factor whose binding sites are enriched for SNPs associated with the traits in Table 14. x-axis is truncated at 10 for ease of display.
Figure 19: Association plots identifying SNPs in footprints. Log Bayes factor (top) and posterior probabilities (bottom) of association to the indicated trait for all genetic variants in the regions containing rs4519508 and rs532436.
Figure 20: Association plots identifying SNPs in footprints from fgwas. For each SNP in Table 13, two plots show the log Bayes factor (top) and posterior probabilities (bottom) of association to the indicated trait for all genetic variants in the region containing the SNP. This Figure continues to next page . . .
Figure 20 ... continued from previous page

Figure 20 continues to next page ...
Figure 20 . . . continued from previous page

Figure 20 continues to next page . . .
Figure 20 continues from previous page.

Figure 20 continues to next page...
Figure 20... continued from previous page

Figure 20 continues to next page...
Figure 20 . . . continued from previous page

Figure 20 continues to next page . . .
Figure 20 continues from previous page

Figure 20 continues to next page . . .
Figure 20... continued from previous page

Figure 20 continues to next page...
Figure 20 ... continued from previous page

Figure 20 continues to next page ...
Figure 20.continued from previous page

Figure 20 continues to next page...
Figure 20 . . . continued from previous page
Figure 20 continues from previous page

Figure 20 continues to next page...
Figure 20 ... continued from previous page

Figure 20 continues to next page ...
Figure 20 . . . continued from previous page
### Table 9: Annotations Used
Experimental and computational annotations used for overlap with the MPRA dataset.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Description</th>
<th>Publication</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASH</td>
<td>Variants with allele-specific hypersensitivity (ASH) in DNase-seq data</td>
<td>Moyerbrailean et al 2016</td>
<td>genome.grid.wayne.edu/centisnps/</td>
</tr>
<tr>
<td>LCL dsQTLs</td>
<td>Variants associated with DnaseI hypersensitivity</td>
<td>Degner et al 2012</td>
<td>eqtl.uchicago.edu</td>
</tr>
<tr>
<td>GTEx eQTL</td>
<td>Variants associated with gene expression changes</td>
<td>Consortium et al 2015</td>
<td><a href="http://www.gtexportal.org">www.gtexportal.org</a></td>
</tr>
<tr>
<td>Effect-SNPs</td>
<td>Variants predicted to affect TF binding</td>
<td>Moyerbrailean et al 2016</td>
<td>genome.grid.wayne.edu/centisnps</td>
</tr>
<tr>
<td>GKM-svm</td>
<td>Predicts the impact of SNPs on DNase I sensitivity (dsQTLs) from gapped k-mers</td>
<td>Lee et al 2015</td>
<td><a href="http://www.beerlab.org/deltasvm/downloads">www.beerlab.org/deltasvm/downloads</a></td>
</tr>
<tr>
<td>CATO</td>
<td>Predicts the effect of SNPs on the energy of TF binding</td>
<td>Maurano et al 2015</td>
<td>Supplement Table</td>
</tr>
<tr>
<td>DeepSEA</td>
<td>Deep learning-based algorithmic framework for predicting the chromatin effects of sequence alterations</td>
<td>Zhou et al 2015</td>
<td>deepsea.princeton.edu/job/analysis/create</td>
</tr>
</tbody>
</table>

### Table 10: QuASAR Results for LCLs
QuASAR p-values for MPRA LCL counts data. File is LCL_quasar_pval_comb.csv


### Table 11: QuASAR Results for HepG2
QuASAR p-values for MPRA HepG2 counts data. File is hepg2_quasar_pval.csv


### Table 12: Logistic Model predictions for LCLs
p-value results from running the logistic model. In all cases where scores had both negative and positive values, absolute scores were used.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GKM-svm</td>
<td>$1.11 \times 10^{-02}$</td>
<td>$1.53 \times 10^{-09}$</td>
</tr>
<tr>
<td>Effect-SNP</td>
<td>$2.75 \times 10^{-02}$</td>
<td>$1.71 \times 10^{-15}$</td>
</tr>
<tr>
<td>LCL dsQTL</td>
<td>$3.10 \times 10^{-02}$</td>
<td>$1.46 \times 10^{-03}$</td>
</tr>
<tr>
<td>CATO</td>
<td>$5.75 \times 10^{-01}$</td>
<td>$2.48 \times 10^{-03}$</td>
</tr>
<tr>
<td>DeepSEA</td>
<td>$-4.41 \times 10^{-02}$</td>
<td>$2.17 \times 10^{-02}$</td>
</tr>
<tr>
<td>GTEx lead SNP</td>
<td>$7.35 \times 10^{-04}$</td>
<td>$7.59 \times 10^{-01}$</td>
</tr>
</tbody>
</table>

### Table 13: Logistic Model predictions for HepG2
p-value results from running the logistic model. In all cases where scores had both negative and positive values, absolute scores were used.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GKM-svm</td>
<td>$-2.07 \times 10^{-04}$</td>
<td>$6.57 \times 10^{-01}$</td>
</tr>
<tr>
<td>Effect-SNP</td>
<td>$1.45 \times 10^{-02}$</td>
<td>$1.18 \times 10^{-10}$</td>
</tr>
<tr>
<td>LCL dsQTL</td>
<td>$1.37 \times 10^{-02}$</td>
<td>$2.88 \times 10^{-02}$</td>
</tr>
<tr>
<td>CATO</td>
<td>$-8.51 \times 10^{-02}$</td>
<td>$4.55 \times 10^{-01}$</td>
</tr>
<tr>
<td>DeepSEA</td>
<td>$-1.64 \times 10^{-02}$</td>
<td>$7.95 \times 10^{-02}$</td>
</tr>
<tr>
<td>GTEx lead SNP</td>
<td>$-8.65 \times 10^{-06}$</td>
<td>$9.95 \times 10^{-01}$</td>
</tr>
</tbody>
</table>
Comparison to QuASAR

We compared our QuASAR-MPRA implementation to the original QuASAR method. We show that while QuASAR may perform well when the DNA proportion distribution is closely centered on 0.5, but not when the proportion is different than 0.5 (Figure 26). While the average DNA proportion for this dataset is 0.499, the range is considerable (0.014 to 0.974), so a method that can handle a wide range of DNA proportions is necessary.

Reverse oligo results

Based on the genomics inflation values, the QuASAR test results in the lowest inflation. The results provided in the main text are for the forward strand constructs but we obtain very similar genomic inflation values in the reverse strand. However we only found 22 SNPs that were significant in both the forward and reverse orientations. This could be as both directions are not equally sampled but it would be interesting to do a more detailed analysis of the mechanism behind the differences seen between the two directions in the future. QuASAR lambda is 1.17 LCL/ 0.76 HepG2, while the binomial test produces the greatest inflation, with lambda values of 56.77 LCL/ 27.07 HepG2. Two tests previously used in high throughput reporter assay, result in intermediate levels of inflation, between these two extremes. A paired $t$-test with independent estimation of variance and Welch’s adjustment, as in Tewhey et al., results in lambda values of 2.66 LCL/ 2.85 HepG2; while Fisher’s exact test, as in (Vockley et al., 2015) results in lambda values of 38.74 LCL/ 9.95 HepG2. In addition, looking at genomic inflation values for only non-effect SNPs, QuASAR lambda is 1.22 LCL/ 0.79 HepG2, the binomial test lambda is 53.18 LCL/ 22.60 HepG2,
t-test lambda is 3.12 LCL/3.36 HepG2, and Fisher’s exact test lambda is 35.98 LCL/11.35 HepG2.

Figure 21: Comparing ASE testing methods in HepG2 from Tewhey et al. (2016). QQplot depicting the p-value distributions from testing for ASE using four different methods in HepG2 for all SNPs (Left) or non-effect SNPs only (Right). $\lambda$ measures genomic inflation deviation from the uniform.

Figure 22: Validating experimental annotations in HepG2. QQ plot depicting the p-value distributions from testing for ASE using QuASAR, overlapping with experimental genomic annotations in HepG2. An annotation enrichment p-value is reported next to their labels, but only for those annotations that are significantly enriched for small QuASAR-MPRA p-values according to the logistic model (see Methods).

Figure 23: Validating computational genomic annotations in HepG2. QQ plot depicting the p-value distributions from testing for ASE using QuASAR, overlapping with computational genomic annotations in HepG2. Effect-SNP scores have a threshold of $<-3$ or $>3$. CATO Maurano et al. (2015) prediction scores have a threshold of $>0.1$. GKM-svm Lee et al. (2015a) gapped kmer sequence-based computational method to predict the effect of regulatory variation has a threshold of $<-6$ or $>6$. DeepSEA Zhou and Troyanskaya (2015) predicts genomic variant effects at the variant position using deep learning-based algorithmic framework. The functional significance predictions have a threshold of $<0.05$. An annotation enrichment p-value is reported next to their labels, but only for those annotations that are significantly enriched for small QuASAR-MPRA p-values according to the logistic model (see Methods).
Figure 24: **Identifying cell type effects in LCLs.** QQ plot depicting the p-value distributions from testing for ASE using QuASAR, overlapping with effect-SNP annotations in LCLs. LCL effect-SNPs (blue) are variants in TFs active in LCLs, HepG2 effect-SNPs (red) are variants in TFs active in HepG2, LCL/HepG2 effect-SNPs (green) are variants in TFs active in LCLs or HepG2.

Figure 25: **Identifying cell type effects in HepG2.** QQ plot depicting the p-value distributions from testing for ASE using QuASAR, overlapping with effect-SNP annotations in HepG2. LCL effect-SNPs (blue) are variants in TFs active in LCLs, HepG2 effect-SNPs (red) are variants in TFs active in HepG2, LCL/HepG2 effect-SNPs (green) are variants in TFs active in LCLs or HepG2.

Figure 26: **Comparing QuASAR and QuASAR-MPRA ASE testing methods in LCLs.** QQplot depicting the p-value distributions from testing for ASE using QuASAR or QuASAR-MPRA in LCLs, SNPs separated by DNA proportion range to be either close to 0.5 or far away from 0.5.
APPENDIX B

Oligo selection and design

Tables 15, 16, 17 report the annotations we have considered with their sources and Table (29, 25) includes the library composition. These included: SNPs predicted to alter transcription factor binding in LCLs and HepG2 (CentiSNPs, (Moyerbrailean et al., 2015)), LCL eQTLs fine-mapped in (Wen et al., 2015), liver eQTLs (Innocenti et al., 2011), significant fgwas SNPs in transcription factor binding motifs for 18 complex traits (Moyerbrailean et al., 2016b), significant fgwas SNPs for base models of functional annotations for 18 complex traits (Pickrell, 2014), ASB SNPs, and strong enhancers with no predicted ASB (regions with footprints containing SNPs that are not predicted to affect binding of TFs) (Moyerbrailean et al., 2015). CentiSNP is an annotation that we recently developed (Moyerbrailean et al., 2016b), and that uses the Centipede framework (Pique-Regi et al., 2011) to integrate DNase-seq footprints with a recalibrated position weight matrix (PWM) model for the sequence to predict the functional impact of SNPs in footprints. SNPs in footprints “footprint-SNPs” are further categorized using Centipede hierarchical prior for each allele as “CentiSNP” if the prior relative odds for binding are $>20$. FASTA sequences with a window of 99 (on each side of the SNP) on the BED file were grabbed using seqBedFor2bit (Moyerbrailean et al., 2015), and 15bp matching sequencing primers used for Illumina NGS were added to each end. This generates an oligo containing 200bp of regulatory region with the SNP centered in the middle, with primers on both ends (Figure 27). Each regulatory region was designed to have two oligos: one for each of the alleles. A second list of the FASTA sequences without the primer ends was generated to use as a cus-
tom reference genome, then converted to fastq using fatofastq (UCSC genomics utilities). The full SNP list was aligned to the hg19 genome with BWA mem (Li, 2013), removing the regions that aligned with a quality score less than 20 (unique alignment probability > 99%). The full SNP list was also aligned to the custom reference genome, and then filtered for a quality score of 190. A total of 39,366 indexes were randomly generated to match this pattern: RDHBVDHBVD. This sequence was chosen to limit the longest possible polyACGT run at any position to 3 nucleotides, and avoid a G in the first and last position (corresponding to a dark cycle on the Illumina NextSeq 500).

**Oligo synthesis and amplification**

DNA inserts 230bp long, corresponding to 200bp of regulatory sequence, were synthesized by Agilent to contain the regulatory region and the SNP of interest within the first 150bp. We performed a first round of PCR using Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) and primers [F_transposase and R_transposase] (Table 30) with cycling conditions: 98°C for 30s, followed by 4 cycles of 98°C for 10s, 50°C for 30s, 72°C for 60s, followed by 6 cycles of 98°C for 10s, 65°C for 30s, 72°C for 60s, followed by 72°C for 5 min. This reaction was used to double strand the oligos and complete the sequencing primers. The PCR product was run on a 2% agarose gel, extracted and purified with the NucleoSpin Gel and PCR Clean-Up Kit (Clontech). A subsequent round of PCR amplified the material using the same reaction as in the first round of PCR, but with cycling conditions: 98°C for 30s, followed by 15 cycles of 98°C for 10s, 65°C for 30s, 72°C for 60s, followed by 72°C for 5min. The PCR product was purified as described above.
Cloning Regulatory regions into pGL4.23

A recent study demonstrated that the ORI can be an active promoter in pGL4.23 plasmids and can function as a stronger promoter in the absence of other promoter sequences (Muerdter et al., 2017). Here we used a design that includes a minimal promoter, thus potentially missing some signal from the weakest enhancer sequences in our library. However, as we focus on allele-specific enhancer activity, the presence of a minimal promoter in addition to the ORI, should affect both alleles similarly and should not induce false positives in the allele-specific signal.

Plasmid pGL4.23 (Promega) was linearized using CloneAmp HiFi PCR Premix (Clontech), primers [STARR_F_SH and STARR_R_SH], and 35 cycles of 98°C for 10s, 60°C for 15s, and 72°C for 5s. The PCR product was purified on a 1% agarose gel as described above. Inserts were cloned into the linear plasmid using standard Infusion (Clontech) cloning protocol. Clones were transformed into XL10-Gold Ultracompetent Cells (Agilent) in a total of 7 reactions. These reactions were pooled and grown overnight in 500ml LB at 37°C in a shaking incubator. DNA was extracted using Endofree maxiprep kit (Qiagen).

Transfection of library

Previous studies (Muerdter et al., 2017; Huerfano et al., 2013) have found that transfection, especially from nucleofection, can lead to activation of type 1 interferon response, which may complicate comparison of enhancer activities between different cell types. In our study design, allelic effects are measured and contrasted within the same cell type, thus any trans-effect is inherently controlled. Furthermore, in LCLs the immune response is already activated because of EBV transformation. DNA library was transfected into
LCLs (GM18507) using standard nucleofection protocol, program DS150, 3\(\mu\)g of DNA and 7.5\(\times\)10\(^6\) cells. A total of 3 sets of transfections were done in triplicate cuvettes, then pooled. We performed nine biological replicates of the transfection from 7 independent cell growth cultures. After transfection, cells were incubated at 37\(^{\circ}\)C and 5% CO2 in RPMI1640 with 15%FBS and 1% Gentamycin for 24h. Cell pellets were then lysed using RLT lysis buffer (Qiagen), and cryopreserved at -80\(^\circ\)C.

**Library preparation**

RNA-libraries. Thawed lysates were split in three aliquotes and total RNA was isolated using RNeasy Plus Mini Kit (Qiagen). Poly-adenylated RNA was selected using Dynabeads mRNA Direct Kit (Ambion) using the protocol for total RNA input. RNA was reverse transcribed to cDNA using Superscript III First-Strand Synthesis kit (ThermoFisher) with primer [Nextera_i7_10N] and following the manufacturer’s protocol. cDNA technical replicates were pooled and SPRI Select beads (Life Tech) were used for purification and size selection at a ratio of 0.9X. PCR Library Enrichment was performed using a nested PCR protocol. For the first round of PCR we used Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) and primers [F_trans_short and Illumina2.1] with cycling conditions: 98\(^\circ\)C for 30s, followed by 15 cycles of 98\(^\circ\)C for 10s, 72\(^\circ\)C for 15s, followed by 72\(^\circ\)C for 5 min. PCR product was purified on a 2% agarose gel as described above. The nested PCR used Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) and primers [fixed N5xx adapter (Illumina) (unique per each library replicate) and Illumina2.1] with cycling conditions: 98\(^\circ\)C for 30s, followed by 5 cycles of 98\(^\circ\)C for 10s, 72\(^\circ\)C for 15s, followed by 72\(^\circ\)C for 5 min. In a side quantitative real-time PCR reaction, 5\(\mu\)L of PCR product, 10X
SYBR Green I, and the same primers and master mix were run in conditions: 98°C for 30s, 30 cycles of 98°C for 10s, 63°C for 30s, and 72°C for 60s. To determine the number of PCR cycles needed to reach saturation, we plotted linear Rn versus cycle and determined the cycle number that corresponds to 25% of maximum fluorescent intensity on the side reaction (Buenrostro et al., 2013). The PCR product (Figure 27) was purified on a 2% agarose gel as described above.

DNA-libraries. We prepared 7 replicates of the DNA library using the PCR protocol as described in (Buenrostro et al., 2013) except using primers [fixed N5xx adapter (Illumina) (unique per each library replicate) and Nextera_i7_10N] and 30ng of input plasmid DNA. PCR product was purified on a 2% agarose gel as described above.

BiT-BUNDLE-seq

We used BiT-BUNDLE-seq, a new version of the BUNDLE-seq protocol (Levo et al., 2015). Input DNA sequences were extracted from the BiT-STARR-seq DNA plasmid library using the same PCR conditions as in preparing the DNA libraries, followed by purification on a 2% agarose gel as described above. We used N-terminal GST-tagged, recombinant human NFKB1 from EMD Millipore. The reaction buffer (0.15 M NaCl, 0.5 mM PMSF [Sigma], 1 mM BZA [Sigma], 0.5X TE, and 0.16 \( \mu \)g/\( \mu \)L PGA [Sigma]) was incubated at room temperature for 2 hours in low binding tubes (ThermoFisher). The tubes were cooled for 30 min at 4°C, and then 0.067 \( \mu \)g/\( \mu \)L BSA (Sigma) was added before adding the NFKB1 protein. One hundred nanograms of DNA were then added, and the protein and DNA were incubated for 1 h at 4°C. Experiments were performed in triplicates for each NFKB1 concentration.
The reaction mix was run with 6\(\mu\)L Ficoll (Sigma) in a 7.5% Mini-PROTEAN TGX Pre-cast 10-well Protein Gel (BIORAD) in cold 0.25X TBE buffer for 2 hours at 100V. The gel was stained for 30 min with 3X GelStar (Lonza). Bound and unbound DNA bands were excised under a blue light transilluminator. The DNA was eluted from the gel using the QIAQuick Gel Extraction Kit with a User-Developed Protocol (Qiagen QQ05). The gel slices were incubated in a diffusion buffer (0.5 M ammonium acetate, 10mM magnesium acetate, 1mM EDTA, pH 8.0 [KD Medical]; 0.1% SDS [Sigma]) at 50°C for 30 minutes. The supernatant was then passed through a disposable plastic column containing packed, siliconized glass wool [Supelco] to remove any residual polyacrylamide. Libraries were then quantified and loaded on the NextSeq 500 for sequencing.

**Library Sequencing**

Pooled RNA and DNA libraries were sequenced on the Illumina Nextseq 500 to generate 125 cycles for read 1, 30 cycles for read 2, 8 cycles for the fixed multiplexing index 2 and 10 cycles for index 1 (variable barcode). Sequencing depth for each replicate can be found in Table 31.

**Data Processing**

Reads were mapped using the HISAT2 aligner (Kim *et al.*, 2015), using the 1Kgenomes snp index so as to avoid reference bias. First we removed variants whose UMI was not possible to be present, given the UMI pattern selected. We then ran UMItools (Smith *et al.*, 2017) using standard flags, as well as a q20 filter. We then ran the deduplicated files through mpileup using a BED file of our full SNP list, the -t DP4, -g, and -d 1000000. DNA reads were processed through a counts filter (on the summed replicates) of more than
7 counts per SNP and at least one count for the reference and alternate alleles in either direction. 50,609 SNPs in the DNA library were used as input to the RNA library. The RNA library was processed following the same procedure as for the DNA library, except that the counts filter required a count of >1 per SNP and at least one count for both reference and alternate alleles. To identify SNPs with allele-specific effects, we applied QuASAR-MPRA (?), where for each SNP the reference and alternate allele counts were compared to the DNA proportion. QuASAR-MPRA results from each replicate were then combined using the fixed effects method, and corrected for multiple tests using BH procedure (Benjamini and Hochberg, 1995). The effect size $\beta_{l,n}$ of each replicate $n$ is weighted by $w_{n,l} = 1/\hat{\sigma}^2_{n,l}$, to calculate the overall effect size and standard error:

$$\beta^*_l = \frac{1}{w^*_l} \sum_n \beta_{n,l} w_{n,l} \quad \text{and} \quad \sigma^*_l = \sqrt{1/w^*_l}$$  \hspace{1cm} (0.13)$$

where $w^*_l = \sum_n w_{n,l}$. We can then calculate the Z-score and $p$-value to test for an overall change between all the RNA replicates combined with respect to the original DNA proportion $\beta_0$:

$$Z_l = \frac{\beta^*_l - \beta_0}{\sigma^*_l}, \quad \beta_0 = \log \frac{r_l}{1 - r_l}, \quad p = 2\Phi(-|Z_l|)$$  \hspace{1cm} (0.14)$$

We used the genomic inflation test to calculate the genomic inflation parameter, $\lambda$, for a set of $p$-values (Yang et al., 2011). For this we calculated the ratio of the median of the $p$-value distribution to the expected median, thus quantifying the extent of the bulk inflation and the excess false positive rate.
**BiT-BUNDLE-seq data analysis**

Counts from both the unbound and bound DNA were combined, and a filter was set so that each SNP direction combination had 5 counts for each allele. This combined count was also used to calculate a reference proportion. Each replicate for the bound and unbound libraries were then run through QuASAR-MPRA using the calculated reference proportion. These were then compared using $\Delta$AST (Moyerbrailean et al., 2016a) to identify ASB in the bound fraction that is differential relative to the unbound fraction. The replicates were combined using Stouffer’s method (STOUFFER et al., 1949) to identify ASB for each NFKB1 concentration, and combined again to identify the total ASB. The unbound and bound libraries counts were additionally analyzed with DESeq2 (Love et al., 2014) to identify over-represented bound enhancer regions (FDR 1% and logFC>1). To better estimate the dispersion parameters, the DESeq2 model was fit on all sequencing data and without merging the replicate libraries:

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i)$$  \hspace{1cm} (0.15)

$$\mu_{ij} = s_j q_{ij}$$  \hspace{1cm} (0.16)

$$\log_2(q_{ij}) = \beta_{i,0} + \beta_{i,C(j)} + \beta_{i,B(j)}$$  \hspace{1cm} (0.17)

For each enhancer region $i$ and sample $j$, the read counts $K_{ij}$ are modeled using a negative binomial distribution with fitted mean $\mu_{ij}$ and an enhancer region-specific dispersion parameter $\alpha_i$. The fitted mean is composed of a sample-specific size factor $s_j$ and a parameter $q_{ij}$ proportional to the expected true concentration of regions for sample $j$. The
coefficient $\beta_0$ represents the mean effect intercept, $\beta_{C(k)}$ represents the lane (NFKB1 concentration:replicate) effect, and and $\beta_{B(j)}$ represents the Bound/Unbound effect for each NFKB1 concentration (High, Medium, and Low).

We then contrasted the bound to the unbound for each concentration (i.e., high concentration bound to high concentration unbound) using the default DESeq2 Wald test for each enhancer region $\beta_{B(j)} \neq 0$, and a Benjamini-Hochberg (BH) adjusted $p$-value was calculated with automatic independent filtering (DESeq2 default setting).

**GWAS overlap**

SNPs nominally significant ($p<0.05$) for ASB (identified with $\Delta$AST) or ASE (identified with QuASAR-MPRA) that were also annotated as CentiSNP were overlapped with SNPs from the GWAS catalogue (V6) (MacArthur *et al.*, 2017), as well as with SNPs fine-mapped with the fgwas software as in (Moyerbrailean *et al.*, 2016b).

Table 14: **BiT-STARR-seq results.** QuASAR-MPRA results for BiT-STARR-seq.

Available at [https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S1.txt](https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S1.txt)

Table 15: **Annotations Used: CentiSNPs.** SNP annotations used for overlap with BiT-BUNDLE-seq and BiT-STARR-seq. First 4 columns are in the same order for each file (chr, pos, pos1, rsID). Column 5 contains the transcription factor with a CentiSNP at that location.

Available at [https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S2.txt](https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S2.txt)

Table 16: **Annotations Used: GWAS.** SNP annotations used for overlap with BiT-BUNDLE-seq and BiT-STARR-seq. First 4 columns are in the same order for each file (chr, pos, pos1, rsID). Column 5 contains the GWAS trait associated with the SNP.

Available at [https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S3.txt](https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S3.txt)

Table 17: **Annotations Used: eQTL.** SNP annotations used for overlap with BiT-BUNDLE-seq and BiT-STARR-seq. First 4 columns are in the same order for each file (chr, pos, pos1, rsID). eQTL SNPs. Column 5 contains the information for whether the eQTL was identified in cells infected with L (*Listeria*), S (*Salmonella*), or NI (not infected). Column 6 contains the gene associated with the eQTL. Column 7 contains the beta for the eQTL association. Column 8 contains the $p$-value for the eQTL association.

Available at [https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S4.txt](https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S4.txt)
Table 18: Distribution of ASE $Z$-scores. For each regulatory category: KS-test results from comparing the $Z$-score distribution for ASE for the category vs the negative control.

<table>
<thead>
<tr>
<th>Reg Cat</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CentiSNPs</td>
<td>$2.44 \times 10^{-6}$</td>
</tr>
<tr>
<td>ASH</td>
<td>$4.28 \times 10^{-4}$</td>
</tr>
<tr>
<td>Liver_eQTLs</td>
<td>$3.19 \times 10^{-4}$</td>
</tr>
<tr>
<td>LCL_eQTLs</td>
<td>0.01</td>
</tr>
<tr>
<td>fgwas_SNPs</td>
<td>$&lt; 2.20 \times 10^{-16}$</td>
</tr>
</tbody>
</table>

Table 19: Transcription factors in BiT-STARR-seq. Number of SNPs in motifs matching the top 10 covered transcription factors in BiT-STARR-seq.

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTCF</td>
<td>4911</td>
</tr>
<tr>
<td>E2F-1</td>
<td>2794</td>
</tr>
<tr>
<td>E2F</td>
<td>4407</td>
</tr>
<tr>
<td>ATF</td>
<td>5567</td>
</tr>
<tr>
<td>AML1</td>
<td>3794</td>
</tr>
<tr>
<td>ATF2:c-Jun</td>
<td>3651</td>
</tr>
<tr>
<td>CREB</td>
<td>12955</td>
</tr>
<tr>
<td>AP1</td>
<td>2673</td>
</tr>
<tr>
<td>ARG RI</td>
<td>3445</td>
</tr>
<tr>
<td>STF1</td>
<td>3561</td>
</tr>
</tbody>
</table>

Table 20: DEseq results: Combined concentrations. Differentially bound regions for combined concentrations. Columns are (identifier(rsID_Direction), adjusted $p$-value, $p$-value, logFC).

Available at https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S7.txt

Table 21: DEseq results: Low concentration. Differentially bound regions for low concentration. Columns are (identifier(rsID_Direction), adjusted $p$-value, $p$-value, logFC).

Available at https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S8.txt

Table 22: DEseq results: Mid. Differentially bound regions for mid concentration. Columns are (identifier(rsID_Direction), adjusted $p$-value, $p$-value, logFC).

Available at https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S9.txt

Table 23: DEseq results: High. Differentially bound regions for high concentration. Columns are (identifier(rsID_Direction), adjusted $p$-value, $p$-value, logFC).

Available at https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S10.txt

Table 24: BiT-BUNDLE-seq results. $\Delta$AST results for BiT-BUNDLE-seq. Columns are identifier, $Z$-score, $p$-value, adjusted $p$-value, rsID.

Available at https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S11.txt
Table 25: **Designed Regulatory Category Content.** For each regulatory category: the number of constructs and how many were significant (FDR 10%) for ASB.

<table>
<thead>
<tr>
<th>Reg Cat</th>
<th>Tested ASB</th>
<th>ASB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASH</td>
<td>180</td>
<td>4</td>
</tr>
<tr>
<td>CentiSNPs</td>
<td>50359</td>
<td>1514</td>
</tr>
<tr>
<td>fgwas_SNPs</td>
<td>5811</td>
<td>285</td>
</tr>
<tr>
<td>Negative_Control</td>
<td>1676</td>
<td>43</td>
</tr>
<tr>
<td>LCL_eQTLs</td>
<td>2753</td>
<td>73</td>
</tr>
<tr>
<td>Liver_eQTLs</td>
<td>29070</td>
<td>1009</td>
</tr>
</tbody>
</table>

Table 26: **Enrichment for ASB and ASE variants in TFs.** For each transcription factor: enrichment results from a Fisher’s test for ASE in the category vs being in the TF, subset for having ASB.

<table>
<thead>
<tr>
<th>TF</th>
<th>OR</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AML1</td>
<td>4.61</td>
<td>0.01</td>
</tr>
<tr>
<td>CREB1</td>
<td>Inf</td>
<td>0.02</td>
</tr>
<tr>
<td>CTCF</td>
<td>2.86</td>
<td>0.07</td>
</tr>
<tr>
<td>CREB</td>
<td>1.31</td>
<td>0.35</td>
</tr>
<tr>
<td>ARG RI</td>
<td>1.37</td>
<td>0.58</td>
</tr>
<tr>
<td>STF1</td>
<td>1.26</td>
<td>0.83</td>
</tr>
<tr>
<td>E2F</td>
<td>0.98</td>
<td>0.84</td>
</tr>
<tr>
<td>ATF</td>
<td>0.94</td>
<td>0.86</td>
</tr>
<tr>
<td>AP1</td>
<td>1.21</td>
<td>1.00</td>
</tr>
<tr>
<td>ATF:c-Jun</td>
<td>1.10</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 27: **ASB and complex traits.** ΔAST results for BiT-BUNDLE-seq. SNPs are nominally significant, associated to a complex trait, and are also CentiSNPs. Columns are rsID, direction, p-value, complex trait.

Available at [https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S14.txt](https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S14.txt)

Table 28: **ASE and complex traits.** QuASAR-MPRA results for BiT-STARR-seq. SNPs are nominally significant, associated to a complex trait, and are also CentiSNPs. Columns are rsID, direction, p-value, complex trait.

Available at [https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S15.txt](https://genome.cshlp.org/content/suppl/2018/10/17/gr.237354.118.DC1/Supplemental_Table_S15.txt)

Table 29: **Designed Regulatory Category Content.** For each regulatory category: the number of constructs and how many were significant (FDR 10%) for ASE and enrichment results from a Fisher’s test for ASE in the category vs the negative control.

<table>
<thead>
<tr>
<th>Reg Cat</th>
<th>Tested ASE</th>
<th>Sig ASE</th>
<th>OR ASE</th>
<th>p-value ASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASH</td>
<td>162</td>
<td>5</td>
<td>1.77</td>
<td>0.23</td>
</tr>
<tr>
<td>CentiSNPs</td>
<td>43615</td>
<td>1806</td>
<td>2.40</td>
<td>1.28 × 10^{-5}</td>
</tr>
<tr>
<td>fgwas_SNPs</td>
<td>4894</td>
<td>338</td>
<td>4.12</td>
<td>1.54 × 10^{-13}</td>
</tr>
<tr>
<td>Negative_Control</td>
<td>1111</td>
<td>20</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>LCL_eQTLs</td>
<td>2307</td>
<td>94</td>
<td>2.36</td>
<td>2.41 × 10^{-4}</td>
</tr>
<tr>
<td>Liver_eQTLs</td>
<td>22943</td>
<td>827</td>
<td>2.08</td>
<td>4.88 × 10^{-4}</td>
</tr>
</tbody>
</table>

Table 30: **Primers used in BiT-STARR-seq.**

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>STARR_F_SH</td>
<td>CCAGGCCCACGAGACCTAGAGTCGGGCGGCCGCCG</td>
</tr>
<tr>
<td>STARR_R_SH</td>
<td>TGACGCTGCGACGAAATTATTACACGGCGATC</td>
</tr>
<tr>
<td>F_transposase</td>
<td>TCGTGGCAGCGTCAGATGTGTATAAGAGACAG</td>
</tr>
<tr>
<td>R_transposase</td>
<td>GTCTCGTGGGCTGGAGATGTGTATAAGAGACAG</td>
</tr>
<tr>
<td>F_trans_short</td>
<td>TCGTGGCAGCGTCAGAT</td>
</tr>
<tr>
<td>I2.1</td>
<td>CAAGCAGAAGACGCGCATACGAG</td>
</tr>
<tr>
<td>Nextera_i7_10N</td>
<td>CAAGCAGAAGACGCGCATACGAGRDHBVDHBVDGTCTCGTGGGCTCGG</td>
</tr>
</tbody>
</table>
Table 31: **Sequencing Depth.** Sequencing depth for each BiT-STARR-seq replicate. Seq depth is total reads, and after deduplication is the number of reads after removing duplicates using UMI tools.

<table>
<thead>
<tr>
<th>Rep</th>
<th>Seq Depth</th>
<th>UMI Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rep1</td>
<td>89,360,505</td>
<td>482,117</td>
</tr>
<tr>
<td>Rep2</td>
<td>55,819,932</td>
<td>182,865</td>
</tr>
<tr>
<td>Rep3</td>
<td>32,784,823</td>
<td>1,487,089</td>
</tr>
<tr>
<td>Rep4</td>
<td>34,141,541</td>
<td>577,343</td>
</tr>
<tr>
<td>Rep5</td>
<td>71,090,681</td>
<td>464,647</td>
</tr>
<tr>
<td>Rep6</td>
<td>36,835,814</td>
<td>987,771</td>
</tr>
<tr>
<td>Rep7</td>
<td>74,991,057</td>
<td>550,080</td>
</tr>
<tr>
<td>Rep8</td>
<td>61,014,562</td>
<td>640,360</td>
</tr>
<tr>
<td>Rep9</td>
<td>31,142,602</td>
<td>404,739</td>
</tr>
</tbody>
</table>
Figure 27: Schematic of oligos in BiT-STARR-seq and BiT-BUNDLE-seq.

<table>
<thead>
<tr>
<th>Name</th>
<th>Details</th>
<th>Length</th>
<th>Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_S1</td>
<td>Seq primer handle</td>
<td>15</td>
<td>GTGTTATAAGAGAGAAG</td>
</tr>
<tr>
<td>H_S2</td>
<td>Seq primer handle</td>
<td>15</td>
<td>CTGCTCTTTATACAC</td>
</tr>
<tr>
<td>S1</td>
<td>Seq primer</td>
<td>33</td>
<td>TGCTGGGCGGTCGATCGATCGATATAGAGAAG</td>
</tr>
<tr>
<td>S2</td>
<td>Seq primer</td>
<td>34</td>
<td>CTGCTCTTTATACACATCGAGCGACAGAGAGAC</td>
</tr>
<tr>
<td>S5</td>
<td>Index</td>
<td>8</td>
<td>NNNNNNNNNN</td>
</tr>
<tr>
<td>UMI</td>
<td>Index</td>
<td>10</td>
<td>RDELHDHBDU</td>
</tr>
<tr>
<td>P5</td>
<td>Adapter</td>
<td>27</td>
<td>AATGATACGCTGCTCCTTG</td>
</tr>
<tr>
<td>P7</td>
<td>Adapter</td>
<td>24</td>
<td>ATCTGATATACGCTCCTTG</td>
</tr>
</tbody>
</table>

Figure 28: Correlation of DNA libraries. Scatterplot of filtered DNA library counts for each replicate plotted against all other replicates in log_{10} scale. Spearman rho correlation range is stated at the top.
Figure 29: **Correlation of RNA libraries.** Scatterplot of filtered RNA library counts for each replicate plotted against all other replicates in $\log_{10}$ scale. Spearman $\rho$ correlation range is stated at the top.
Figure 30: **BiT-STARR-seq effect by regulatory category.** $Z$-score distribution for SNPs in each designed regulatory category. Absolute $Z$-score (y axis) for each regulatory category (x axis) is shown in the boxplot, center line of the boxplot is the median.

Figure 31: **DNase window centering of BiT-STARR-seq variants.** QQplot depicting the $p$-value distributions from QuASAR-MPRA based on how far the regulatory variant is from the center of the DNase window.
Figure 32: **DNase peak size of BiT-STARR-seq variants.** QQplot depicting the p-value distributions from QuASAR-MPRA based on the DNase peak size.

Figure 33: **Enrichment of NF-kB complex footprints in BiT-BUNDLE-seq bound regions.** Fisher’s exact test was performed to identify enrichment (x axis is the OR) for significant differentially bound regions (logFC > 1 and FDR < 1%). In red are the regions containing a SNP in a NF-kB complex footprint, in blue the regions containing a SNP in footprints for other transcription factors.
Figure 34: *Overlap between constructs with significant ASB for each concentration in BiT-BUNDLE-seq.* In purple are the number ASB at low concentration, in yellow are the number ASB at mid concentration, and in blue are the number ASB at high concentration.

Figure 35: *Overlap between constructs with significant ASB or ASE.* A) Overlap at 10% FDR. B) Overlap at 20% FDR. C) Overlap at 30% FDR. D) Overlap at nominal $p$-value ($p$-value < 0.05).
Figure 36: Depletion of ASE with TFs that repress NFKB1 binding. QQplot depicting the ASE p-value distribution from QuASAR-MPRA for SNPs with CREB1 or AML1 binding (green) or not with CREB1 or AML1 binding (grey)
REFERENCES


Degner, J. F., Pai, A. a., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon,


Mangravite, L. M., Engelhardt, B. E., Medina, M. W., Smith, J. D., Brown, C. D., Chasman,


Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield,


Vietri, Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D., Tanay, A., and Hadjur,


ABSTRACT

VALIDATING FUNCTIONAL MECHANISMS FOR NON-CODING GENETIC VARIANTS ASSOCIATED WITH COMPLEX TRAITS

by

CYNTHIA A. KALITA

May 2019

Advisors: Dr. Francesca Luca and Dr. Roger Pique-Regi

Major: Molecular Genetics and Genomics

Degree: Doctor of Philosophy

Genome-wide association studies (GWAS) have identified a large number of genetic variants associated with disease as well as normal phenotypic variation for complex traits. However challenges remain in determining the functional relevance of human DNA sequence variants. Even after fine mapping, most variants are located in non-coding regions making it difficult to infer mechanisms linking individual genetic variants with the disease trait. In addition, we do not know under which environmental conditions the sequence variants have a functional impact, and whether they become one of many factors involved in complex phenotypes at the organismal level.

Chapter 1 describes computational methods to predict causal GWAS variants, validation of a subset for ASE using a traditional reporter assay, and development of a method to identify ASE in high throughput assays. These methods improved positive detection of enhancer activity and ASE, and this analysis pipeline will continue to be useful as more researchers begin using high throughput assays to identify allelic effects. Chapter 2 improves upon chapter 1 with the development of a new modification of STARR-seq in order
to streamline the assay and improve power to detect ASE through the addition of an UMI. Additionally, by integrating BiT-STARR-seq with a high throughput allele-specific EMSA, we are able to identify the mechanism behind many ASE variants.

Studying GxE in human studies is extremely difficult, so our approach of using an *in-vitro* method and modeling molecular phenotypes is a useful alternative. Chapter 3 describes the investigation of GxE with complex traits. Using GEMMA, we were able to identify environments that were enriched for complex traits. With ATAC-seq data we were able to identify differentially accessible regions, TF footprints, and differential TF footprints. Integrating this data with BiT-STARR-seq, we were able to identify enrichment for these differential chromatin accessibility regions with ASE.

Overall, these chapters show the integration of computational predictions with experimental validation in order to identify allelic effects. This design is a useful approach to validate the molecular mechanism for specific transcription factors, and link these to the context of human health.
It's always been my life pursuit to be a research scientist and over the past eleven years, I've taken strides to achieve that goal. While earning a Bachelor's of Science in Biotechnology (Magna Cum Laude in three years) from the State University of New York, Environmental Science and Forestry, I completed two honors research projects under the guidance of Dr. William Powell, working towards the restoration of the American Chestnut tree. For my part in the project, I completed transformation of American Chestnut embryos using an Agrobacterium-mediated transformation, a natural genetic engineer to insert genes for blight resistance, as well as a GFP marker.

After graduation, I took several non-academic positions, including quality control at Eastman Kodak and Arch Chemicals. Quickly, I realized I would not be satisfied doing routine testing, so I accepted an opportunity to join a research team at Dow Chemical. There I was part of a team researching how to synthesize carbon fiber from inexpensive polymer. This project was particularly challenging as we couldn't simply use whatever protocol worked to create the fiber, we also had to take into account the cost of the process. Over my two years on the team, we were able to generate carbon fiber that matched the strength and quality of carbon fiber currently on the market, but due to market fluctuations, the project was dropped as not economically viable. Interacting with doctoral level scientists there led me to apply to Wayne State University's Molecular Biology and Genetics (MBG) PhD program. I joined the laboratory of Dr. Francesca Luca, working on non-coding regulatory variation and was trained in computational analysis by co-PI Dr. Roger Pique-Regi. My work here resulted in several publications, including 1 co-first author, and 2 first author papers.