Characterizing human regulatory genetic variation using CRISPR/Cas9 genome editing

Margot Brandt

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

ABSTRACT

Characterizing human regulatory genetic variation using CRISPR/Cas9 genome editing

Margot Brandt

Rare gene-disrupting variants and common regulatory variants play key roles in rare and common disease, respectively. These variants are of great interest for investigation into genetic contributions to disease, but experimental methods to validate their impact on gene expression levels are lacking. In this study, we utilized CRISPR/Cas9 genome editing to validate regulatory variants including *cis*-eQTLs, rare stop-gained variants in healthy and disease cases and one immune-response *trans*-eQTL master regulator.

For investigation into common and rare regulatory variants within transcribed regions, we developed a scalable CRISPR-based polyclonal assay for experimental assessment. First, we applied this assay to nine rare stop-gained variants found in the general population, in GTEx. After editing, the stop-gained variants show a significant allele-specific depletion in transcript abundance, as expected. Next, we utilized the assay to validate 33 common eQTLs found in GTEx. After editing, the eQTL variants show higher variance in effect size than control variants, indicating a regulatory effect. Finally, we applied the polyclonal editing approach to clinical and new stop-gained variants in two disease-associated genes. The results follow the expected trend, with NMD being triggered by variants upstream of the NMD threshold but not by those beyond. This method demonstrates scalable experimental confirmation of putative causal regulatory variants, and improved interpretation of regulatory variation in humans.

Next, we sought to experimentally validate an immune-response eQTL for *IRF1* in *cis* and many genes in *trans* under LPS stimulation. We used CRISPRi to repress the enhancer locus and

found that the enhancer is active in our immune cell system. Next, we used CRISPR-Cas9 genome editing and isolation of monoclonal cell lines to target this variant locus. After LPS stimulation, we performed RNA-sequencing on wild type and edited clones, showing that the effect size of the genes which are associated with the *trans*-eQTL are correlated with differential expression between the edited and wild type cell lines for the same genes. Additionally, we find that the differential expression between edited clones is correlated with CRISPRi repression of the IRF1 promoter and enhancer. In this way, we were able to identify a common genetic variant which modifies the transcriptomic immune response to LPS and validate the *trans*-eQTL signal.

TABLE OF CONTENTS

List of figures and tables

Chapter 1

Chapter 2

Chapter 3

Acknowledgements

I would first like to thank my advisor Tuuli whose whole-hearted passion for her research is felt by all those lucky enough to work with her. Tuuli is forever optimistic and very understanding that science (especially in the wet lab) does not always go as planned. I would next like to thank my thesis committee members for five years of wisdom and guidance. I'd also like to thank Ana V., the magician who miraculously created the first faculty wet lab at NYGC out of thin air, and our past technicians and rotation students Ana P., Aaron, Yocelyn and Nora for invaluable contributions to wet lab projects. I'd also like to thank Sarah for allowing me to adopt the IRF1 project and providing insight along the way. Thanks to our current technician Alper for being my right-hand man in the lab. Many an hour were spent side by side in the fume hood pipetting trizol. The RNA may degrade, but the memories won't.

I would also like to thank all current and past members of the Lappalainen lab for their advice and support throughout my PhD and more importantly for humoring my obsession with discussing different Trader Joe's products. Now that we're here I'd also like to thank Trader Joe's and all its cheerful checkout people for keeping me fed and happy throughout the thesis writing process. Bet you didn't think I'd incorporate Trader Joe's into my thesis! Not unlike scoring cauliflower gnocchi from the frozen aisle on a Sunday evening, these six years of grad school have been no easy feat. I would not have been able to do it without support (and distraction) from my friends. I am grateful for trying to set the record for most countries visited during grad school with Stephanie, wine nights (a.k.a. therapy sessions) with Chelsea, Clau and Ruth, and reliving college with Wake friends who somehow all keep moving to NYC. Also, thanks to Mike for making me dinner when I had too much work to do and tolerating my "feedback" on his cooking technique.

Finally, I’d like to acknowledge my parents and brother for supporting all my endeavors, attending every graduation ceremony, taking me to Broadway shows and filing my taxes.

Dedication

This work is dedicated to every first-year grad student who is terrified that they won’t be able to do it and aren’t good enough. You will and you are.

Chapter 1: Introduction

I. Genetic variation and human traits

Genetic variation in the genome is ubiquitous and diverse. Investigating genetic variation can provide insights into disease etiology, phenotypic variation, gene function, evolution, and more. Variants can be broadly classified into four categories: structural variants, tandem repeat variants, single nucleotide variants (SNV), and small insertion/deletion variants (indel). Structural variants cause genomic rearrangements or copy number variants of greater than 50 bp. They can have severe consequences and are quite rare, occurring several thousand times per human genome. Tandem repeat variants are variants which affect the number of repeats in a repetitive region of the genome. Due to the difficulty in sequencing and mapping these repetitive regions, the exact frequency is unknown. SNVs and indels are by far the most common form of variation with 3-4 million SNVs and 0.4-0.5 million indels occurring per human genome.

SNVs and indels can be further classified as coding or noncoding. The protein-coding portion of the genome is composed of the segments of genomic DNA which translate to the protein sequence of the roughly 20,000 proteins encoded in the genome. Coding variants are defined as those falling into this protein-coding portion of the genome, while noncoding variants fall into the 99% of the genome which does not code for protein. The vast majority of SNVs and indels have no functional effect, but many do. Noncoding variants have typically been thought of as having an effect on the dosage of protein, while coding variants are thought of as affecting protein structure by changing the protein coding sequence. As we study more of these variants, we find that both coding and noncoding variants can impact dosage or structure through a variety of mechanisms (figure 1-1).

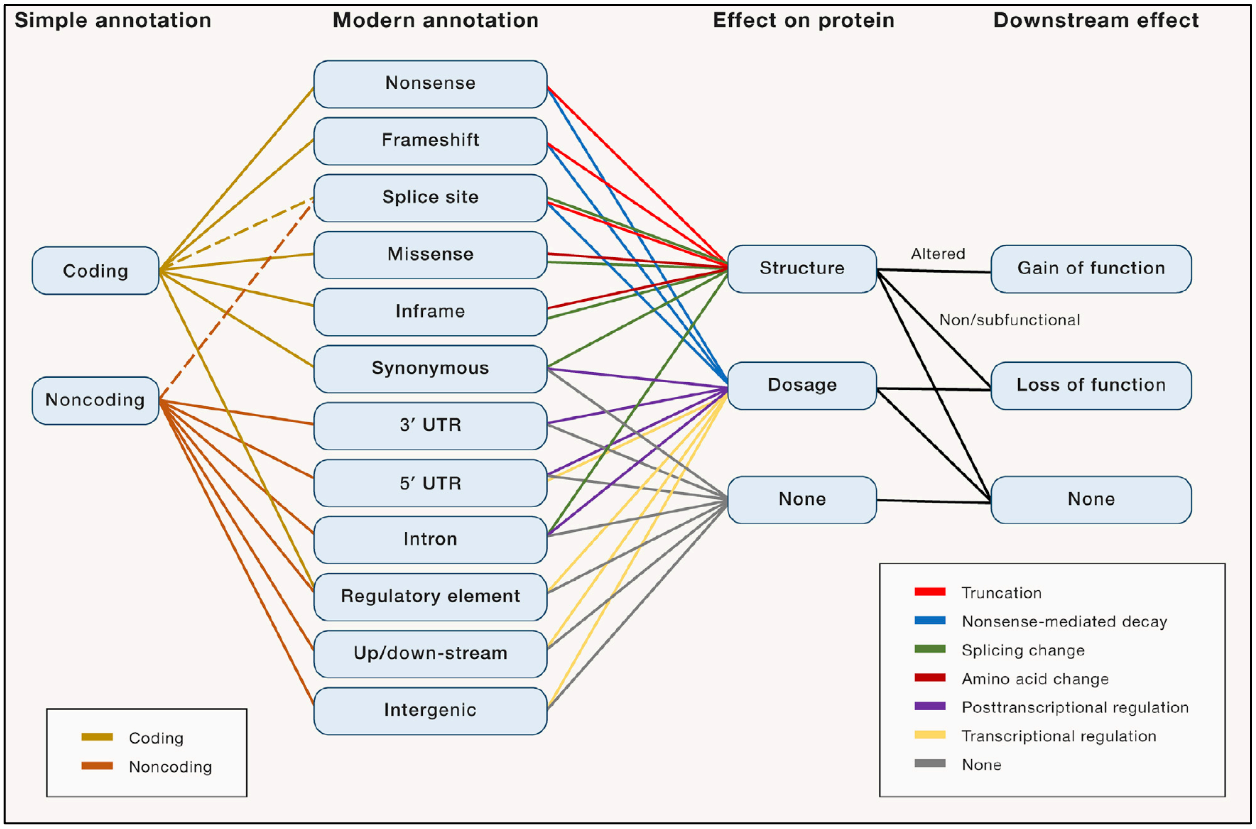


Figure 1-1. Different types of genetic variants and their potential effects on protein production in the cell.

Non-coding variants can impact dosage or structure of protein or have no effect, whereas coding variants can affect dosage, structure or have no effect at all. Adapted from (Lappalainen et al., 2019).

Understanding phenotypic consequences of variation of the genome has been a priority in the field of genetics for as long as the field has existed. Variants associated with mendelian disorders tend to be rare and have high penetrance. Historically, linkage analysis was utilized to identify the location of mendelian variants by tracking disease cases in families. In the age of next generation sequencing, these variants can be discovered by whole exome sequencing of unrelated patients to look for damaging mutations within the coding region of genes. However, a challenge to this approach is distinguishing harmless variants from gene-disrupting.

Common variants which typically have low penetrance and have small contributions to disease require a much larger sample size to detect their effects. The current standard approach to understand the effects of common variants is associating variants in the human population with human traits through genome-wide association studies (GWAS). GWAS have been performed on a multitude of complex human traits (Welter et al., 2014). Not surprisingly, the traits being studied by GWAS with the most interest and attention are human diseases. The rationale behind GWAS design is that alleles which are enriched in disease subjects over controls are likely to be contributing to disease. A variant which is found to be associated with disease would ideally have a clear mechanism of effect on a gene, e.g. by disrupting the protein-coding sequence of a gene by introduction of a missense or premature stop codon or by affecting splicing. Unfortunately, about 90% of GWAS variants are not found within protein-coding regions but instead within the noncoding portion of the genome (Maurano et al., 2012). This phenomenon has created a need for a better understanding of how noncoding variants affect genome function, regulation of proximal target genes, and subsequent pathways that affect disease risk.

A *cis*-regulatory variant in the genome is defined by one allele causing a higher expression of a proximal gene than the other allele (figure 1-2a). Noncoding regulatory variants can affect expression of genes through a multitude of mechanisms. For example, they can act by disrupting the binding site of a transcription factor (TF) in an enhancer or affecting the chromatin state of the region surrounding the variant. One obstacle is determining which gene a variant is affecting. Assuming that the variant affects the nearest gene is a flawed approach since it has been reported that approximately one-third of enhancers affect genes which are not the nearest (Gasperini et al., 2019) and up to two-thirds of GWAS variants are associated with a gene which is not the nearest (Zhu et al., 2016). An understanding of the effect of an associated variant on gene expression gives the first clue into the mechanism of disease. Further experimental follow up can deduce how this expression regulation then affects the cell or organism but the first step is to understand how the sequence modification affects gene expression. Thus, systematic interrogation of the effects of common variants on gene expression is essential. Expression quantitative trait loci (eQTL) studies have emerged to fill this void, associating variants with expression transcriptome-wide in order to identify regulatory variants.

II. Discovering common regulatory variants

Expression quantitative trait loci (eQTL) studies

eQTL studies aim to identify common regulatory variants which have an effect on the transcriptome. A *cis*-eQTL study is designed to detect regulatory variants that affect the expression of genes nearby the variant. The studies are typically designed so that genotype data, either from genotyping arrays or whole genome sequencing, and gene expression data, either from microarray or RNA-sequencing, are collected for a population of individuals. To test for association of a variant with expression of a proximal gene, a linear regression is performed between the genotype of the variant and expression of the gene (figure 1-2b). This test is repeated for all variants within a specific window on either side of the transcription start site (TSS) of the gene, resulting in a p-value of association for each variant (figure 1-2c). This is then repeated for every gene in the genome. A gene which is associated with a significant eQTL is referred to as an eGene, while all variants that are significantly associated with the eGene are referred to as eVariants. With large enough sample sizes, eQTL studies have discovered an eQTL association for almost every gene in the genome, and multiple independent eQTLs for many genes (GTEx Consortium et al., 2017). The results help illuminate elements in the genome that regulate gene expression, as well as assist in understanding the mechanisms of human disease. eQTLs have been reported to be enriched for

GWAS loci (Lappalainen et al., 2013) and GWAS loci that are also eQTLs are more likely to be true associations (Nicolae et al., 2010). Overlap between eQTLs and GWAS loci is therefore utilized to interpret the mechanism of noncoding GWAS variants (Zhu et al., 2016).

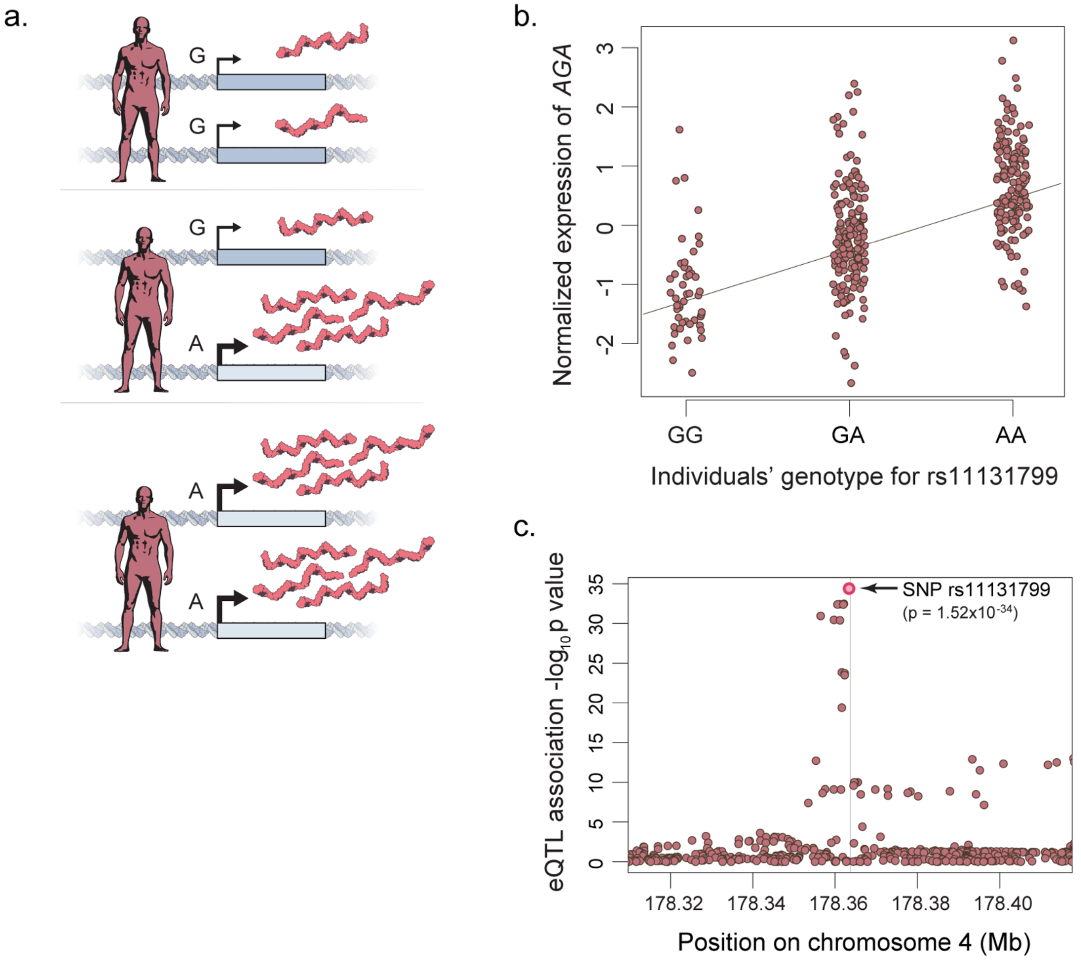


Figure 1-2. An example eQTL.

a) The A allele results in higher expression of the *AGA* gene than the G allele. Individuals homozygous for the G allele have the lowest overall expression, while those homozygous for the A allele have the highest expression. Heterozygous individuals have intermediate expression. b) The effect of the variant is detected by performing a linear regression between genotype and expression across all individuals in the study. This test is repeated for all variants within 1 Mb of the transcription start site of *AGA*. c) The top variant has the strongest association with gene expression, but many variants may have significant associations due to linkage disequilibrium with the causal variant. Adapted from (Brandt and Lappalainen, 2017).

Cis-eQTLs regulate expression in an allelic fashion, meaning that the variant affects expression only of the haplotype of the gene on which it falls. Consider an example where a TF binding site is disrupted by the alternative allele of the variant, resulting in decreased expression of the gene. The TF's binding on the reference allele is not affected and therefore the expression of that haplotype remains unchanged. *Trans*-eQTL studies, on the other hand, aim to detect regulatory variants that act on distal genes. As in *cis*-eQTL studies, variants are tested for association using linear regression between genotype and expression. A *trans*-eQTL is defined as an association between a variant and a gene which is enacted through a *trans* mediator. *Trans*-eQTLs in practice are detected as a variant which resides on a different chromosome or greater than 5 Mb from the TSS of the eGene.

One possible mechanism through which a *trans*-eQTL can act is by affecting the expression of another gene in *cis*, which then affects the expression of the tested gene (example in figure 1-3). Thus, the causal variant for a *cis*-eQTL can also be the causal variant for a *trans*-eQTL. This is not the only mechanism through which a *trans*-eQTL can act. In addition to the *cis*-eGene directly regulating the *trans*-eGene, it could regulate another gene which in turn regulates the *trans*-eGene. Additionally, a *trans*-eQTL might not involve a *cis*-eQTL at all. A variant in the protein-coding region of a TF gene affecting its activity could in turn affect expression of any downstream targets of that gene. *Trans*-eQTLs by nature do not exert an allelic effect. As in the example in figure 1-3, the *trans*-eQTL effect exists because the amount of TF A is reduced, and that affects both alleles of Gene B equally.

Unlike *cis*-eQTL studies where variants are limited to a window around the TSS of a given gene, *trans*-eQTL studies test all variants genome-wide for association with each gene. Thus, the multiple testing burden on *trans*-eQTLs is much higher than for *cis*-eQTLs, which has made them much more difficult to detect with the limited size of a typical eQTL study. Even studies with sample sizes of greater than 5,000 individuals have discovered modest numbers of *trans*-eQTLs, on the order of a couple of hundred (Brynedal et al., 2017; Westra et al., 2013; Yao et al., 2017). By increasing the sample size to over 31,000 and testing only trait-associated variants, a recent study was able to detect a few thousand *trans*-eQTLs (Võsa et al., 2018). Additionally, *trans*-eQTL false positives are more likely than in *cis*-eQTLs and can be caused by batch effects, population structure or even RNA-sequencing alignment errors (Saha and Battle, 2018).

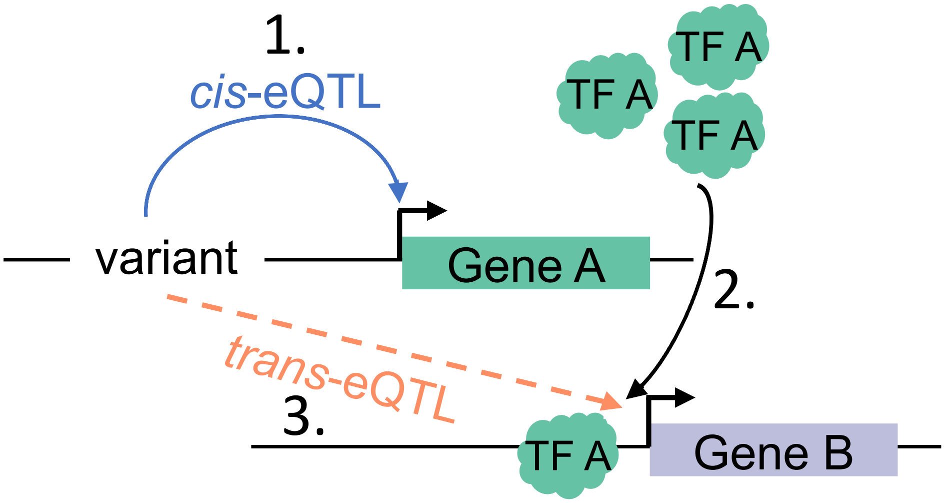


Figure 1-3. An example mechanism of a *trans*-eQTL association.

(1) The variant is a *cis*-eQTL variant for Gene A, which codes for a transcription factor, TF A. (2) When the protein TF A is translated from Gene A, it in turn affects expression of Gene B. (3) Thus, the variant exhibits a 293Tassociation for Gene B.

Context-specificity of eQTLs

Most of the eQTL studies mentioned thus far have performed RNA-seq on either one type of cell line isolated from individuals or an easily accessible tissue, such as whole blood. Recently, emphasis has been placed on context-specific eQTLs: either tissue-specific or condition-specific. For example, the GTEx consortium has performed eQTL analysis on 44 tissues from hundreds of post-mortem donors (GTEx Consortium et al., 2017). The resulting eQTLs demonstrate a u-shaped distribution where eQTLs tend to be either shared by all tissues or unique to one tissue. Interestingly, it appears that *trans*-eQTLs tend to be more tissue-specific than *cis*-eQTLs.

In addition to tissue-specific eQTLs, studies have focused on eQTLs which are active under specific conditions. Innate immune cells, such as monocytes, have been used for looking at context-specificity of eQTLs under stimulation (Chen et al., 2016; Fairfax et al., 2014; Kim-Hellmuth et al., 2017). Monocytes lend themselves to this type of study because they can be isolated from blood and have a transcriptional response to exposure to viral or bacterial stimuli. Immune-response eQTLs (reQTL) are eQTLs where the association between the variant and eGene varies based on stimulation. Figure 1-4 shows an example of a reQTL where a variant is not associated with expression of a gene under normal baseline conditions, but with exposure to a stimulus, individuals with the A allele have higher gene expression of the eGene than those with the G allele. Immune reQTLs have been found to overlap GWAS signals for diseases related to infection and immunity (Chen et al., 2016; Fairfax et al., 2014; Kim-Hellmuth et al., 2017). This demonstrates the utility of immune reQTLs to help determine the mechanism of GWAS loci in the noncoding genome, perhaps revealing mechanisms that would be missed if only utilizing cells without stimulation.

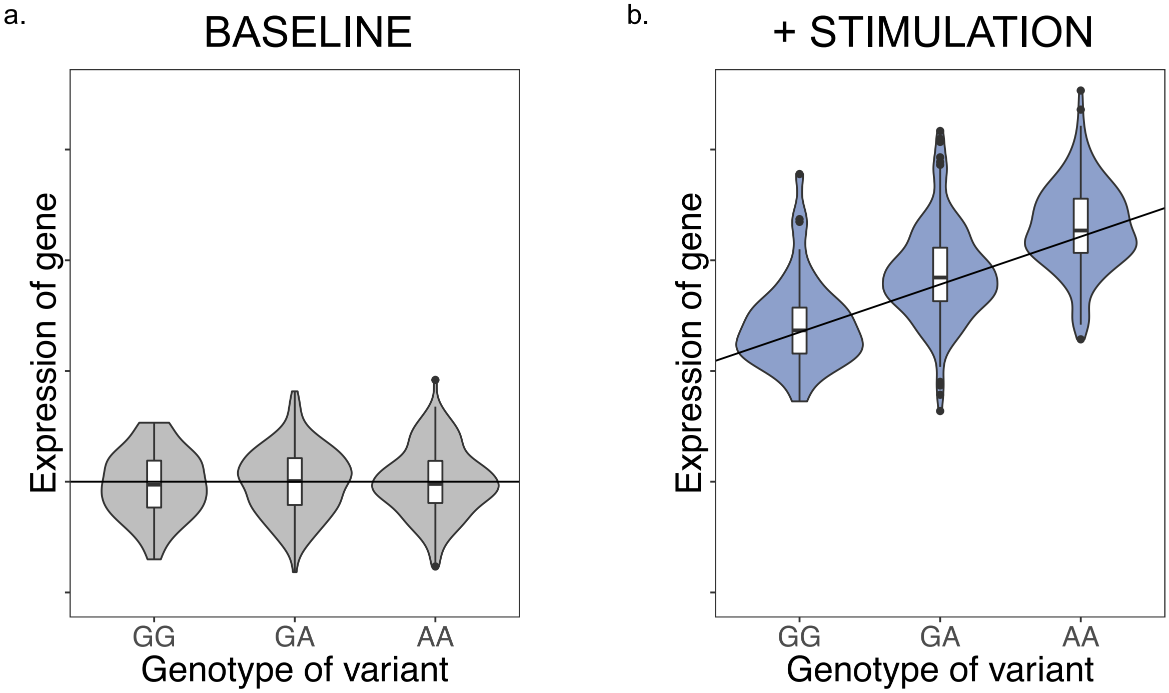


Figure 1-4. A simulated example immune-response eQTL in human immune cells. (a) Under normal conditions, the variant does not show an association with expression of the gene, but with immune stimulation (b), there is an effect of the genotype of the variant on expression of the gene. Thus, the variant is considered an immune-reQTL.

Linkage disequilibrium in eQTL studies

Linkage disequilibrium (LD) is the phenomenon that variants near each other on the same chromosome tend to be inherited together. Recombination during meiosis, which occurs more often at recombination hotspots, can separate variants on the same chromosome. But if there is little recombination between two variants, they are often found in the same individuals and are therefore difficult to statistically disentangle in association studies. LD is advantageous in association studies where variants which are not directly measured are imputed, but it is a problem for identifying the causal variant at an eQTL locus. Any variant with a significant association with an eGene is considered an eVariant, but there is often only one causal eVariant at an eQTL. The

LD phenomenon is exemplified in figure 1-2c where several variants have highly significant p-values for association with the *AGA* gene, while only one variant is likely to be causal.

III. Approaches to determine causal variants

Statistical fine-mapping of eQTLs

In order to distinguish causal variants from those in LD with causal variants, several computational and experimental methods have been developed. The computational methods generally involve statistical fine-mapping of the eQTL locus to assign a probability of causality to each variant (Hormozdiari et al., 2014; Pickrell, 2014; Wellcome Trust Case Control Consortium et al., 2012; Wen et al., 2015). The methods incorporate LD information between variants with the genetic association statistics to determine probability of causality, while taking into account noise that can be introduced by small sample sizes. Some of these methods assume a single causal variant at the eQTL locus (Pickrell, 2014; Wellcome Trust Case Control Consortium et al., 2012), which can lead to false negatives or positives. CAVIAR (Hormozdiari et al., 2014), the fine-mapping approach utlilized in chapter 2, provides posterior probabilities of association for each eVariant, without assumption of the number of eQTLs present at an eGene. Even with fine-mapping, it is often still unclear which is the causal variant. In instances where two variants are in perfect or close to perfect LD, it is impossible computationally to distinguish which is causal. Thus, the door is open to experimental approaches to detect causality of eVariants.

Experimental approaches for validating regulatory variants

The high-throughput experimental approaches to validating regulatory signals fall into two broad categories: (1) those which focus on the difference in expression between multiple alleles of a regulatory variant, incorporating the sequence into a reporter assay, and (2) those which confirm activity of a regulatory element within which a variant may fall, maintaining the native genomic context.

The first category of high-throughput approaches is dominated by massively parallel reporter assays (MPRAs), reviewed in (Inoue and Ahituv, 2015; Santiago-Algarra et al., 2017). MPRAs consist of incorporating a potential regulatory DNA sequence into a reporter plasmid, upstream of a weak promoter and a reporter gene such as GFP or luciferase, often followed by a unique barcode which can be linked to the regulatory DNA sequence. This is done in parallel with thousands of DNA sequences either composed of endogenous gDNA or with mutations introduced. The constructs are pooled and introduced into cells in culture. To quantify the effect of each DNA sequence, the barcodes are sequenced from the mRNA of the cells. An over- or under-representation of a particular barcode suggests that the matching DNA sequence in indeed regulatory. A comparison of the effect of two alleles of the same DNA sequence can identify an allele-specific effect of a variant on expression of a gene, and therefore point to a causal regulatory variant. The major advantage to these assays is certainly the high-throughput capacity; It has been extended to genome-wide scale (van Arensbergen et al., 2019). The ability to compare alleles of the same variant in isolation is also an advantage but taking the variant out of its endogenous genomic context into an artificial construct can result in false-positive or false-negative results. In addition, the results from these assays often do not correlate well with population trends (Tewhey et al., 2016; van Arensbergen et al., 2019).

The second category of high-throughput approaches involves confirming regulatory elements while maintaining the native genomic context. Most recently, the emergence of CRISPR interference (CRISPRi) technology has led groups to utilize the technology to perturb regions of the genome to detect regulatory effects on expression of genes (Gasperini et al., 2019; Gilbert et al., 2013; Yeo et al., 2018). The advantage to this approach is that it maintains the endogenous genomic context, which can be important for enhancer activity. However, these assays do not provide allele-specific effects and therefore cannot directly confirm the effect of a variant on expression of a gene.

All above high-throughput assays, both the reporter-based and endogenous approaches, are relatively good at identifying regulatory elements in enhancers. A major drawback is that they are unable to assay variants which are found within the transcript. These variants can affect the expression of genes either through stability of the transcript, splicing disruption or introduction of frameshift or premature stop codons.

Introduction of a specific variant into human cells using genome editing and homologous recombination is another approach with which to validate single regulatory variants. CRISPR/Cas9 genome editing, a tool harnessed from an adaptive immune system in bacteria, provides an easily-programmable method to target a specific locus in the genome for editing (Cong et al., 2013; Jinek et al., 2012; Shalem et al., 2014). In practice, genome editing can be achieved by introducing Cas9 and a locus-targeting guide RNA (gRNA) into a cell line. This induces the Cas9 enzyme to make a double stranded break (DSB) in the genomic DNA at the locus which the gRNA targets. In order to achieve a specific change in the DNA sequence, for example converting a specific variant to the alternative allele, a homologous template containing the variant is also provided. The cell has two pathways for repair of a DSB: non-homologous end joining (NHEJ) and homology directed repair (HDR). The editing of the cells will yield a combination of results from both repair pathways: NHEJ will manifest as insertions or deletions, and HDR will manifest as a seamless repair of the genome using the provided template. In order to assess particular effects of the variant, generally one would isolate single cells, expand each into a monoclonal cell line and select those cell lines which have the desired mutation. All in all, this is a fairly time consuming and low-throughput process, and therefore is only practical in cases where a single variant or handful of variants are of interest (Gupta et al., 2017; Soldner et al., 2016; Zhu et al., 2018).

A less labor-intensive genome editing approach involves editing cells as described above but analyzing the polyclonal edited cell population instead of isolating monoclonal cell lines. Groups have utilized this approach with CRISPR/Cas9 genome editing and homologous recombination followed by sequencing for allelic expression to validate the effects of rare variants (Li et al., 2017) and all possible mutations in a particular exon by using saturation mutagenesis (Findlay et al., 2014). The advantages to this approach of variant confirmation are that the endogenous genomic locus of the variant is maintained and the allelic effect of the variant is measured. This approach is limited to validation of variants within the transcript, thus providing a complimentary method to validate variants which are unable to be validated with the MPRA method. This approach takes advantage of the fact that regulatory variants within the transcript are transcribed in the mRNA and therefore can be quantified by sequencing the mRNA. In parallel, the gDNA is sequenced in order to quantify the editing efficiency. A significant difference between the proportion of the variant in gDNA versus mRNA suggests a regulatory effect of the variant. An application of this approach to confirm regulatory transcript variants genome-wide would be an integral contribution to the field, and one of the aims of this thesis.

IV. Premature stop-gained variants

Nonsense-mediated decay

Like common transcript regulatory variants, rare stop-gained variants often exert an allelic effect on the expression level of the gene in which they reside. This expression effect is mediated through the nonsense mediated decay (NMD) pathway. NMD is a quality control mechanism utilized by the cell to detect transcripts with premature stop codons and degrade them before they are translated to potentially damaging truncated proteins. The NMD pathway has been extensively studied (reviewed in (Hug et al., 2016)). Studies find that premature stop codons prior to 50-55 bp before the last exon-exon junction tend to be degraded by the NMD pathway, whereas stop codons beyond this cutoff tend to escape NMD (Nagy and Maquat, 1998). As transcription of a gene occurs in the nucleus, an exon-junction complex is applied to each pre-mRNA 20-24 bp upstream of each exon-exon junction. This complex remains intact through spicing and transport to the cytoplasm, where NMD machinery recognizes a premature stop-codon and determines whether it is more than 50-55 bp upstream of the last exon-exon junction. The biological rationale for this cutoff is that a protein with a small truncation at the 3' end is perhaps less damaging to the cell than a reduction in protein levels, which would be the result from NMD.

In RNA-sequencing data, a heterozygous stop-gained variant which triggers NMD can be detected as allele-specific expression of the gene, with a depletion of the stop-gained allele (Rivas et al., 2015) (exemplified in figure 1-5). In a heterozygous cell where one allele of a gene contains a premature stop codon, both alleles are transcribed in the nucleus. Once the mRNA reaches the cytoplasm, NMD machinery recognizes the premature stop codon and specifically degrades those transcripts. When mRNA is isolated from the cell for RNA-sequencing, there is a strong enrichment for the wild type allele.

Stop-gained variants in rare disease

Stop-gained variants are enriched in rare disease-causing variants (Holbrook et al., 2004). These variants can result in disease whether they trigger NMD or not. Variants which trigger NMD can result in an up to 50% reduction in the amount of a given protein in a cell. If the gene is haploinsufficient, this can result in disease. On the other hand, if the stop-gained variant escapes NMD (which is likely if it falls after the 50-55 bp cutoff), it can result in a truncated version of the protein which can result in dominant negative effects of the protein which can be severe.

For some disease-associated genes, there are stop-gained variants both at the beginning of the transcript, likely triggering NMD, and at the end of the transcript, likely escaping NMD. Depending on where the variant is within the transcript, the disease can manifest with different symptoms or method of inheritance (Miller and Pearce, 2014).

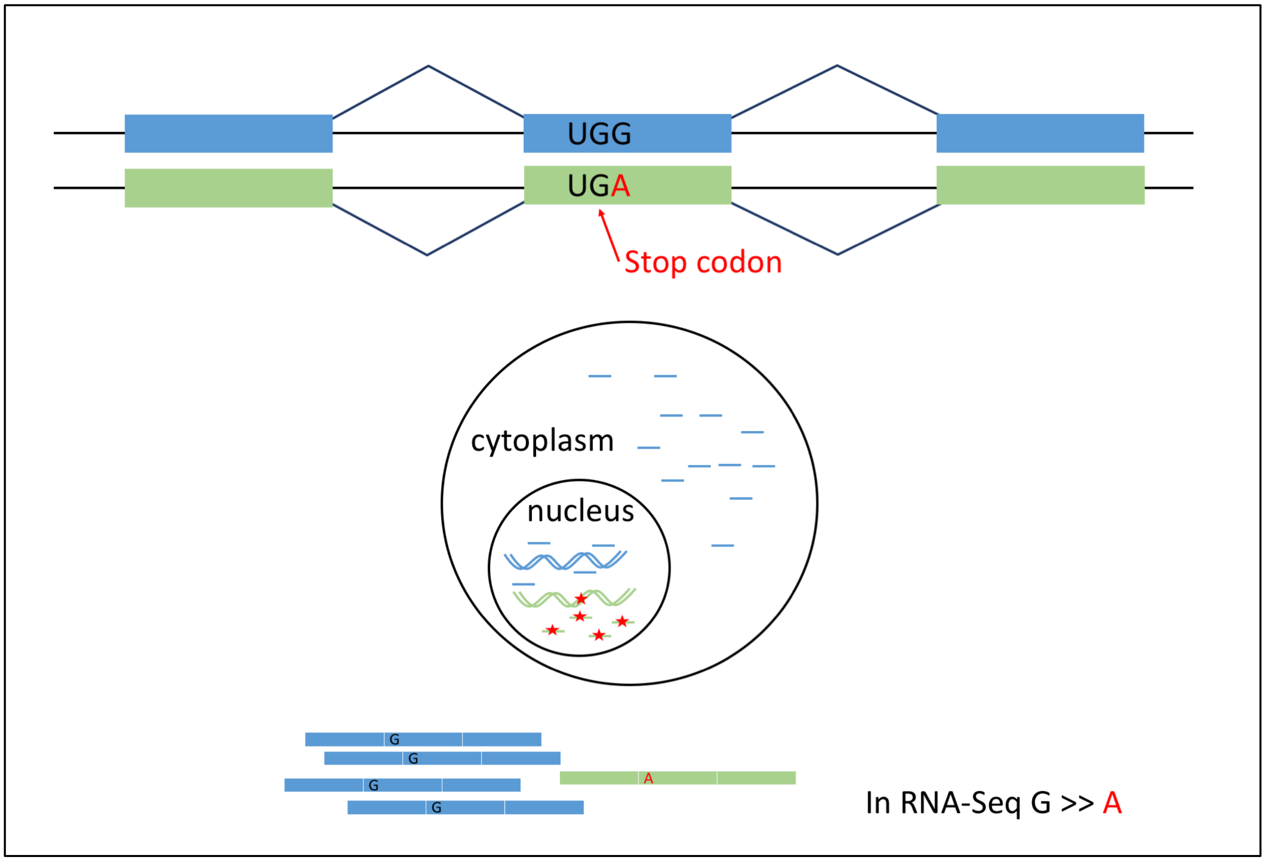


Figure 1-5. Stop-gained variants trigger nonsense mediated decay, which is detected in RNA-sequencing. A visual example of a cell which is heterozygous, with one reference allele and one stop-gained allele. The stop-gained variant introduces a premature stop codon into the transcript. In the cell, both

alleles are transcribed normally, but after transcription, the NMD complex degrades the transcripts with the stop-gained allele. Thus, the resulting effect of the stop-gained variant is a decreased presence of the stop-gained allele in the mRNA, which can be observed in RNA-sequencing data.

V. Innate immune response to LPS

The innate immune system is integral to the body's rapid response to invading pathogens. Cells of the innate immune system are responsible for recognizing pathogens, quickly activating a transcriptional response, and releasing signaling molecules to inform other cells throughout the body of the infection. Many immune cell-surface receptors responsible for recognizing particular pathogenic molecules belong to the Toll-like receptor (TLR) family. The members of the TLR family each recognize specific molecules characteristic of bacteria, fungi, protozoa and viruses (Takeda and Akira, 2005).

Extracellular recognition of lipopolysaccharide (LPS), a cell-surface marker on gram-negative bacteria, is transduced through the *trans*-membrane protein TLR4. TLR4 binds to extracellular LPS cooperatively with extracellular proteins CD14 and MD2 (Rosadini and Kagan, 2017). The binding of LPS triggers dimerization of TLR4, resulting in activation of two parallel signaling pathways: MyD88-dependent and MyD88-independent (figure 1-6). The MyD88-dependent pathway is responsible for inducing transcription of pro-inflammatory cytokines and the MyD88-independent pathway induces transcription of type-I interferons and their targets. Notably, both pathways act through activation of the TF NF-kB, whose targets include a broad array of pro-inflammatory chemokines and cytokines (Liu et al., 2017). In macrophages and monocytes, this release of cytokines in response to LPS stimulation is important for activating and recruiting cells of the innate and adaptive immune system as well as stimulating local tissue inflammation (Turner et al., 2014).

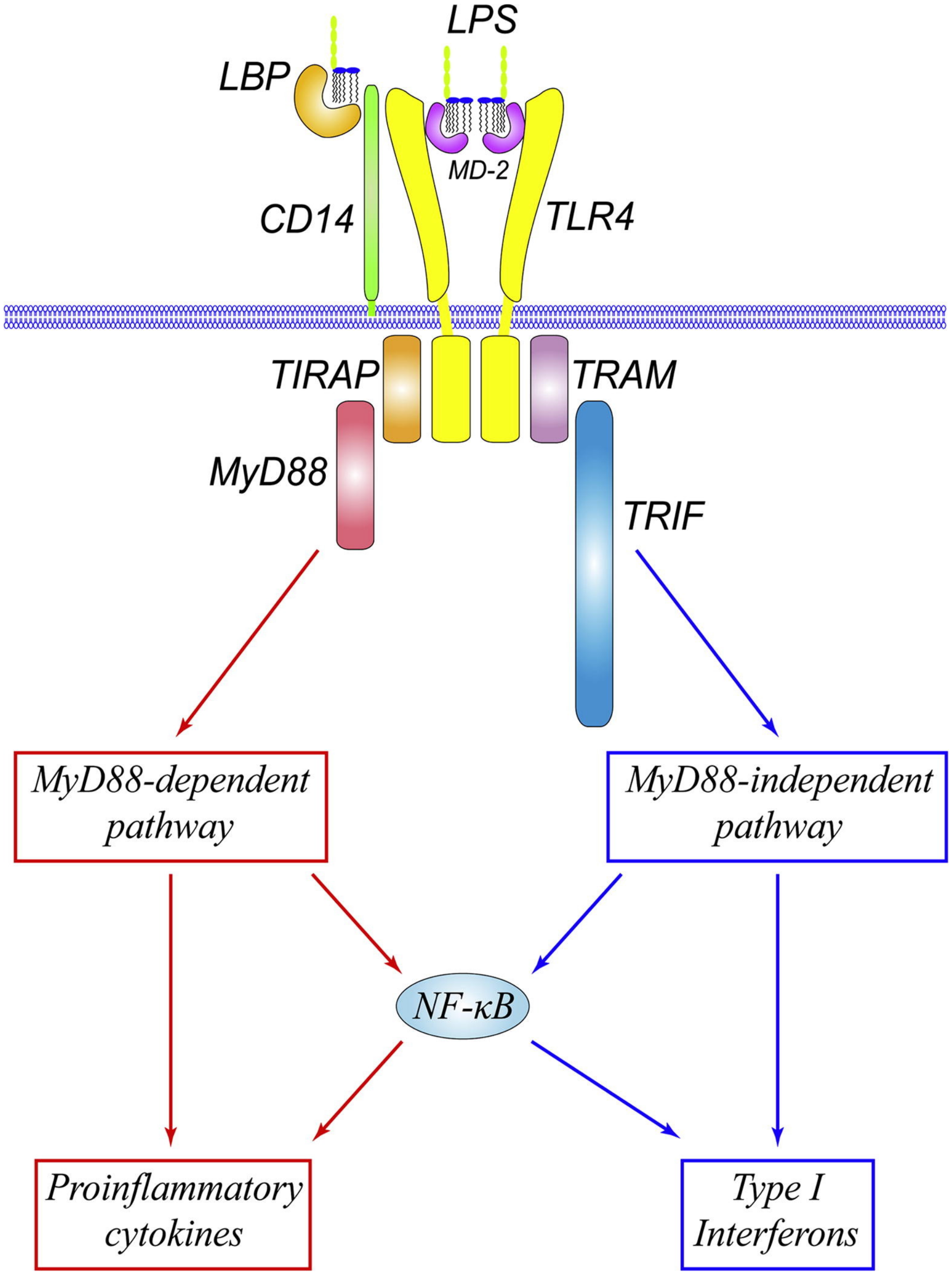


Figure 1-6. LPS-triggered signal transduction through the TLR4 receptor. TLR4, located on the surface of innate immune cells, binds LPS with cooperation from CD14 and MD-2. This binding activates two branches of signaling, the MyD88-dependent pathway and the My88-independent pathway mediated by TRIF. Both pathways activate NF-kB and induce expression of cytokines and type I interferons. Adapted from (Lu et al., 2008).

VI. Interferon-regulatory factor 1

Interferon-regulatory factor 1 (*IRF1*) is a member of a family of nine IRF transcription factors. The IRF TFs were first discovered for their regulation of expression of type I interferon genes and they recognize variations on the interferon-stimulated response element in the genome. In addition to interferon regulation, IRFs are involved in many facets of innate and adaptive immunity, oncogenesis and metabolism (Honda and Taniguchi, 2006; Tamura et al., 2008; Zhao et al., 2015).

IRF1 was the first member of the IRF family to be identified and was originally shown to be necessary for type-I interferon induction upon viral stimulation (Matsuyama et al., 1993). In addition, it has been implicated in lymphoid and myeloid lineage development, with abnormalities in immune cell development observed in irf1 knockout mice (Abdollahi et al., 1991; Matsuyama et al., 1993; Ogasawara et al., 1998). *IRF1* has additionally been shown to be involved in activation of macrophages via the IFN-γ receptor (Langlais et al., 2016). Also in macrophages, *IRF1* has been shown to be essential for induction of inducible nitric oxide synthase (iNOS), which produces nitric oxide, an important cell signaling molecule released by innate immune cells in response to infection (Kamijo et al., 1994). In the same study, *irf1* knockout mice were shown to be more susceptible to infection with mycobacterium. Another study showed that *IRF1*-deficient macrophages have reduced IL-12 cytokine induction after stimulation with LPS (Liu et al., 2003). Taken together, past research suggests an important role for *IRF1* in inducing immune-response genes in response to infection.

IRF1 is thought to be a target of NF-kB due to conserved NF-kB TF binding sites in its promoter (Harada et al., 1994; Iwanaszko and Kimmel, 2015). Additionally, *IRF1* is strongly induced after stimulation with LPS in monocytes (Kim-Hellmuth et al., 2017). Taken together with the fact that NF-KB is a major signal transducer upon TLR4 activation, it is possible that *IRF1* transcription is induced by LPS via NF-kB regulation. Additionally, *IRF1* is further implicated in TLR signaling, through direct activation by MyD88 (Negishi et al., 2006). Altogether, there is plausibility for *IRF1* being involved in TLR4 signaling through induction by NF-KB and activation by MyD88. This connection is further bolstered by the finding that *irf1* knockout mice have decreased survival after infection with LPS (Pan et al., 2013), implicating *IRF1* in LPS signaling. However, the specific role *IRF1* plays in LPS signaling has yet to be revealed.

VII. Summary and thesis aims

In genetic association studies, such as eQTL mapping or GWAS, it is not always clear whether an associated variant is causal or linked to a causal variant through LD. Computational methods to identify causal variants, such as various fine-mapping approaches, can improve our estimation of causality, but fall short in cases of high LD and provide probabilistic estimates rather than empirical evidence. Therefore, experimental methods to validate causal regulatory variants are essential. MPRA approaches attempt to validate these variants in high throughput, but in doing so remove the variants from their endogenous loci, and cannot address variants within the transcript. In studies of rare stop-gained variants, there is often a strong hypothesis for the mechanism of the causal variant, but since the variant is often observed in only one or few individuals, statistical evidence is difficult to obtain from the extremely small sample size. This makes it difficult to establish the functional mechanism without experimental approaches.

In this thesis, we aimed to validate and characterize regulatory variants found in the human population using CRISPR/Cas9 genome editing. In chapter 2, we first focused on transcript variants from the general population which are associated with gene expression. This includes rare stop-gained variants which are suspected to act through the NMD pathway, and common eQTL variants which lie in the transcript. Next, we tested our ability to validate the effects of disease-causing regulatory variants, focusing on premature stop-gained variants in two disease genes, *ROR2* and *GLI3*. The validation of causal regulatory variants is a central goal of functional genomics research. Here we demonstrate a reliable medium-throughput technique to detect regulatory effects of variants, which can be utilized as a valuable tool for variant validation.

Understanding how genetic variation affects expression of genes across the genome in *trans* can give us a deeper understanding of how variation affects broad transcriptional networks in the cell. In chapter 3, we interrogate a particularly interesting immune-response eQTL, which is a *cis*-eQTL for *IRF1* early after LPS stimulation and a *trans*-eQTL for many genes several hours after stimulation. To experimentally validate this association, we employed CRISPR/Cas9 genome editing and isolation of monoclonal cell lines followed by LPS stimulation and RNA-sequencing. Here we demonstrate the power of experimental investigation of master-regulatory variants, which contribute to inter-individual differences in cellular response to immune stimulus at a pathway level, potentially indicating mechanisms of disease-associated variants.

Chapter 2: A polyclonal allelic expression assay for detecting regulatory effects of transcript variants

The experiments described in this chapter were conceived and designed by Margot Brandt with guidance from Tuuli Lappalainen. All experiments and analyses described in this chapter were performed by Margot Brandt with the following exceptions: 1) Alper Gokden and Marcello Ziosi performed a replicate of the polyclonal assay on the stop-gained, eQTL and control edited variants. The results presented in this chapter will be published as a paper with Margot Brandt as the first author (Brandt et al, in submission. https://www.biorxiv.org/content/10.1101/794081v1).

Introduction

A method for scalable functional validation of regulatory variants associated with gene expression in human populations remains largely unaccomplished. Methods such as massively parallel reporter assays (MPRAs), which couple regulatory sequences with an expression reporter, are high-throughput and can be effective for finding active regulatory variants outside of the gene body, such as in enhancers. However, the results of the assays show low concordance with the direction of the population associations of the variants (Tewhey et al., 2016; van Arensbergen et al., 2019), perhaps due to taking the variant out of its genomic context. In addition, MPRAs are unsuited to validating variants found within the transcript: variants in exons, 5' UTRs or 3' UTRs. These variants are not compatible with MPRAs because they can act through post-transcriptional regulatory mechanisms affecting the stability of the transcript, as opposed to rate of transcription. A regulatory transcript variant can act by affecting splicing, inducing nonsense mediated decay (NMD) or impacting a miRNA binding site, for example. Posttranscriptional mechanisms have been shown to be important for regulatory variants, with eQTLs being strongly enriched for transcript annotations (Aguet et al., 2019; Lappalainen et al., 2013). In order to validate these transcript variants, it is essential to introduce the variant into the native genomic context so that it can be incorporated into the transcript when the gene is transcribed by the cell.

The advent of CRISPR/Cas9 genome editing technology (Cong et al., 2013; Jinek et al., 2012; Shalem et al., 2014) has provided an avenue with which to introduce specific variants into the genome of cells in order to validate their effects on expression. However, editing one variant at a time, isolating hundreds of single cell clones, genotyping and expanding clones and measuring transcript abundance in both edited and wild type clones is a hugely time-consuming, expensive process. In addition to the resource cost of completing such an experiment, undetected large on-target mutations (Kosicki et al., 2018), off-target mutations and other clone-specific genomic abnormalities can create noise which requires many replicates of each desired genotype in order to detect the effects of variants. To avoid undesirable clone-specific effects, we employed a method to validate the effects of regulatory variants in a polyclonal population of edited and wild type cells.

We first set out to validate eQTL variants from GTEx fibroblasts using CRISPR in a human embryonic kidney cell line (293T). In addition to the variants being significantly associated with expression of genes, the eQTLs were also fine-mapped using CAVIAR (Hormozdiari et al., 2014). The CAVIAR posterior probability of association (PPA) assigned to each eQTL-associated variant gives the probability that the variant is causal, which helps distinguish causal eQTL variants from those variants in high linkage disequilibrium (LD) with the causal variant. By selecting variants for editing that have strong associations with the eQTLs as well as a high PPAs from fine-mapping, the chances of the variant truly being causal are increased.

In addition to validating eQTL variants, this assay can also be used for validating the effects of stop-gained variants on transcript stability. Stop-gained variants located 50-55 bp or more before the last exon junction are thought to induce NMD, while variants located beyond this threshold are thought to escape NMD and therefore produce truncated protein (Nagy and Maquat, 1998). In order to establish this application, we first used the assay to analyze rare variants from individuals in GTEx which introduce a premature stop codon in the transcript prior to the 55-bp threshold.

Understanding the effect of stop-gained variants on the transcript could have clinical utility as well. If a patient has a variant of unknown significance that introduces a stop codon prematurely into the transcript, it could be beneficial to validate the variant experimentally to be sure of its mechanism of action. This could be especially useful in cases where the disease mode of inheritance or severity of symptoms are dependent upon whether the variant induces NMD or results in the production of a truncated protein. To demonstrate the utility of the assay in this context, we selected two disease-associated genes *GLI3* and *ROR2*. Stop-gained variants towards the beginning of *GLI3* are associated with Greig cephalopolysyndactyly, while variants towards the end of the gene are associated with the clinically distinct Pallister-Hall syndrome (Johnston et al., 2005). It is hypothesized that Greig cephalopolysyndactyly is caused by haploinsufficiency of *GLI3*, while Pallister-Hall syndrome is caused by a dominant negative effect of truncated *GLI3* protein. Similarly, stop-gained variants towards the beginning of *ROR2* are associated with the autosomal recessive Robinow syndrome, while variants towards the end of the transcript are associated with autosomal dominant Brachydactyly type B1 (Schwabe et al., 2000). Brachydactyly type B is thought to be caused by a dominant negative effect of truncated protein, since the recessive inheritance of Robinow Syndrome indicates haplosufficiency of the gene. The association between variant position in the transcript and clinical manifestation of disease is hypothesized to be dependent upon whether the variant induces NMD or produces a truncated protein. For these two disease genes, we edited both disease-associated variants from ClinVar (Landrum et al., 2018) and artificial stop-gained variants falling on both sides of the NMD cutoff.

Methods

fgwas enrichment

First, we sought to establish the relevance of testing eQTL effects driven by variants within transcripts by analyzing the extent of cis-eQTL enrichment in functional elements of the genome. We used GTEx v6 fibroblast eQTL data and a diverse set of annotations: Gene annotations were obtained from GENCODE (Harrow et al., 2012), and regulatory annotations (CTCF-binding site, enhancer, open chromatin region, promoter, promoter-flanking region, and TF binding site) were obtained from the Ensembl regulatory build release 80 (Zerbino et al., 2015). Additional annotations include CADD variant consequence scores (Kircher et al., 2014), SPIDEX machine-learning based prediction of splicing effects (Xiong et al., 2015), experimentally validated miRNA binding sites from Tarbase (Vergoulis et al., 2012), 3' UTR regulatory elements (Oikonomou et al., 2014), and RNA-binding protein sites from CLIPdb (Yang et al., 2015). Significant fibroblast eQTLs were analyzed for enrichment in these functional annotations using fgwas (Pickrell, 2014), with each annotation tested separately. Annotations that are significantly enriched in eVariants over other non-significant variants include many annotations found within the transcript (figure 2-1a). These include annotations involved in splicing, such as splice region variant, splice acceptor variant and SPIDEX predicted splicing. Additionally, variants are enriched for annotations that affect transcript stability either through triggering nonsense mediated decay (stop-gained variant, frameshift variant) or by disrupting a binding site of a regulator within the transcript (miRNA binding site, RNA binding protein site, 3' UTR variant). Rare stop-gained variants in GTEx within the transcript affect transcript stability in an allele-specific manner (figure 2-1b). The bimodal pattern of allelic expression in stop-gained variants suggests that some variants affect the abundance of the transcript they are found in, as demonstrated by a high reference ratio (calculated as ref reads/ total reads), while some maintain a roughly 0.5 reference ratio. These findings emphasize the importance of regulatory variants within the transcript and the need for a method to validate them.

Assay design

In order to validate transcript regulatory variants' allelic effects on transcript abundance, we utilized CRISPR/Cas9 genome editing with a gRNA specific to the locus of the variant of interest and a single-stranded DNA (ssDNA) template containing the alternative allele for homology-directed repair (HDR) (Figure 2-1c). For each variant of interest, we transfected the gRNA and ssDNA template into a well of inducible Cas9 293T cells. After editing, cells were harvested for gDNA and mRNA, followed by amplicon sequencing of the locus of interest in each. A regulatory effect of the variant is detected as a difference in the ratio of the alternative allele between gDNA and mRNA (figure 2-1d). This effect size is calculated as the log ratio of the alternative allele in cDNA over the ratio of the alternative allele in gDNA: log2(cDNA alt/ref / gDNA alt/ref), or the allelic fold change (aFC).

Variant selection

In this study, we edited five types of variants: GTEx stop-gained, GTEx eQTL, disease gene stop-gained, non-eQTL synonymous control, and synthetic control variants.

Stop-gained variants from the general population were obtained from the GTEx v6 data release. Starting with all stop-gained variants that were singletons in GTEx v6, we used allele-specific expression (ASE) data from the fibroblast sample of the individual carrying the variant to select those that are likely triggering NMD. The selected variants have RNA-seq coverage of >=20 reads, a reference ratio Ref/(Ref+Alt) > 0.7, and are located in a gene with > 5 RPKM in a published HEK293 RNA-seq dataset (Sultan et al., 2014). Additionally, we required ASE data in at least 5 tissues and a first quartile of ASE across tissues of > 0.7 to select variants where NMD does not appear to be highly tissue-specific. Finally, we selected variants > 30 bp from the end of an exon for primer design. Nine variants were used for editing.

eQTL variants were obtained from the GTEx v8 data release. Significant eQTL variants in fibroblasts were filtered for being within at least one protein-coding transcript, having a CAVIAR fine-mapping posterior probability of association > 0.8, an eGene with > 1 RPKM in HEK293 cells, and an effect size in the top quartile of effect sizes of all associations (aFC > 0.30). The top 33 highest effect size variants with successful gRNA and primer design were chosen for editing.

Ten stop-gained variants for each of the disease genes *GLI3* and *ROR2* were created by changing a codon in the transcript to a stop codon. The stop codons were spaced 20 bp apart in both directions from the NMD cutoff point (55 bp upstream of last the exon-exon junction). The 6 disease variants tested were obtained from ClinVar (Landrum et al., 2018), choosing disease-associated variants in the two genes on either side of the NMD threshold.

We selected 30 non-eQTL negative control variants from common synonymous variants in GTEx v8 data with an eQTL association p > 0.1 with the gene in which they reside. The templates for the 35 synthetic control variants were designed by introducing a nucleotide other than the reference or alternative allele at the stop-gained variant locus, which does not create a premature stop codon.

All variants edited with the assay can be found in Table 2-2.

Cell culture

Genome editing was carried out in a doxycycline-inducible Cas9 293T cell line, transduced with pCW-Cas9 plasmid (Addgene plasmid #50661 (Wang et al., 2014)), courtesy of the Sagi Shapira lab. 293T cells were cultured in OptiMEM (Gibco) supplemented with 5% HyClone Cosmic Calf Serum (Fisher), 1% Glutamax (Gibco), 1% NaPyr (Corning), and 1% penicillin/streptomycin (Corning). The cells were passaged and maintained following standard techniques in 5% CO2 and 95% air.

Genome editing

The protocol for the polyclonal editing assay can be found at dx.doi.org/10.17504/protocols.io.7c6hize. gRNAs were designed with E-CRISP version 5.3 (Heigwer et al., 2014) using medium settings, with an NGG PAM, a 5' G, excluding designs with more than 5 off-targets, and classifying off-targets as having up to 3 mismatches in 5' region of the gRNA. gRNAs were ordered as gBlocks gene fragments (IDT): a U6 promoter sequence followed by the specific gRNA and tracr sequence (Arbab et al., 2015). The gBlocks were amplified using Q5 high fidelity 2X master mix (NEB) and locus-independent gBlock amplification primers (Arbab et al., 2015). Homologous templates were designed by extracting the sequence 50 bp upstream and downstream of each variant and substituting the reference allele with the alternative allele. Stop-gained control templates have another nucleotide substituted in the variant position which does not create a stop codon. Homologous templates were synthesized as ultramers by IDT. If possible, primers which amplify both cDNA and gDNA were designed using IDT primer quest, choosing those that cover the PCR target (region spanning the variant and DSB) with at least 15 bp between the PCR target and one primer and at least 60 bp to the other primer. Otherwise, cDNA- and gDNA-specific primers were designed using either the cDNA or gDNA sequence as the template. Nextera adapter sequences were appended to forward and reverse primer sequences as follows:

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG+ForwardPrimerSequence

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG+ReversePrimerSequence

Primers were ordered as standard oligos from IDT.

Twenty-four hours before transfection for CRISPR editing, iCas9 293T cells were plated in 24-well plates and induced with 5 ug/mL of doxycycline, with a separate well for each targeted variant. Cells were transfected with 500 ng homologous template and 500 ng gRNA gblock using Lipofectamine MessengerMAX transfection reagent (Thermo Fisher Scientific). After 24 hours, transfection reagent was removed and replaced with new media. Cells were split after 4 days and 6 days, and DNA and RNA were extracted from the polyclonal edited cultures at 9 days. 75% of the 24-well culture was harvested for RNA using IBI Isolate DNA/RNA Reagent (IBI Scientific) according to the manufacturer’s instructions. Purified RNA was quantified by Nanodrop (Thermo Fisher). cDNA was synthesized with ~200 ng of purified RNA using 1/4 reactions of SuperScript IV VILO Master Mix with EZ DNase (Invitrogen). Another 10% of the cell culture was used for DNA extraction using 15 uL of QuickExtract (Lucigen). For the timecourse optimization experiment, mRNA and gDNA was extracted as above at days 4, 6 and 9.

Library preparation

Amplicon libraries from cDNA and gDNA were created using either the same nextera primers (if possible) or separate nextera primers for cDNA and gDNA. 1 uL of cDNA or gDNA was amplified using Q5 High Fidelity 2X Master Mix (NEB). An indexing PCR was performed next using Nextera XT index kit primers (Illumina) and NEBNext High-Fidelity 2X PCR Master Mix (NEB) resulting in dual barcoded amplicons with illumina adapters. cDNA and gDNA libraries were mixed in equal volume and sequenced on the MiSeq using 150 bp paired-end reads. We obtained a median coverage of 85,000 reads per sample.

Sequencing analysis

Fastqs generated from Illumina software were trimmed for adapter sequences and quality using trimmomatic. Reads were aligned to the gDNA or cDNA sequence specific for each amplicon and categorized as HDR, no edit, or NHEJ using EdiTyper (Yahi et al. in prep). Variants were eliminated if HDR in gDNA was greater than 30% (suggesting the cell line is in fact heterozygous for the variant). Samples were filtered out if they had fewer than 1000 reads covering the locus of interest. Additionally, samples were filtered out if they had an outlier NHEJ rate of greater than 80%, indicative of an alignment error. The effect size for each variant was calculated as the $\log_2$((Alt/Ref in cDNA) / (Alt/Ref in gDNA)), or allelic fold change (aFC). An effect size of zero means the variant has no effect on transcript abundance.

Statistical analysis

Significance between the control variant distribution and the other experimental variant types was determined using a two-sided Wilcoxon rank sum test. An F test was utilized to detect a difference in variance of aFC between non-eQTL control and synthetic control variants, and eQTL and control variants. For each individual regulatory variant, a p-value was calculated from the z-score of the variant's effect size based on the mean and standard deviation of the control distribution. The p-values were then bonferroni corrected and variants with a corrected p-value of less than 0.05 were considered significant.

eQTL effect size in GTEx

For the GTEx effect sizes for the eQTLs, we used the allelic fold change (aFC) estimates from the GTEx v8 data release (Aguet et al., 2019; GTEx Consortium et al., 2017; Mohammadi et al., 2017). For eQTL effect size in each GTEx tissue, we used the aFC estimates calculated from eQTL data. For stop-gained variants, we calculated aFC as the $\log_2$ ratio of the alternative and reference allele counts in the RNA-seq data in GTEx. To analyze the variation of eQTL variants' effects across GTEx individuals, we calculated the aFC for each eQTL variant across heterozygous individuals in GTEx using the alternative and reference allele counts in the gene body (Mohammadi et al., 2017). Samples were filtered for those with greater than 50 reads covering heterozygous sites in the gene.

eQTL variant characteristics analysis

LocusZoom plots were created for the eQTL variants using the p-values obtained from the GTEx v8 eQTL dataset and R^2 linkage disequilibrium estimates from the same dataset. Predictions of the molecular effects of the eQTL variants for Table 2-1 were obtained using the Ensembl Variant Effect Predictor tool (McLaren et al., 2016).

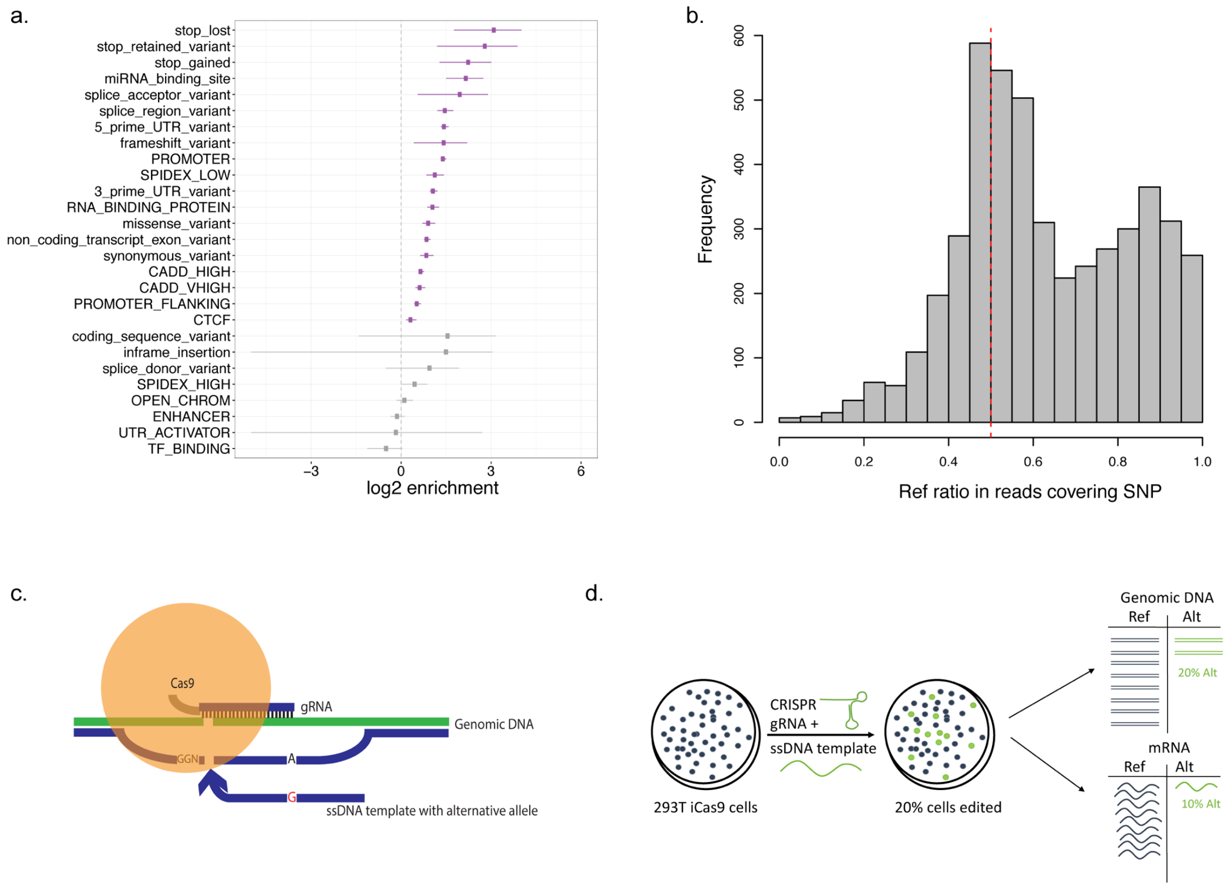


Figure 2-1. Variants found in the transcript are important for regulation of transcript abundance. (a) fgwas enrichment of functional annotations in GTEx fibroblast eVariants. Significant annotations are colored in purple. (b) Distribution of reference allelic ratio (reference reads/ total reads) in rare stop-gained variants found in individuals in GTEx. (c) For each potential regulatory variant of interest, we designed a gRNA and ssDNA template specific to the variant locus. (d) Polyclonal allelic expression assay to validate regulatory variants found in transcript. Inducible Cas9 293T cells undergo homologous recombination after transfection with the gRNA and ssDNA template in order to introduce the alternative allele to the cells. Editing is followed by targeted sequencing of gDNA and mRNA to detect the ratio of ALT/REF alleles in the polyclonal population of cells.

Results

Polyclonal allelic expression assay is a replicable method to validate regulatory variants

First, we assessed the right timepoint to harvest mRNA after transfection with CRISPR constructs. Since mRNA is likely to remain in the cell for hours to days after editing has occurred, we expect to see a depletion in edited mRNA molecules early after transfection. To find the optimal timepoint, we edited 17 control variants in 17 different genes which are not expected to have an effect on expression and harvested at three timepoints post-transfection: 4 days, 6 days and 9 days. At four days, we do see a depletion in the edited allele in the mRNA (figure 2-2a). However, this effect is lessened after 6 days and gone by 9 days. Therefore, we used the 9-day timepoint for the assay in order to analyze only the mRNA which has been transcribed post-editing.

In order to determine the optimal set of negative control variants, we compared the distribution of effect sizes of the synthetic control variants (new variants created in the same genes as the stop-gained variants) and non-eQTL variants (common synonymous variants where eQTL effects were tested in GTEx and not observed). The synthetic control variants have several outlier variants with large effect sizes that are consistent in replicates (figure 2-2b). This suggests that a subset of synthetic variants affect transcript levels and are thus not ideal negative controls. The non-eQTL control variants, however, have effect sizes consistently close to zero (median aFC = -0.009), demonstrating the utility of population data in selecting nonfunctional negative control variants. The variance of the synthetic controls was significantly greater than the variance of the non-eQTL controls (1.02 versus 0.038; F test $p = 3.7 \times 10^{-8}$). The non-eQTL variants were thus utilized as the control distribution for comparison with the stop-gained and eQTL variants tested with the assay.

Next, we analyzed how technical variation in editing efficiency or PCR amplification may affect the robustness of the assay. We compared the HDR rate with standard deviation of effect size between editing replicates of 62 variants (2 replicates for 23 and 3 replicates for 32 variants).

We found that low HDR appears to be associated with a higher standard deviation in the calculated effect size from the assay (figure 2-2c). Therefore, moving forward, we discarded any variant with an HDR rate of less than 0.4% as determined by amplicon sequencing of the gDNA. The effect size is well correlated between two replicates of the same variants (Spearman's rho = 0.53, p = 2.8x10^{-3}, figure 2-2d). The HDR rate is variable between loci, but very well correlated between replicates (Spearman's rho = 0.95, p = 7x10^{-16}, figure 2-2e), suggesting that the results of the assay are not strongly influenced by PCR amplification bias or variation in transfection efficiency.

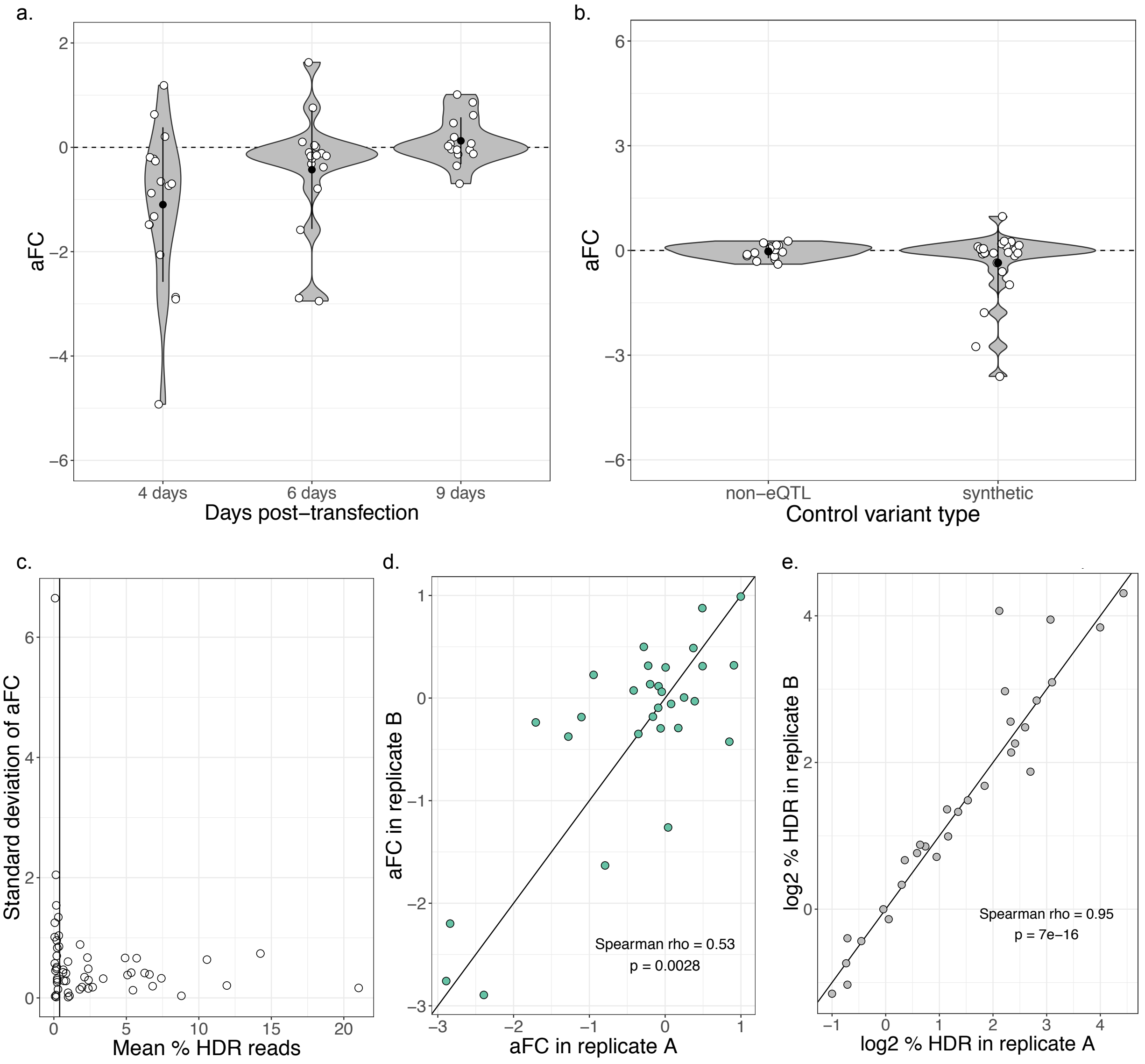


Figure 2-2. The polyclonal allelic expression assay is a replicable measure of variants' allelic effects on transcript abundance. (a) Distribution of effect size for control variants over a timecourse post-transfection. (b) Distribution of effect size for different control variant types: Synthetic and GTEx synonymous non-eQTL. (c) Homologous recombination rate versus standard deviation of effect size for variants replicated 2-3 times with assay. Vertical line shows 0.4% HDR cutoff which was used to filter variants for subsequent analysis. (d) Scatter plot showing reproducibility of effect size (allelic fold change) detected by polyclonal allelic expression assay in two editing replicates of the same variants. (e) Scatter plot showing correlation of HDR between the two replicates.

Edited stop-gained variants cause an allele-specific decrease in mRNA

The first variants selected for evaluation using the polyclonal allelic expression assay were rare stop-gained variants from GTEx. The alternative allele of these variants introduces a stop codon into the gene and shows evidence of allelic expression in GTEx heterozygous individuals as demonstrated by a decreased presence of the alternative allele in the mRNA (reference reads/total reads > 0.7 in GTEx).

The stop variants were expected to have a negative effect size by triggering NMD in the cell, which is what we observed (figure 2-3a), with the distribution of the effect size of the stop variants being significantly lower than the non-eQTL control variants (Wilcoxon signed rank $p = 3.26x10^{-5}$). The fact that the effect size of the edited stop-gained variants is in the expected direction consistent with NMD, an effect which is absent in the control variants, lends support for the assay capturing NMD effects of variants. As compared to the control variants, five of the stop-gained variants deviate significantly from the control distribution (Bonferroni-corrected z-test $p<0.05$).

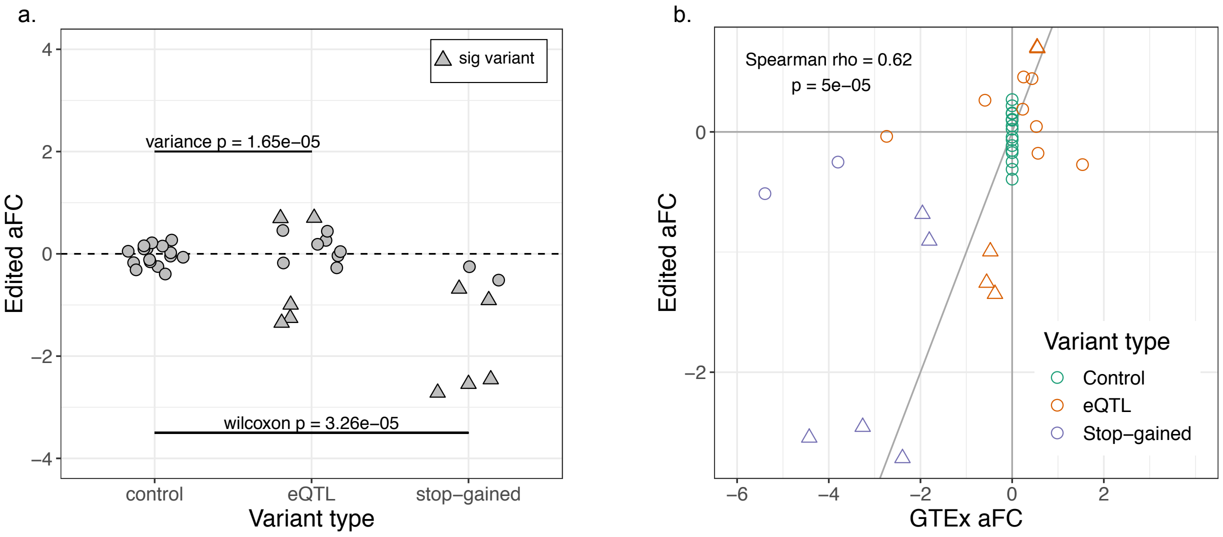


Figure 2-3. Stop-gained and eQTL variants from GTEx show allele-specific regulatory effects on expression. (a) Effect size of non-eQTL control, eQTL and stop-gained variants after editing with the polyclonal assay. Triangular points mark variants whose effect sizes significantly deviate from the control distribution. (b) Correlation between effect size observed in GTEx and effect size resulting from polyclonal assay for non-eQTL control, eQTL and stop-gained variants. Triangular points mark variants whose effect sizes significantly deviate from the control distribution.

eQTL variants have a larger regulatory effect than control variants

Next, we extended the assay to assess putatively causal eQTL variants within transcripts using GTEx fibroblast eQTLs. We chose fibroblasts because GTEx fibroblast transcriptome expression is highly correlated with that of HEK293 cells (rho = 0.68, $p < 2.2x10^{-16}$). The 33 eQTL variants chosen for editing are located within the transcript of the eGene with which they are associated and have a high posterior probability of causality based on CAVIAR fine mapping. After editing and QC filtering, 13 eQTL variants remained. The variance of the effect size of the eQTL variants was significantly higher than that of the control variants (0.49 versus 0.038; F-test $p = 1.65x10^{-5}$; figure 2-3a), which suggests that the edited eQTL variants as a whole have a greater regulatory effect than the edited control variants. Ten of the 13 variants have an effect in the same direction as the GTEx eQTL effect. Five of the eQTL variants are individually significantly different from the control distribution (figure 2-3a), and all five of these variants have an effect in the same direction as in GTEx. Additionally, there is a significant correlation between the effect size of the edited stop-gained, non-eQTL control and eQTL variants and their effects in GTEx (Spearman's rho = 0.62; p = 5x10^{-5}; figure 2-3b), again indicating that the assay captures regulatory effects seen in the population.

eQTL variant effects vary across tissues and individuals

The lack of effect observed for some of the eQTL variants could be due to our cell line not perfectly recapitulating the genetic regulatory effects in GTEx fibroblast samples. To investigate this, we looked at variation in effect size between GTEx tissues for each of the eQTL variants (figure 2-4a). We also looked at inter-individual variation within fibroblast samples in GTEx, which may reflect more subtle cell type-specific genetic effects as well as the effects of other regulatory variants in the GTEx individuals. We measured the effect size in eQTL heterozygotes in GTEx based on allelic imbalance within the gene body (figure 2-4b), with eleven of the eQTL variants having sufficient data for this analysis. For all five significant variants, there is agreement in direction between the polyclonal aFC, median heterozygous aFC across individuals and median eQTL aFC across tissues. For several of the other variants, figures 2-4a and b demonstrate a large range of effects both across tissues and across individuals. The observed effect of a variant in the cell line of our assay, like an individual or tissue, is likely to fall somewhere in a spectrum of possible effects. We note that for some of the variants the effect detected in our assay is consistent with the eQTL effect but the assay does not currently have sufficient sensitivity to assign this as a significant effect.

Another even more likely possible explanation for the lack of effect in some eQTL variants is that the top variant was incorrectly identified by fine mapping and there is another true causal variant. To test this hypothesis, we looked at the p-value distributions for the edited eQTL variants to see if there are other highly significant variants that could be driving the eQTL association. There does not appear to be a difference in the p-value distributions between the significant (figure 2-4c and d) and non-significant eQTL variants (figure 2-4e and f). In both cases, there are examples of distributions which appear to have other likely causal variants (figure 2-4c and e) and those that seem to have a clear top variant (figure 2-4d and f).

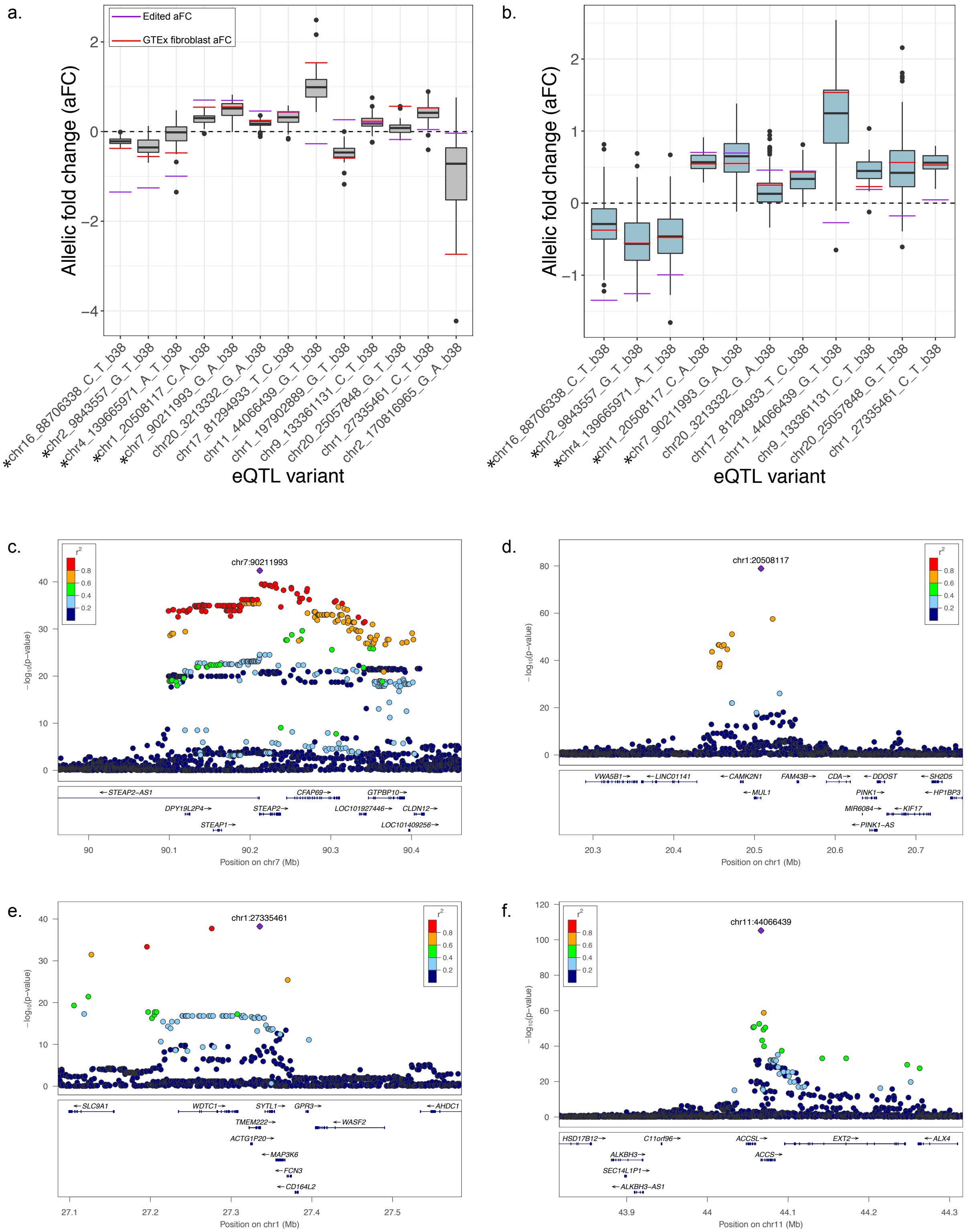

a.
Edited aFC
GTEx fibroblast aFC
Allelic fold change (aFC)
2
0
−2
−4
*chr16_88706338_C_T_b38
*chr2_9843557_G_T_b38
*chr4_139665971_A_T_b38
*chr1_20508117_C_A_b38
*chr7_90211993_G_A_b38
chr20_3213332_G_A_b38
chr17_81294933_T_C_b38
chr11_44066439_G_T_b38
chr1_197902889_G_T_b38
chr9_133361131_C_T_b38
chr20_25057848_G_T_b38
chr1_27335461_C_T_b38
chr2_170816965_G_A_b38
eQTL variant
b.
Allelic fold change (aFC)
2
1
0
−1
*chr16_88706338_C_T_b38
*chr2_9843557_G_T_b38
*chr4_139665971_A_T_b38
*chr1_20508117_C_A_b38
*chr7_90211993_G_A_b38
chr20_3213332_G_A_b38
chr17_81294933_T_C_b38
chr11_44066439_G_T_b38
chr9_133361131_C_T_b38
chr20_25057848_G_T_b38
chr1_27335461_C_T_b38
eQTL variant
c.
chr7:90211993
r^2
0.8
0.6
0.4
0.2
$-\log_{10}$(p−value)
40
30
20
10
0
←STEAP2−AS1
CFAP69→
GTPBP10→
DPY19L2P4→
STEAP2→
LOC101927446→
CLDN12→
STEAP1→
LOC101409256→
90
90.1
90.2
90.3
90.4
Position on chr7 (Mb)
d.
chr1:20508117
r^2
0.8
0.6
0.4
0.2
$-\log_{10}$(p−value)
80
60
40
20
0
VWA5B1→
←LINC01141
←CAMK2N1
FAM43B→
CDA→
←DDOST
←SH2D5
←MUL1
PINK1→
←HP1BP3
MIR6084→
←KIF17
←PINK1−AS
20.3
20.4
20.5
20.6
20.7
Position on chr1 (Mb)
e.
chr1:27335461
r^2
0.8
0.6
0.4
0.2
$-\log_{10}$(p−value)
40
30
20
10
0
←SLC9A1
WDTC1→
SYTL1→
GPR3→
←AHDC1
TMEM222→
←WASF2
ACTG1P20→
←MAP3K6
←FCN3
←CD164L2
27.1
27.2
27.3
27.4
27.5
Position on chr1 (Mb)
f.
chr11:44066439
r^2
0.8
0.6
0.4
0.2
$-\log_{10}$(p−value)
120
100
80
60
40
20
0
HSD17B12→
C11orf96→
ACCSL→
EXT2→
←ALX4
ALKBH3→
ACCS→
SEC14L1P1→
←ALKBH3−AS1
43.9
44
44.1
44.2
44.3
Position on chr11 (Mb)

Figure 2-4. Characteristics of the eQTL edited variants. (a) eQTL effect size (aFC) in GTEx tissues for the 13 edited eQTL variants shown as boxplots, with lines indicating the median effect size in GTEx fibroblasts (red) and in the assay (purple). Asterisks mark variants which were significant in the assay. (b) aFC in GTEx fibroblasts, measured in eQTL heterozygous individuals for 11 of the edited eQTL variants. (c-d) Locus zoom plots showing p-values and LD for variants tested for eQTL association for two representative significant eQTL variants from the polyclonal assay. (e-f) Locus zoom plots showing p-values and LD for variants tested for eQTL association for two representative non-significant eQTL variants from the polyclonal assay.

To further test the hypothesis that some of the non-significant eQTL variants are not causal, we used the Ensembl variant effect predictor (VEP) (McLaren et al., 2016) to determine the likely molecular consequences of the eQTL variants on their respective genes (Table 2-1). All five of the significant eQTL variants fall into the promoter region of the gene, while five out of eight non-significant variants fall into promoters. Interestingly, three out of five significant eQTLs fall into high-info positions within transcription factor (TF) motifs, suggesting they have an impact on binding of the TF, while none of the nonsignificant variants fall into high-info positions (fisher test, p = 0.035). This result suggests that the assay can distinguish true causal variants disrupting genetic regulatory elements from eQTL variants that are not causal.

To compare our assay to another method of eQTL validation, we compared our results to the results from a large MPRA study (van Arensbergen et al., 2019). Of the three of our significant eQTL variants that are found in their data, one variant (chr4_139665971_A_T_b38) has a nominally significant negative effect in both cell types used in that study, consistent with our results. Three of our non-significant eQTL variants (chr17_81294933_T_C_b38, chr20_25057848_G_T_b38 and chr2_170816965_G_A_b38) are nominally significant in that study and two are not. The overlap of significant variants is not expected to be especially high since the assays likely have different sensitivities and in the MPRA the variants are tested out of the genomic context.

	Variant	5' UTR	Promoter	TF motif	High-info	3' UTR	Splice region
Significant variants	chr1_20508117_C_A_b38	X	X	X	X		
	chr16_88706338_C_T_b38	X	X				
	chr2_9843557_G_T_b38		X	X	X		X
	chr4_139665971_A_T_b38	X	X	X	X		
	chr7_90211993_G_A_b38	X	X				
Non-significant variants	chr1_27335461_C_T_b38						
	chr1_197902889_G_T_b38		X				
	chr11_44066439_G_T_b38	X	X	X			
	chr17_81294933_T_C_b38	X	X				
	chr2_170816965_G_A_b38	X	X	X			
	chr20_3213332_G_A_b38						
	chr20_25057848_G_T_b38		X				
	chr9_133361131_C_T_b38						

Table 2-1. Variant effect predictions for the eQTL edited variants. Predicted effects of the edited eQTL variants based on Ensembl Variant Effect Predictions.

The polyclonal assay distinguishes disease variants which cause NMD from those that do not

In order to apply our assay to the detection of nonsense-mediated decay triggered by disease-associated variants, we introduced stop-gained variants into two disease-associated genes: *ROR2* and *GLI3*. Seven of the edited stop-gained variants fall before the 55 bp threshold and were therefore expected to trigger NMD. Of these variants, all seven resulted in negative effect sizes and the distribution of these variants was significantly different from that of both the four variants which were not expected to trigger NMD (Wilcoxon p = 6.1x10^{-3}) and the non-eQTL control variants (Wilcoxon p = 5.8x10^{-4}, figure 2-5a). When tested individually, six of the seven expected

NMD variants are significantly different from the control distribution, indicating that we can sensitively detect NMD and NMD- escape across the 55-bp boundary in these two genes.

In addition to the newly created stop-gained variants, we also included disease-causing stop-gained variants from ClinVar. The Arg442Ter mutation in *ROR2* results in a stop-codon right before the predicted NMD cutoff and is associated with the recessively inherited Robinow syndrome. We observe a significant negative effect of this variant (aFC = -1.39, Bonferroni corrected z-test p = $2.2\text{x}10^{-11}$), which is consistent with NMD and the clinical manifestation of disease (figure 2-5b). In contrast, the variant Trp749Ter is associated with dominant Type B brachydactyly and falls after the NMD cutoff in the transcript. Our assay shows that Trp749Ter does not affect the expression level of *ROR2* and therefore does not appear to be triggering NMD (aFC = -0.17, corrected p = 1). The one disease variant tested in *GLI3*, Arg792Ter, falling immediately before the predicted border of NMD escape, shows evidence of triggering NMD with a negative effect size in the assay (aFC = -1.02, corrected p = $3.6\text{x}10^{-6}$). This result is consistent with the clinical features of this variant, since it is associated with Grieg cephalopolysyndactyly syndrome which is thought to be caused by haploinsufficiency in the gene *GLI3*. The results of editing stop-gained variants in these disease genes indicate that there is a sharp cutoff of NMD / NMD-escape at the previously described 50-55 bp threshold and pinpoint the immediate molecular mechanism of NMD / NMD-escape for these disease variants. Additionally, the results demonstrate the potential for utilizing this assay to assess whether a variant of clinical interest triggers NMD when it falls close to the threshold of NMD escape.

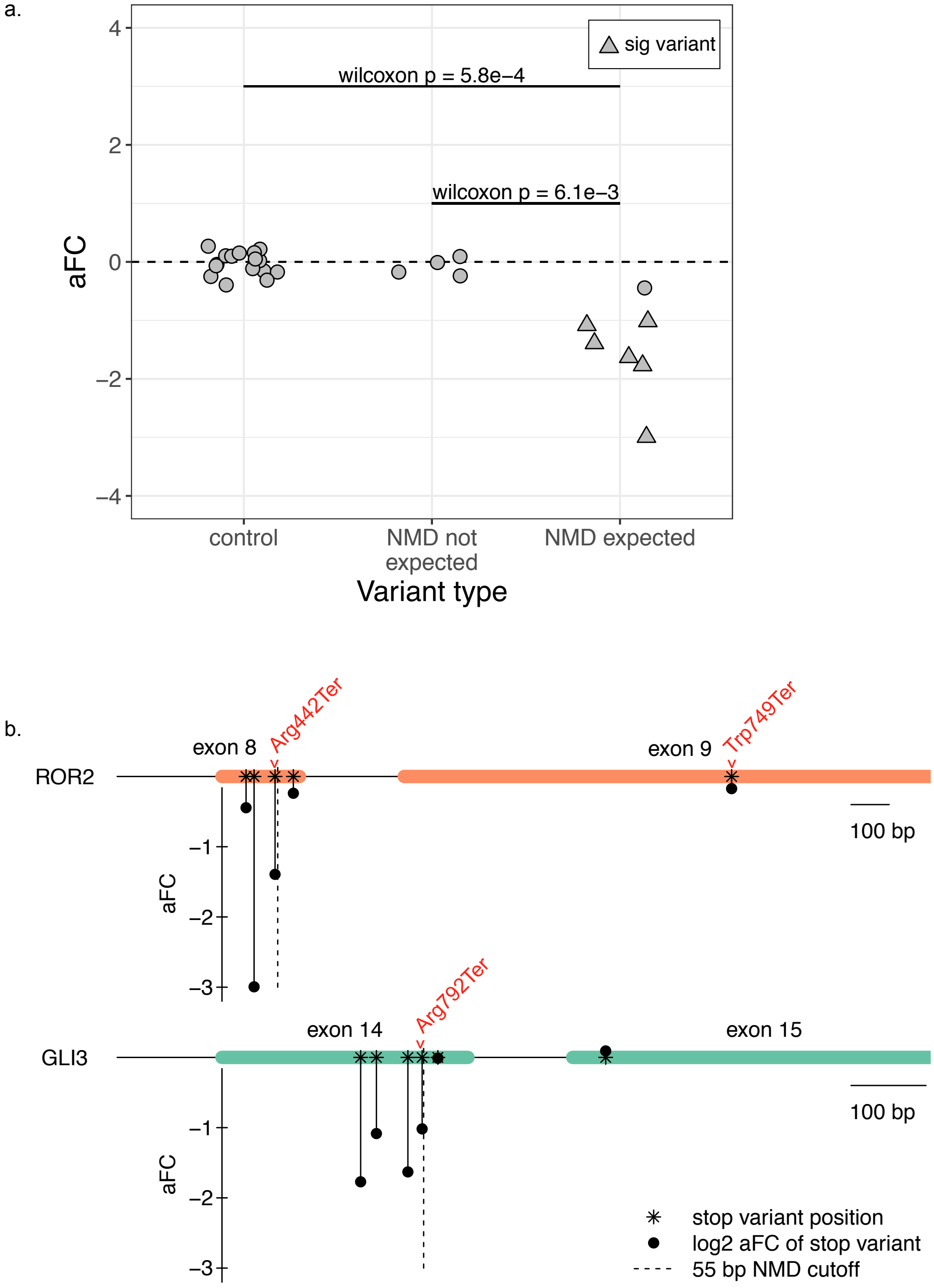


Figure 2-5. Stop-gained variants in disease-associated genes show expected regulatory effect based on position in transcript. (a) Effect size in control variants, stop-gained variants after NMD threshold, and stop-gained variants before NMD threshold. Triangular points mark variants whose effect size significantly deviates from the control distribution. (b) Diagram of the last two exons of NMD disease genes *ROR2* and

GLI3, showing the effect size (y-axis) and position in the transcript (x-axis) for each successfully edited variant. Disease- associated variants from ClinVar are labeled in red.

Discussion

In this study, we described a method utilizing CRISPR/Cas9 genome editing and targeted sequencing to validate regulatory variants without the need for isolating monoclonal cell lines. We demonstrated our ability to reliably detect the effects of stop-gained variants in the general population and in disease cases with the assay. The ability to experimentally assess the effect of potentially disease-causing stop-gained variants could lead not only to better understanding of the rules of NMD / NMD-escape, but also more accurate diagnosis and prognosis. The American College of Medical Genomics recommends caution in interpreting pathogenicity of stop-gained or frameshift variants of unknown significance, especially in cases where the variant is in an exon which might be alternatively spliced, or close to the 3' end of the transcript (Richards et al., 2015). Even though RNA analysis from patients is increasingly used to support variant interpretation (Ben-Shachar et al., 2009; Cummings et al., 2017; Kremer et al., 2017), establishing causality has been difficult since lower expression of a mutant haplotype or gene could be driven by other genetic or environmental factors. Our approach provides evidence that introduction of the specific variant in question underlies transcript level changes, thus reducing ambiguity. Furthermore, for genes where NMD / NMD-escape is clinically relevant, saturation editing at the 50-55-bp border could build a high-resolution reference for variant interpretation.

This polyclonal assay has the ideal throughput for identifying causal variants from a list of a few to several dozen candidate variants discovered from a rare genetic study. It would be feasible to perform the polyclonal assay on a number of potential regulatory variants, sequencing mRNA and gDNA from the polyclonal culture, and then sort monoclonal cell lines from the same polyclonal culture for only the variants which demonstrate allele-specific regulatory activity.

When we applied the polyclonal assay to eQTL variants, we detected increased effects on expression levels as compared to controls, often in the same direction as the GTEx eQTL effect. Five of 13 variants had significant effects, all consistent with the GTEx eQTL data. This clearly demonstrates the ability of our assay to capture common regulatory variant effects. Some of the non-significant eQTL variants appear to have edited effect sizes consistent with GTEx, but we lack the sensitivity to detect these small effects with confidence. In addition, some of the inconsistencies between the assay results and eQTL data are likely to originate from the eQTL data. Since we do not expect fine mapping to always succeed in identifying the true causal variants at these loci, the undetected effects could represent these situations. Furthermore, with multiple eQTLs for the same gene being common (GTEx Consortium et al., 2017), it is possible that eQTL effect sizes observed in populations reflect multiple regulatory variants in partial LD. Therefore, editing a single variant may not yield the same results as the full haplotype. When we looked at the aFC in heterozygous individuals for these variants in GTEx, we found a broad range of effect sizes, suggesting the presence of effects from multiple variants and potential modifiers that may not be captured by editing a single variant. Finally, assessing genetic regulatory effects even in closely matched cell lines does not necessarily capture effects measured in tissue samples. While this is likely to contribute to some of the differences, cis-eQTLs, especially in the transcribed region, are often highly robust across different tissues (GTEx Consortium et al., 2017), and are expected to replicate in cell lines as well. We highlight that our approach maintains the genomic context of variants and native gene regulation. Thus, it does not suffer from the limitations of massively parallel approaches where discrepancies between eQTL and experimental data may be due to measuring genetic regulatory effects in artificial constructs (Tewhey et al., 2016; van Arensbergen et al., 2019). Altogether, more experimentation and further comparison of population and experimental results are required to fully understand differences between experimental and population data.

Finally, we note that our assay is somewhat limited by HDR efficiency, which varies greatly between loci. Capturing the specific effect of the edited variant requires discarding any reads in the gDNA or cDNA which contain indels created through non-homologous end joining (NHEJ). Since NHEJ often dominates HDR in efficiency, this can result in low numbers of HDR reads. Research in improving the HDR rates in editing is ongoing (Aird et al., 2018; Chu et al., 2015; Maruyama et al., 2015), and likely HDR efficiency will be greatly improved in the future. Additionally, future improvements on base editor technology, which avoids the introduction of double stranded breaks and therefore minimizes the risk of indels (Gaudelli et al., 2017; Komor et al., 2016), could also benefit this system and increase sensitivity of the assay.

Table 2-2. All variants edited with the polyclonal allelic expression assay in chapter 2.

Variant type	Variant ID	chr	Chr pos GRCh38	REF allele	ALT allele	gRNA seq
stop gained	stop_1	1	10463557	G	A	GTGTGTGCTGCAGCCGCTGGA
stop gained	stop_2	1	113981146	C	T	GGGTGAAGTCACGCAGCCTT
stop gained	stop_3	1	946463	G	T	GCAGCTGCTTGGGAAGGTTC
stop gained	stop_5	19	18784264	G	A	GTGGAAGGACCCGCGAAACG
stop gained	stop_6	2	200651022	C	T	GGCATTGATCTCTTGTTTGTA
stop gained	stop_7	20	44948809	C	A	GGCAATGAGCTTGTAAAGAA
stop gained	stop_8	22	36512560	G	A	GAAGCTGGACTCTCAGCGAG
stop gained	stop_9	3	49122992	G	A	GCCCATCCTCATCTCGACAGC
stop gained	stop_10	5	154821355	C	T	GAGGGCCAAGGAAAACCACA
eQTL	chr1_20508117_C_A_b38	1	20508117	C	A	GTTTCCGGTCAGGTTAGGCC
eQTL	chr10_16817401_G_C_b38	10	16817401	G	C	GCGTGTTCGCTGTTCAGTGC
eQTL	chr11_44066439_G_T_b38	11	44066439	G	T	GGCTCCAGGTTTCCAGGCAG
eQTL	chr15_64156166_C_T_b38	15	64156166	C	T	GTGGAAGCAGGAGGGCATGG
eQTL	chr16_70157320_G_A_b38	16	70157320	G	A	GCAGCCTATTAGTTCTGGTG
eQTL	chr19_32972322_C_T_b38	19	32972322	C	T	GGTTCCTGCCGGCTGTATTC
eQTL	chr19_32972339_A_G_b38	19	32972339	A	G	GCTGTATTCGGGCCTTGGAC
eQTL	chr19_984554_C_G_b38	19	984554	C	G	GCAAGAATTACATCAGCGCC
eQTL	chr2_170816965_G_A_b38	2	170816965	G	A	GCCCAGCGATCCGCTCGGCT
eQTL	chr2_9843557_G_T_b38	2	9843557	G	T	GCCTCCTTACCGCCTCCTCG
eQTL	chr20_1325648_T_C_b38	20	1325648	T	C	GCTTTAAACTCCCCTGGCCT
eQTL	chr20_3213332_G_A_b38	20	3213332	G	A	GATAAGTGCCGGAGTACCAG
eQTL	chr22_36507049_C_T_b38	22	36507049	C	T	GCGCGGCCTCATTAGACCAC
eQTL	chr4_41990735_G_C_b38	4	41990735	G	C	GCTCCGTCTGCGATGCAGGG
eQTL	chr8_144414270_T_C_b38	8	144414270	T	C	GGCAGTGGGTGCAGTCACTG
eQTL	chr5_56909530_A_G_b38	5	56909530	A	G	GGTGCCAGGAACACTGAGAG
eQTL	chr3_33218930_C_T_b38	3	33218930	C	T	GCGCTCGGCTCACGAATCGC
eQTL	chr7_4768773_G_A_b38	7	4768773	G	A	GAGCGGGGAGAGTGGTGAGG
eQTL	chr19_9324196_C_T_b38	19	9324196	C	T	GCGTGGGCGCATGCGCATAA
eQTL	chr8_544804_C_G_b38	8	544804	C	G	GCCGCAGGCAGAGCGTCCGG
eQTL	chr16_89972488_G_C_b38	16	89972488	G	C	GATCAAACCCTCGAACGGTC
eQTL	chr4_139665971_A_T_b38	4	139665971	A	T	GAACTATTTGTAGAGCGCAC

eQTL	chr1_147647471_G_A_b38	1	147647471	G	A	GAGTTTGAGAGCAGAGTGCG
eQTL	chr1_27335461_C_T_b38	1	27335461	C	T	GGCCACCGAGCAGCCATCAC
eQTL	chr1_197902889_G_T_b38	1	197902889	G	T	GAGCGAAGAGTTAACCGCGG
eQTL	chr20_25057848_G_T_b38	20	25057848	G	T	GTCCCAGACGGTGTGGTAGG
eQTL	chr7_90211993_G_A_b38	7	90211993	G	A	GGGCGAGCCTTGCAGCTCCC
eQTL	chr3_100709512_A_G_b38	3	100709512	A	G	GAGAACTTGGGCTCTGTACG
eQTL	chr9_133361131_C_T_b38	9	133361131	C	T	GGCATGTGCTTTTATTAACC
eQTL	chr12_51199833_T_G_b38	12	51199833	T	G	GCTGAAGGTGGCAATGGCAG
eQTL	chr16_88706338_C_T_b38	16	88706338	C	T	GCGCGGGCCTGGCCCCGGGA
eQTL	chr19_57840921_G_A_b38	19	57840921	G	A	GACAGGTGTGTCTCCCAAGA
eQTL	chr17_81294933_T_C_b38	17	81294933	T	C	GGTCATAGTGAGAGGTCTAG
ROR2 stop	ROR2_exp_1	9	91725057-91725059	CAG	CTA	GTCTGCGGTGAGGTTCATGG
ROR2 stop	ROR2_exp_2	9	91725078-91725080	CGC	CTA	GCCATGAACCTCACCGCAGACAG
ROR2 stop	ROR2_exp_3	9	91725099-91725101	GAG	CTA	GCCATGAACCTCACCGCAGACAG
ROR2 stop	ROR2_exp_4	9	91726553-91726555	GTT	CTA	GGCATGGAGACCTGTTTGTGC
ROR2 stop	ROR2_exp_5	9	91726574-91726576	GTC	CTA	GCCATCAGCTGTCGCCGCTG
ROR2 stop	ROR2_exp_6	9	91726616-91726618	GGA	CTA	GCCATCAGCTGTCGCCGCTG
ROR2 stop	ROR2_exp_7	9	91726637-91726639	ATT	CTA	GCCATCAGCTGTCGCCGCTG
ROR2 stop	ROR2_exp_8	9	91726658-91726660	GAA	CTA	GAAAAGGCAAGCGATGACCAG
ROR2 stop	ROR2_exp_9	9	91726679-91726681	CAG	CTA	GAAAAGGCAAGCGATGACCAG
ROR2 stop	ROR2_exp_10	9	91726700-91726702	GAC	CTA	GAAAAGGCAAGCGATGACCAG
ROR2 stop	ROR2_exp_Arg442Ter	9	91726603	G	A	GCCATCAGCTGTCGCCGCTG
ROR2 stop	ROR2_exp_Trp720Ter	9	91724334	C	T	GCGGGACAGTCATCGGGGCA
ROR2 stop	ROR2_exp_Trp749Ter	9	91724248	C	T	GTAGTTGGAAAGGTTGCCCC
GLI3 stop	GLI3_exp_1	7	41966592-41966594	GAG	CTA	GGGCCCATGACGCTTCTCCC
GLI3 stop	GLI3_exp_2	7	41966613-41966615	CAA	CTA	GGGCCCATGACGCTTCTCCC
GLI3 stop	GLI3_exp_3	7	41966635-41966637	CTG	CTA	GTCCAACAACACCTGCAGCT
GLI3 stop	GLI3_exp_4	7	41967609-41967611	AGG	CTA	GAGAGACCGCAGGGGCTTTA
GLI3 stop	GLI3_exp_5	7	41967630-41967632	AGG	CTA	GAGAGACCGCAGGGGCTTTA

GLI3 stop	GLI3_exp_6	7	41967672-41967674	TTG	CTA	GAAAGGCTAAAACAAGTGAA
GLI3 stop	GLI3_exp_7	7	41967693-41967695	TAC	CTA	GAAACCCGGCAGGGACCAAA
GLI3 stop	GLI3_exp_8	7	41967714-41967716	CCC	CTA	GAAACCCGGCAGGGACCAAA
GLI3 stop	GLI3_exp_9	7	41967735-41967737	TTG	CTA	GGCTTGCAAAGCAAGGGCTG
GLI3 stop	GLI3_exp_10	7	41967756-41967758	TGC	CTA	GGCTTGCAAAGCAAGGGCTG
GLI3 stop	GLI3_exp_Ser856Ter	7	41966506	G	T	GTAGGCCGAGCTGATGGTGC
GLI3 stop	GLI3_exp_Arg792Ter	7	41967653	G	A	GGTAGAATGGGGTTCAGTCG
GLI3 stop	GLI3_exp_Gln717Ter	7	41967878	G	A	GTTGGAATAGTTGCTGATGG
synthetic control	stop_con_1	1	10463557	G	T	GTGTGTGCTGCAGCCGCTGGA
synthetic control	stop_con_2	1	113981146	C	A	GGGTGAAGTCACGCAGCCTT
synthetic control	stop_con_3	1	946463	G	A	GCAGCTGCTTGGGAAGGTTC
synthetic control	stop_con_5	19	18784264	G	T	GTGGAAGGACCCGCGAAACG
synthetic control	stop_con_6	2	200651022	C	A	GGCATTGATCTCTTGTTTGTA
synthetic control	stop_con_7	20	44948809	C	T	GGCAATGAGCTTGTAAAGAA
synthetic control	stop_con_8	22	36512560	G	T	GAAGCTGGACTCTCAGCGAG
synthetic control	stop_con_9	3	49122992	G	T	GCCCATCCTCATCTCGACAGC
synthetic control	stop_con_10	5	154821355	C	G	GAGGGCCAAGGAAAACCACA
non-eQTL control	syn1	11	88294225	A	C	GGTCTCTTAGACCAGTGTGG
non-eQTL control	syn2	14	104887068	G	A	GTGAGGGGGGCAGCACCCCG
non-eQTL control	syn3	17	40822211	G	A	GGAGGCGGCTTTGGTGGAGG
non-eQTL control	syn4	17	79086238	C	T	GAAATTCCATGCGACGATCC
non-eQTL control	syn5	17	80010342	C	T	GGATGTCACCGAGGAGGGGC
non-eQTL control	syn6	19	58294717	T	C	GGATGCGCTCATGCTGGACG

non-eQTL control	syn7	22	50523830	G	A	GGCAACCTGTTTGGTGGAGC
non-eQTL control	syn8	7	143863472	T	G	GAGTGGCCTCCTCGCAGTGG
non-eQTL control	syn9	1	27366484	G	T	GCCGGCTCCGCGCGCAGCCC
non-eQTL control	syn10	1	3883678	A	G	GTCACAGTTGAGCTTGTGGG
non-eQTL control	syn11	10	69180435	C	T	GTTTGGTATTTTGGAGCCAC
non-eQTL control	syn12	11	126304143	C	T	GCGCGAATTGGACGTGGAGG
non-eQTL control	syn13	14	76776147	C	T	GCGGGTCGTGTGACACGCTC
non-eQTL control	syn14	16	69330127	G	A	GGTTCACGAACACGCGCAGG
non-eQTL control	syn15	16	88715671	G	A	GCGGTAGAGGAAGATGAGCT
non-eQTL control	syn16	16	88721329	C	T	GGAGGGGCCAGGGGTGCCTG
non-eQTL control	syn17	17	15945284	G	T	GGACGCGCTGTACGTGGCGC
non-eQTL control	syn18	17	15945323	C	T	GGTCATCGCCGCGCTTTCGG
non-eQTL control	syn19	17	15999680	C	G	GGAGGATCCTGACCCCCCGC
non-eQTL control	syn20	17	59213022	G	A	GCCTGTCGATCAACGAAGTT
non-eQTL control	syn21	17	75909713	C	T	GGGGAGAGGCACGTGGCCAG
non-eQTL control	syn22	18	23529181	G	A	GGTTTTTTTCTTTCAGGCGG
non-eQTL control	syn23	19	3959446	G	A	GGCTGCGCTGCAGCTCGCGC

non-eQTL control	syn24	21	42903351	G	A	GTCTTTCTTGACAAAAATTT
non-eQTL control	syn25	5	10239241	C	T	GTTCTTACTTACTGGGCTTG
non-eQTL control	syn26	6	108074445	C	A	GCCCCAGCCCTCCACCCTGC
non-eQTL control	syn27	6	139167023	C	T	GCCGCCGAAGGAACTGCACG
non-eQTL control	syn28	7	156950085	C	A	GAACAGGGTCCCGGCCTGGG
non-eQTL control	syn29	7	726792	G	A	GGCCGAGACGGCTGAGGCGG
non-eQTL control	syn30	8	143650949	G	A	GCTGGGTGAAGTTTGACGTC
synthetic control	ROR2_con_1	9	91725057-91725059	CAG	TAG	GTCTGCGGTGAGGTTCATGG
synthetic control	ROR2_con_2	9	91725078-91725080	CGC	TGC	GCCATGAACCTCACCGCAGACAG
synthetic control	ROR2_con_3	9	91725099-91725101	GAG	AAG	GCCATGAACCTCACCGCAGACAG
synthetic control	ROR2_con_4	9	91726553-91726555	GTT	ATT	GGCATGGAGACCTGTTTGTGC
synthetic control	ROR2_con_5	9	91726574-91726576	GTC	ATC	GCCATCAGCTGTCGCCGCTG
synthetic control	ROR2_con_6	9	91726616-91726618	GGA	TGA	GCCATCAGCTGTCGCCGCTG
synthetic control	ROR2_con_7	9	91726637-91726639	ATT	GTT	GCCATCAGCTGTCGCCGCTG
synthetic control	ROR2_con_8	9	91726658-91726660	GAA	AAA	GAAAAGGCAAGCGATGACCAG
synthetic control	ROR2_con_9	9	91726679-91726681	CAG	TAG	GAAAAGGCAAGCGATGACCAG
synthetic control	ROR2_con_10	9	91726700-91726702	GAC	TAC	GAAAAGGCAAGCGATGACCAG
synthetic control	ROR2_con_Arg442Ter	9	91726603	G	T	GCCATCAGCTGTCGCCGCTG
synthetic control	ROR2_con_Trp720Ter	9	91724334	C	G	GCGGGACAGTCATCGGGGCA
synthetic control	ROR2_con_Trp749Ter	9	91724248	C	G	GTAGTTGGAAAGGTTGCCCC
synthetic control	GLI3_con_1	7	41966592-41966594	GAG	AAG	GGGCCCATGACGCTTCTCCC
synthetic control	GLI3_con_2	7	41966613-41966615	CAA	TAA	GGGCCCATGACGCTTCTCCC

synthetic control	GLI3_con_3	7	41966635-41966637	CTG	TTG	GTCCAACAACACCTGCAGCT
synthetic control	GLI3_con_4	7	41967609-41967611	AGG	GGG	GAGAGACCGCAGGGGCTTTA
synthetic control	GLI3_con_5	7	41967630-41967632	AGG	GGG	GAGAGACCGCAGGGGCTTTA
synthetic control	GLI3_con_6	7	41967672-41967674	TTG	CTG	GAAAGGCTAAAACAAGTGAA
synthetic control	GLI3_con_7	7	41967693-41967695	TAC	CAC	GAAACCCGGCAGGGACCAAA
synthetic control	GLI3_con_8	7	41967714-41967716	CCC	TCC	GAAACCCGGCAGGGACCAAA
synthetic control	GLI3_con_9	7	41967735-41967737	TTG	CTG	GGCTTGCAAAGCAAGGGCTG
synthetic control	GLI3_con_10	7	41967756-41967758	TGC	GGC	GGCTTGCAAAGCAAGGGCTG
synthetic control	GLI3_con_Ser856Ter	7	41966506	G	A	GTAGGCCGAGCTGATGGTGC
synthetic control	GLI3_con_Arg792Ter	7	41967653	G	T	GGTAGAATGGGGTTCAGTCG
synthetic control	GLI3_con_Gln717Ter	7	41967878	G	T	GTTGGAATAGTTGCTGATGG

Chapter 3: A distant *cis*-eQTL for *IRF1* is a *trans*-eQTL master regulatory variant in immune response

All experiments described in this chapter were conceived and designed by Margot Brandt with guidance from Tuuli Lappalainen. All experiment and analyses described in this chapter were performed by Margot Brandt with the following exceptions: 1) Sarah Kim-Hellmuth performed the immune-respone *cis*- and *trans*-eQTL study in which the IRF1 signal was discovered. 2) Aaron Wollman transfected the hTLR4 cells and performed the T7E1 assay to confirm editing in polyclonal cell population 3) Alper Gokden performed the long-range PCR on the clonal cell lines. 4) Marcello Ziosi performed the CRISPRi transfection and LPS stimulation of the hTLR4 cells, 5) Yocelyn Recinos tested the optimal time point for hTLR4 cell immune stimulus.

Introduction

Understanding the link between noncoding variants and their effects on genes can help interpret variants associated with disease which fall in the noncoding genome. Doing so in a static environmental context is how classic eQTL studies are done. However, looking for associations in specific environmental contexts can further reveal how regulatory variants act in a cell under variable conditions. This is especially relevant for immune cells, such as monocytes, whose physiological function is dependent upon rapid transcriptional changes in response to stimulus. Immune-response eQTL (reQTL) studies are designed with this goal in mind: to identify regulatory variants whose effects are dependent upon the presence or absence of a stimulus. Such variants can contribute to genetic inter-individual variation in immune response – a widespread phenomenon that underlies many human infectious and autoimmune diseases (Fairfax et al., 2014; Kim-Hellmuth et al., 2017; Lee et al., 2014).

The innate immune response to lipopolysaccharides (LPS), molecules found on the outer membrane of gram-negative bacteria, is transduced through toll-like receptor 4 (TLR4), which is found on the surface of many immune cells, including monocytes. TLR4 cooperates with MD2 and CD14 to recognize LPS outside the cell. Once bound to LPS, TLR4 undergoes a conformational change which triggers a signaling cascade, activating NF-κB and increasing transcription of its targets, including proinflammatory cytokines and type I interferons (Lu et al., 2008).

Immune-reQTL studies in monocytes have identified variants which affect the expression of genes involved in the innate immune response, which can explain quantitative differences in immune response between individuals (Chen et al., 2016; Fairfax et al., 2014; Kim et al., 2014). A *cis*-reQTL is where a variant is associated with expression of a proximal gene, whereas a *trans*-reQTL is where a variant is associated with a distal gene. A previous immune-reQTL study (Kim-Hellmuth et al., 2017) discovered 126 significant *cis*-reQTLs in human monocytes which were active under stimulation with LPS. One such *cis*-reQTL for *IRF1* is not significant under baseline but is a significant eQTL under stimulation with LPS. In addition to being implicated in immune cell lineage development, *IRF1* has been shown to be involved in activation of macrophages via the IFN-γ receptor (Langlais et al., 2016). Another study showed that *IRF1*-deficient macrophages have reduced IL-12 cytokine induction after stimulation with LPS (Liu et al., 2003), implicating *IRF1* in TLR4 signaling in response to LPS. Furthermore, *IRF1*-knockout mice have been shown to have increased survival after infection with LPS as compared to matched control mice (Pan et al., 2013). However, the full extent of the involvement of *IRF1* in TLR4 signaling has not yet been elucidated.

Trans-eQTLs, associations between variants and genes which are mediated by a secondary factor, are more difficult to detect than *cis*-eQTLs. In addition to the multiple testing burden imposed by testing all variants against all genes, *trans*-eQTLs also tend to have smaller effect sizes (Aguet et al., 2019). However, there is evidence that *trans*-eQTLs are particularly important for understanding mechanisms of disease-associated variants, as GWAS associations for complex traits are strongly enriched for trans-eQTLs (Aguet et al., 2019). Furthermore, since they provide information on how genetic perturbations affect transcriptional networks, trans-eQTLs have the potential to elucidate cellular pathways affected by genetic variants. However, while *trans*-eQTL detection has somewhat improved recently with larger sample sizes (Võsa et al., 2018), there has still been very little validation of these associations.

In this study, we first associated the *cis*-reQTL variants from Kim-Hellmuth et al. with all expressed genes in monocytes in order to detect *trans* associations. One *trans*-eQTL near *IRF1* is of particular interest due to its association with *IRF1* in *cis* early after LPS stimulation and delayed association with a large number of genes in *trans*. This result suggests the association is a condition-specific *trans*-eQTL with potentially broad effects on an individual's response to LPS. Therefore, we chose to follow up on this *trans*-eQTL master regulatory variant using CRISPRi gene expression perturbation and CRISPR/Cas9 genome editing in an immune cell line to replicate the variant's effects.

Methods

eQTL discovery

The *IRF1 cis*-reQTL was discovered in a previous immune-response eQTL study (Kim-Hellmuth et al, 2017). Briefly, primary monocytes were isolated from 134 donors and treated with LPS, MDP, IVT RNA, or no treatment. To assess early and late immune response, RNA expression was measured with a microarray chip after 90 min and 6 h. To detect a significant *cis*-reQTL, the effect size β of association between genotype and expression of genes were compared between untreated and treated samples. rs17622517 was found to be a significant *cis*-reQTL variant for *IRF1* at 90m after LPS treatment, i.e. under early immune stimulus.

Trans-eQTL discovery was done with the MatrixEQTL R package, performing a linear regression between the genotype of the top variant for each of the 126 significant *cis*-reQTL at LPS 90 min and all expressed probes in monocytes at 6 hours (22,657). Benjamini-Hochberg correction was performed across p-values for all variants and probes. A significance threshold of FDR < 0.5 was used for downstream analyses. Since rs17622517 was found to be associated with many genes in *trans* at 6 hours, in addition to the *cis* association with *IRF1*, it was selected for further follow up.

Enrichment of IRF1 targets

IRF1 targets were obtained from the molecular signatures database (MSigDB) gene set *IRF1*-01, which comprises genes which contain at least one *IRF1* motif in the 4kb upstream and downstream from their transcription start site (Liberzon et al., 2011). *Trans*-eGenes from the rs17622517 *trans*-eQTL (FDR < 0.5) and CRISPRi differentially expressed gene set (FDR < 0.05) were tested for enrichment in *IRF1* target genes using a Fisher's exact test comparing the number of genes in the set which overlap *IRF1* target genes with the number of background expressed genes which overlap *IRF1* targets.

Cell culture

HEK293/hTLR4A-MD2-CD14 Cells (Invivogen) were selected for functional follow up of the *IRF1 trans*-eQTL because of the ease of transfection of HEK293 cells and the addition of the TLR4 receptor, which is essential for cellular response to LPS. Cells were cultured in DMEM supplemented with 4.5 g/l glucose (Corning), 10% fetal bovine serum (Sigma-Aldrich), 1% penicillin/ streptomycin (Corning) and 1% L-glutaMAX (gibco). Cells were passaged using cell scraping to avoid damaging the cell surface receptors.

CRISPRi of IRF1 promoter and enhancer locus

In order to determine whether the region of the eQTL variant regulates IRF1, and the effect of IRF1 perturbation in our cell line, we performed CRISPRi experiments targeting both the variant locus and the IRF1 promoter. HEK293-TLR4 cells were plated in three 24-well plates with 120K cells/well in 1ml of DMEM. 24 hours later, the medium was replaced with 0.5ml of OptiMem (Gibco) and cells were transfected with 1.5ul of lipofectamine 3000 (Thermo Fisher Scientific), 200ng of CRISPR-KRAB-MeCP2 vector (Addgene 110821) and 50 ng of gRNA gblock (IDT) (4 wells each received a gRNA targeting EGFP as a neutral control ('GGTGGTGCAGATGAACTTCA'), 14 bp downstream of the IRF1 promoter ('GTCTTGCCTCGACTAAGGAG') or the enhancer ('TTCTCTGTAGCCCTTGTATT')). After 28 hours, 1ug/mL of LPS (Invivogen)was added to the 90 m and 12 h samples and nothing to the control samples, with four biological replicates of each gRNA for each LPS treatment (36 total samples). Cells were collected at their respective timepoint with TRIzol reagent (Thermo Fisher Scientific) and stored at -80C before RNA extraction.

Genome editing

In order to validate the *cis* and *trans* reQTL associations of rs17622517, we edited the HEK293-TLR4 cells using CRISPR/Cas9 genome editing. A gRNA (‘TTCTCTGTAGCCCTTGTATT’) was designed with an NGG pam and a cut site 1 bp downstream from rs17622517. The gRNA was ordered as a single stranded oligo gblock from IDT and amplified using 2 50 uL reactions of Q5 High Fidelity 2X Master Mix (NEB). Cells were transfected with 0.5 ug gRNA gblock and 2.5 ug px458 plasmid (Addgene plasmid # 48138) containing spCas9 and GFP with lipofectamine 3000 (Thermo Fisher Scientific). 24-hours later, cells then underwent fluorescence-activated cell sorting for GFP+ cells using a Sony SH800Z cell sorter to enrich for transfected cells. Efficiency of editing was tested using a T7E1 assay and electrophoresis gel to detect presence of NHEJ. The GFP+ cells were also sorted as single cells into 15 96-well plates and expanded into monoclonal cell lines.

Clones were genotyped by creating an amplicon library for each clone from gDNA using nextera primers capturing a 218 bp amplicon containing the variant locus. Indexing PCR was performed next using primers specific to the constant sequence on the nextera primers, resulting in dual-barcoded amplicons with illumina adapters. Libraries were mixed in equal volume and sequenced on the MiSeq using 150 bp paired-end reads. Fastq files generated by the Illumina software were trimmed for adapter sequences and quality using trimmomatic. Reads were aligned to the genomic locus and categorized as no edit or NHEJ using Edityper (Yahi et al. In prep). Since the cell line is triploid at this genomic locus, clones were considered heterozygous if they had NHEJ between 20-70%. Clones with less than 10% NHEJ were considered wild type and clones with greater than 90% NHEJ were considered homozygous edited. Five each of wild type, heterozygous and homozygous edited were selected for follow up. BAM files from the selected clones were also visually inspected to confirm genotype. In addition, a 3,496 bp PCR followed by electrophoresis gel was performed on the clones in order to check for larger indels not captured by the shorter amplicon libraries.

LPS treatment of edited clones

To detect the reQTL effect of the edited locus, we performed LPS treatment followed by RNA-sequencing on each isolated edited and wild type cell line. Each clonal cell line was plated into three wells (one well each for control, 90m and 12h treatments) in 24-well plates with 180,000 cells per well. 24 hours later, 1ug/mL of LPS (Invivogen) was added to the 90 m and 12 h samples. RNA was extracted at the designated timepoint by adding 500 uL of IBI Isolate DNA/RNA Reagent (IBI Scientific) directly to cells on the plate and stored at -80C until extraction.

RNA extraction and RNA-seq library preparation

RNA was extracted following the Direct-zol RNA MicroPrep kit (Zymo Research) manufacturer's instructions. RNA was treated with DNAse I (Ambion) and enriched for mRNA using Dynabeads mRNA DIRECT Purification Kit (Thermo Fisher). cDNA was generated using a custom scaled-down modification of the SMART-seq protocol (Picelli et al., 2014). cDNA was synthesized from RNA input using Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific). It was then amplified using Kapa HiFi 2X Ready Mix (Kapa Biosystems). cDNA was then cleaned using 0.9X Ampure beads. Finally, cleaned cDNA samples were tagmented and indexed using the Nextera XT DNA Library Prep Kit (Illumina). Library size and tagmentation was confirmed using the TapeStation HS D1000 kit (Agilent). Libraries were pooled in equal molarity and sequenced with the NextSeq 550 High-Output kit (Illumina) with paired end 2X75 reads.

RNA-seq analysis

Reads were first trimmed of adapters using trimmomatic, then aligned to the hg19 genome using STAR 2-pass mapping. Gene counts were calculated with FeatureCounts using gencode v19 gene annotations.

Differential expression analysis

For analysis of the effects of CRISPRi on the transcriptome, a count matrix was created in R from the raw counts of all samples. Principal component analysis was performed on the count matrix, after transformation and normalization with vst(). Differential expression was performed using an interaction model (~gRNA + condition + gRNA:condition) with DESeq2. Prior to p-value correction, genes were discarded if they did not have an average expression of greater than 5 read counts across samples and an annotation of protein coding or lncRNA in gencode. P-values for differential expression were corrected using Benjamini-Hochberg correction and a 5% FDR significance threshold was used.

For analysis of the effect of CRISPR editing the variant locus, a count matrix was created in R from the raw counts of all samples. Since library prep and RNA-Seq were performed twice, the count matrices from the two runs were summed together. Differential expression analysis was performed using the R package DEseq2 using a nested interaction model (expression ~ genotype + genotype:clone + genotype:condition) accounting for the same clone being found in each treatment group and testing for differential expression in the interaction between condition and genotype. Prior to p-value correction, genes were discarded if they did not have an average expression of greater than 5 read counts across samples and an annotation of protein coding or lncRNA in gencode. P-values for differential expression were corrected using Benjamini-Hochberg correction and a 5% FDR significance threshold was used.

Enrichment analysis on significantly differentially expressed genes was done using DAVID biological process gene ontology enrichment, using Benjamini-Hochberg corrected p values and 5% false discovery rate.

Comparison between edited clone and CRISPRi differential expression

For comparison of the differential expression fold changes between the edited clonal RNA-seq and the CRISPRi RNA-seq experiments, the genes in the edited differential expression analysis with an FDR of 25% were intersected with the genes in the CRISPRi differential expression analysis and a Spearman correlation test was performed.

Trans-reQTL comparison to differential expression of edited clones and CRISPRi samples

Some of the expression probes in the monocyte study measure the same gene. Therefore, the mean beta from the *trans*-eQTL study for probes overlapping the same gene was taken. The intersection of this gene list and the genes tested for differential expression was obtained. A Spearman correlation test was performed on the *trans*-reQTL betas and fold changes from DESeq2.

Results

Top variant for *IRF1 cis*-reQTL is in a likely enhancer

In the *cis*-reQTL study, the most significant variant for *IRF1* expression 90 minutes after stimulation with LPS is rs17622517, which is the top variant associated with *IRF1* expression under stimulation with LPS, but not at baseline (figure 3-1a). This variant is located about 23,000 bp downstream of the *IRF1* TSS in an intron of the gene *C5ORF56*. There are other neighboring variants which are significantly associated with *IRF1* expression as well (figure 3-1b). However, the genomic context of rs17622517 lends further support to its causality. In ENCODE DNase hypersensitivity assays, this locus overlaps a peak in many cell types, including monocytes (Thurman et al., 2012). Furthermore, ENCODE H3K27Ac ChIP-Seq in the GM12878 cell line, a lymphoblastoid cell line, suggests an active immune cell enhancer in this region (pink peaks in H3K27Ac track in figure 3-1c). Finally, ENCODE transcription factor ChIP-Seq data also show significant peaks at this locus for many transcription factors in lymphoblastoid cell lines (figure 3-1c). These include IRF4, another interferon regulatory factor, and RELA, a subunit of NF-κB, a key regulator in TLR4 signaling. Additionally, this variant has been found as a hit in a GWAS for chronic inflammatory diseases: ankylosing spondylitis, Crohn's disease and ulcerative colitis (Ellinghaus et al., 2016), lending further support for its involvement in immunity.

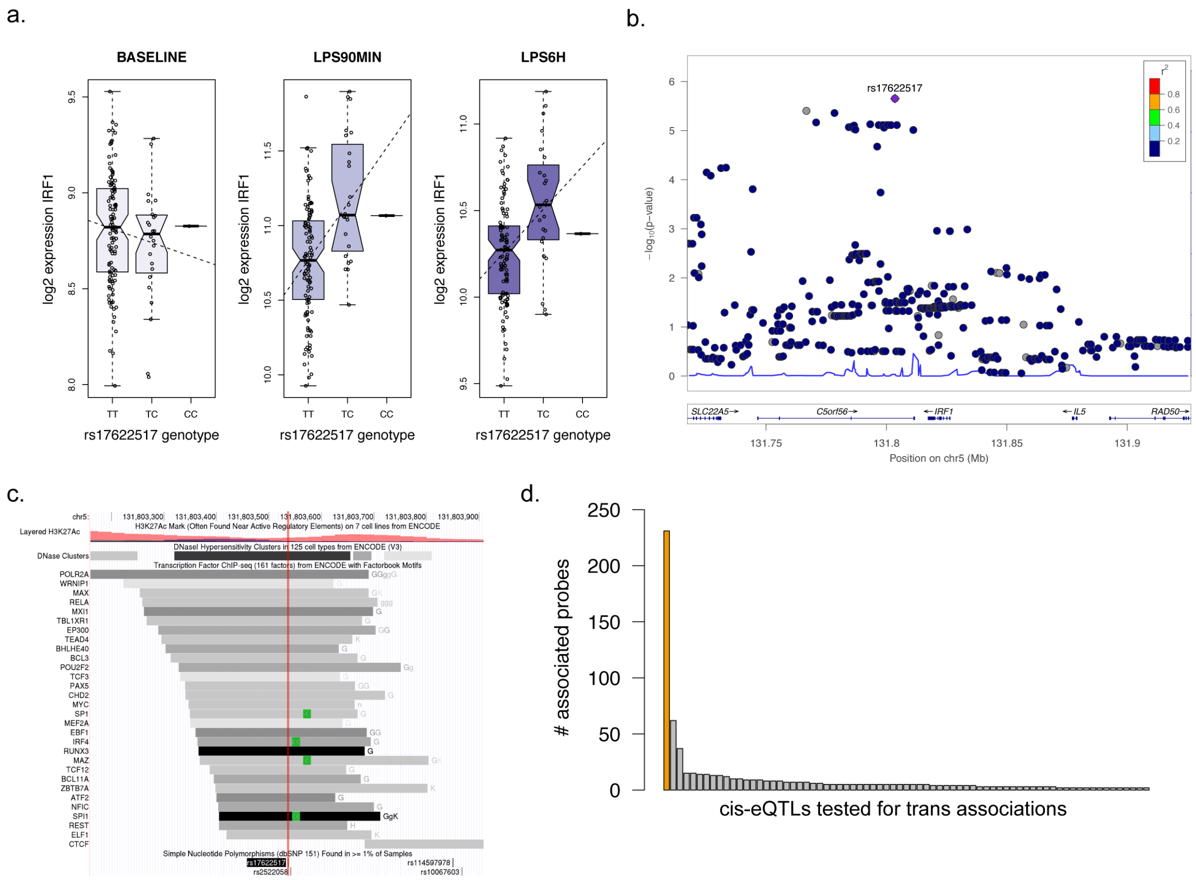


Figure 3-1. rs17622517 is the top variant associated with the *IRF1 cis-* and *trans-* reQTL under LPS stimulation. (a) *cis*-reQTL boxplots demonstrating *IRF1* expression versus genotype of rs17622517 in monocytes harvested from 132 individuals at baseline, 90 min and 6 hours after LPS stimulation. (b) LocusZoom plot of the *IRF1 cis*-reQTL, highlighting top variant rs17622517. (c) UCSC genome browser view of top variant rs17622517, demonstrating overlap with transcription factor binding, DNase hypersensitivity and H3K27Ac peaks. The red line shows the variant location. (d) The number of gene expression probes (y-axis) associated with *cis*-eQTL top variant (x-axis) at FDR 0.5 for those *cis*-eQTL variants with greater than one *trans* association. *IRF1* top *cis*-eQTL variant rs17622517 is highlighted in orange.

IRF1 cis-reQTL variant is a significant *trans*-reQTL for many target genes

The top variants for all 126 of the significant *cis*-reQTLs at LPS 90 min from Kim-Hellmuth, et al. was tested for association with the expression of all 22,657 expressed probes in monocytes. The *IRF1* variant is associated with 232 expression probes with a lenient FDR threshold of 0.5 (table 3-3), substantially more than the other *cis*-reQTLs (figure 3-1d). These results suggest that the *IRF1* variant acts as a master regulator *trans*-eQTL to modify expression of many genes. The FDR 0.5 *trans*-eQTL genes are enriched for MSigDB *IRF1* target genes, defined as having an *IRF1* motif within 4 kb of their TSS (Fisher test p = 1.8e-4). These *trans*-reQTLs are mostly not detected until 12 hours after LPS stimulation (figure 3-2a), suggesting a sequential expression effect with the *cis*-eQTL detectable at 90 minutes and the *trans*-eQTL detectable at 12 hours. This timing is consistent with the hypothesis that the *cis*-eQTL affects expression of *IRF1* first, which then affects the expression of downstream direct and indirect targets of *IRF1*. With a more stringent FDR threshold of 0.05, the *IRF1* variant is associated with 12 *trans*-eGenes, all but one of which have higher expression in individuals with the alternative allele (figure 3-2b), which is the same direction as the *cis*-eQTL. The discovery of the vast number of *trans*-associations for this variant inspired the next part of this study: replicating the *trans*-reQTL for *IRF1* experimentally.

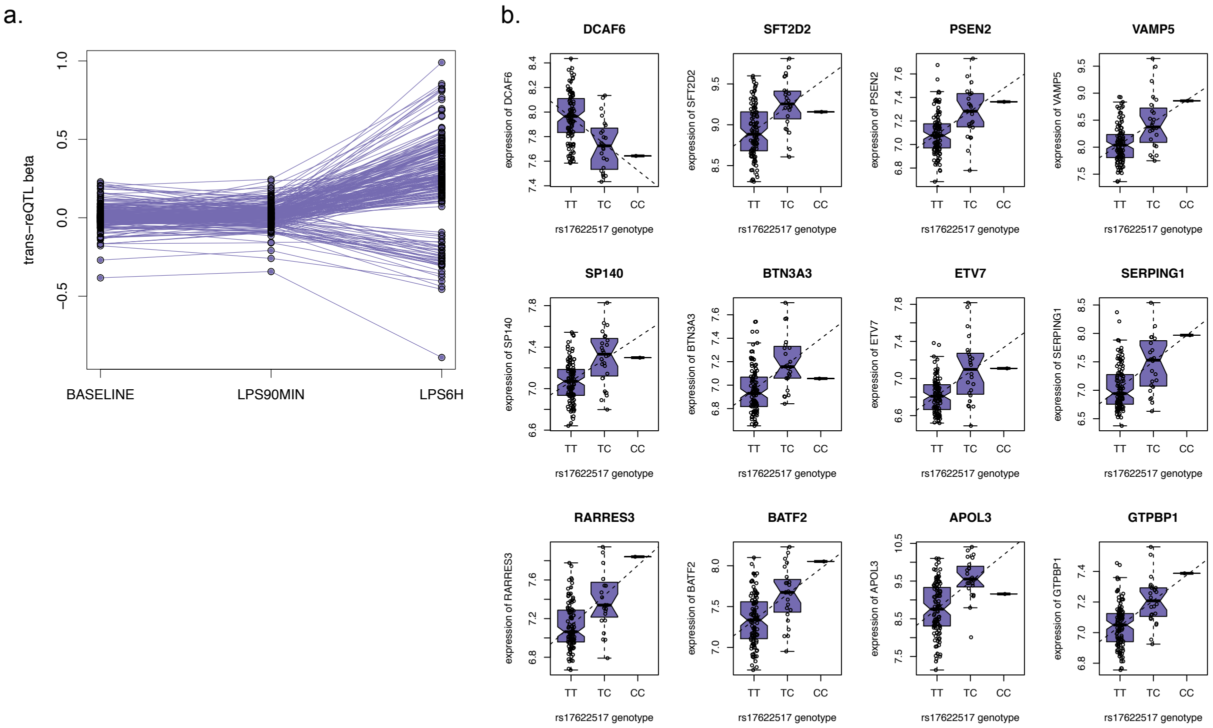


Figure 3-2. The *trans*-association for rs17622517 is detectable 6 hours after LPS stimulation, after the 90 min cis-eQTL for IRF1. (a) Beta of association with rs17622517 for the 216 *trans*-reQTLs (FDR 0.5) at baseline, 90 min and 6 hours after LPS stimulation. (b) Boxplots for the 12 significant *trans*-reQTLs (FDR 0.05) showing expression of *trans*-eGenes versus genotype of rs17622517 6 hours after LPS stimulation.

IRF1 enhancer locus appears to be an active enhancer in HEK-TLR4 cells

In order to investigate the effect of repressing the enhancer in which the IRF1 eQTL SNP is located, we transfected a gRNA at the variant locus along with a dCas9-KRAB-MeCP2 vector in HEK-TLR4 cells. In this CRISPRi system, the dCas9 enzyme is guided to the genomic locus of interest by the gRNA but lacks the enzymatic capacity for inducing double stranded breaks and is coupled to KRAB-MeCP2, a repressor which has been shown to strongly decrease expression of proximal genes (Yeo et al., 2018). To test the effect of repressing expression of IRF1 under LPS stimulation, we also transfected a gRNA targeting 14 bp downstream of the IRF1 TSS along with the dCas9-KRAB-MeCP2 vector (schematic in figure 3-3a). Additionally, we included a gRNA targeting EGFP as a control. After transfection with CRISPRi constructs, we treated four replicates of each transfection with LPS for 0h, 90m or 12h. We then performed RNA-sequencing on the 36 samples (3 gRNAs x 3 LPS conditions x 4 replicates) and mapped reads to known genes.

Upon principle component analysis, samples tend to cluster both by LPS condition, captured in principal component 1, and gRNA, captured in principal component 2 (figure 3-3b). This result suggests that both promoter and enhancer gRNAs are inducing strong effects on the transcriptome. When we look at the expression of IRF1 across samples, we see significantly higher IRF1 expression in LPS 90m samples (Wilcoxon p = 0.024) and LPS 12h samples (Wilcoxon p = 0.038) as compared to 0h samples. Additionally, we see a strong repression of *IRF1* in the promoter gRNA samples across LPS conditions as compared to the control (Wilcoxon $p = 7.4x10^{-7}$, figure 3-3c), indicating a successful repressive effect of the promoter gRNA. The enhancer gRNA samples do not have significantly different *IRF1* expression than the controls (Wilcoxon p = 0.8), but it is not expected to have as large of an effect as in the promoter gRNA.

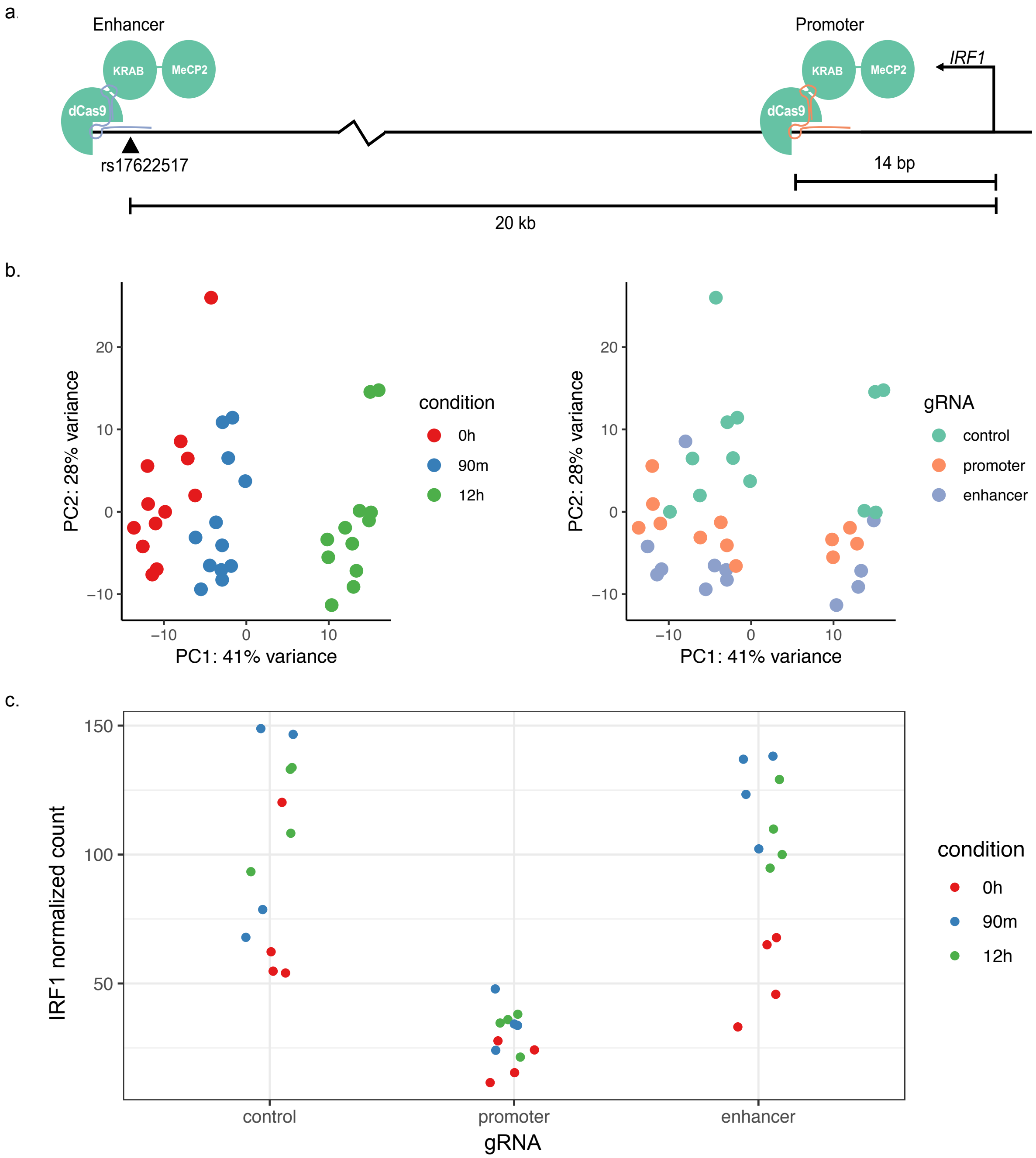


Figure 3-3. RNA-sequencing of hTLR4 cells with CRISPRi targeting IRF1 promoter and rs17622517 supports causality of the variant locus. (a) Schematic of the promoter (orange) and enhancer (purple) gRNAs guiding dCAS9-KRAB-MeCP2 to their respective genomic loci. (b) PCA plots of normalized RNA-sequencing counts coloring samples by LPS condition and CRISPRi gRNA. (c) *IRF1* expression in all samples split by CRISPRi gRNA.

Next, we performed differential expression analysis on the gene counts from the RNA-seq of the 36 samples to detect the transcriptional effect of LPS, silencing of the enhancer and the IRF1 promoter and the interaction between the two. First, we looked at the effect of LPS treatment in control gRNA samples using a model of differential expression between the LPS conditions. The 90m LPS versus 0h effect includes 30 differentially expressed genes with an FDR of 0.05. The top enriched GO terms for this set of genes are inflammatory response (corrected p-value = 5.6×10^{-8}), chemokine-mediated signaling pathway (corrected p-value = 1.7×10^{-5}) and response to LPS (corrected p-value = 7.4×10^{-4}). The 12h versus 0h effect includes 476 differentially expressed genes. The top enriched GO terms for this set of genes are inflammatory response (corrected p-value = 1.9×10^{-5}), cell-cell signaling (corrected p-value = 8.0×10^{-3}), chemokine-mediated signaling pathway (corrected p-value = 8.5×10^{-3}) and NF-κB signaling (corrected p-value = 1.1×10^{-2}). Between these two sets of differentially expressed genes, 22 genes overlap. This result demonstrates that the TLR4 cells are responding to LPS stimulation by activating pro-inflammatory cellular pathways.

The promoter versus control effect, using a model of differential expression between gRNAs in 12h LPS samples, includes 1016 genes. The enriched terms for this gene set fall into a broad spectrum of cellular processes (table 3-1), perhaps due to the large range of functions of IRF1 in the cell. The 19th highest enriched GO term for this set of genes is NF-κB signaling (corrected p-value = 3.2×10^{-4}). The enhancer versus control main effect includes 225 genes. Interestingly, 161 of these enhancer differentially expressed genes are also differentially expressed under the promoter gRNA, suggesting that the enhancer effects on the transcriptome are primarily mediated by IRF1. Among other cellular processes (table 3-2), this set of enhancer differentially expressed genes is enriched for MyD88-independent TLR signaling pathway (p-value = 9.4×10^{-4}), NF-κB signaling (corrected p-value = 7.3x10^{-3}), TRIF-dependent TLR signaling pathway (corrected p-value = 0.015), MyD88-dependent toll-like receptor signaling pathway (corrected p-value = 0.02). The MyD88 and TRIF-dependent pathways are the two parallel pathways triggered by LPS binding and TLR4 activation. These enrichments suggest that repression of the enhancer affects the immune cell's response to LPS. Neither enhancer nor promoter differentially expressed gene lists were significantly enriched for msigdb IRF1 targets (Fisher test p = 1 and 0.245, respectively).

Finally, we used a ~gRNA + condition + gRNA:condition model to capture the interaction effects of gRNA and LPS condition. We see very few significant genes for the interaction effect between gRNA and LPS condition: 2 genes for promoter:condition90m, 15 for enhancer:condition90m, 1 for promoter:condition12h and 17 for enhancer:condition12h. We hypothesize that perhaps we are underpowered to detect subtle interaction effects. Nevertheless, due to the large number of genes whose expression is affected by the enhancer CRISPRi, and high overlap between these genes and the genes affected by the IRF1 promoter CRISPRi, we felt confident that the enhancer is active in these cells and the causality of the lead *cis*- and *trans*-eQTL variant is worth pursuing.

Table 3-1. Promoter versus control enriched GO terms in differentially expressed genes in 12h LPS samples

GO Term	Count	%	p-value	Fold Enrichment	BH adjusted p-value
SRP-dependent cotranslational protein targeting to membrane	68	6.7	1.73E-63	13.2	5.21E-60
viral transcription	68	6.7	1.12E-55	11.1	1.68E-52
translational initiation	73	7.2	2.95E-54	9.7	2.96E-51
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	67	6.6	7.60E-52	10.3	5.72E-49
rRNA processing	79	7.8	6.57E-44	6.7	3.96E-41

translation	81	8.0	9.27E-40	5.8	4.65E-37
mitochondrial electron transport, NADH to ubiquinone	18	1.8	3.72E-10	6.7	1.60E-07
cell-cell adhesion	42	4.2	2.91E-09	2.8	1.10E-06
mitochondrial respiratory chain complex I assembly	19	1.9	4.04E-09	5.5	1.35E-06
ribosomal small subunit assembly	11	1.1	1.34E-08	10.6	4.02E-06
mitochondrial electron transport, cytochrome c to oxygen	11	1.1	2.54E-08	10.0	6.96E-06
cytoplasmic translation	12	1.2	2.70E-08	8.8	6.77E-06
cell division	47	4.6	2.99E-08	2.5	6.92E-06
mitochondrial ATP synthesis coupled proton transport	11	1.1	4.61E-08	9.6	9.93E-06
positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition	18	1.8	5.42E-07	4.3	1.09E-04
mitotic nuclear division	35	3.5	7.08E-07	2.6	1.33E-04
anaphase-promoting complex-dependent catabolic process	18	1.8	9.70E-07	4.2	1.72E-04
negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	17	1.7	1.05E-06	4.4	1.75E-04
NIK/NF-kappaB signaling	16	1.6	2.02E-06	4.4	3.20E-04
proteasome-mediated ubiquitin-dependent protein catabolic process	29	2.9	5.93E-06	2.6	8.92E-04
positive regulation of canonical Wnt signaling pathway	21	2.1	7.51E-06	3.2	1.08E-03
ribosomal small subunit biogenesis	8	0.8	1.06E-05	9.1	1.44E-03
ATP biosynthetic process	10	1.0	1.56E-05	6.3	2.04E-03
protein polyubiquitination	26	2.6	2.39E-05	2.6	3.00E-03
stimulatory C-type lectin receptor signaling pathway	18	1.8	5.22E-05	3.1	6.27E-03
viral process	34	3.4	9.91E-05	2.1	1.14E-02
DNA damage response, detection of DNA damage	10	1.0	1.04E-04	5.1	1.15E-02
Wnt signaling pathway, planar cell polarity pathway	16	1.6	1.31E-04	3.2	1.40E-02

antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	13	1.3	1.34E-04	3.8	1.38E-02
regulation of mRNA stability	17	1.7	1.42E-04	3.0	1.42E-02
error-prone translesion synthesis	7	0.7	3.85E-04	6.7	3.67E-02
hydrogen ion transmembrane transport	12	1.2	4.15E-04	3.6	3.83E-02
covalent chromatin modification	17	1.7	4.22E-04	2.7	3.78E-02
cellular response to DNA damage stimulus	25	2.5	4.44E-04	2.2	3.85E-02

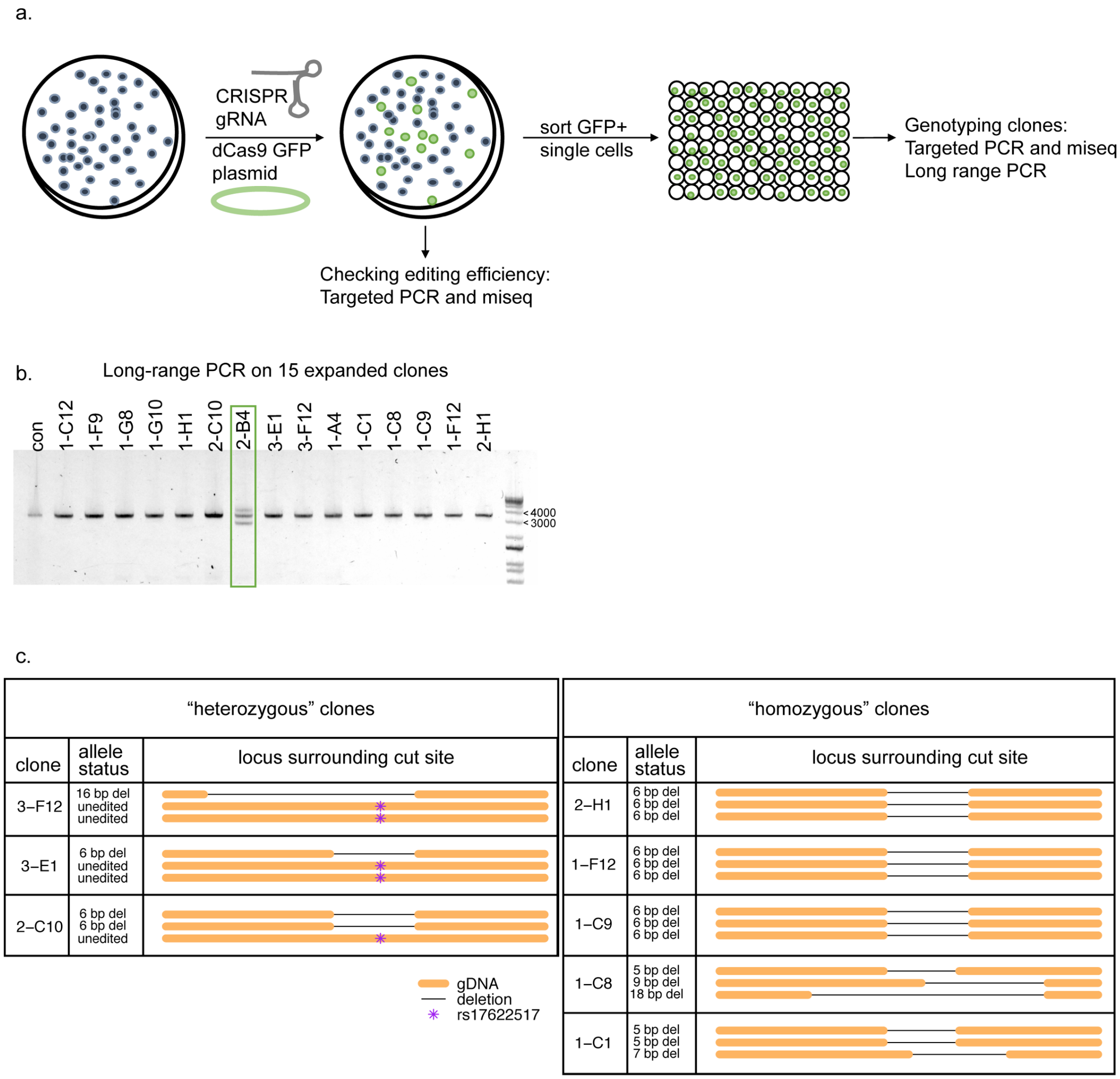
Table 3-2. Enhancer versus control enriched GO terms in differentially expressed genes in 12h LPS samples

GO Term	Count	%	p-value	Fold Enrichment	BH adjusted p-value
SRP-dependent cotranslational protein targeting to membrane	71	32.0	4.58E-119	62.8	4.94E-116
viral transcription	70	31.5	6.50E-108	52.0	3.50E-105
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	70	31.5	3.45E-105	48.9	1.24E-102
translational initiation	70	31.5	3.30E-99	42.5	8.88E-97
rRNA processing	70	31.5	2.37E-82	27.2	5.10E-80
translation	72	32.4	6.52E-80	23.7	1.17E-77
cytoplasmic translation	13	5.9	2.79E-17	43.2	4.30E-15
ribosomal small subunit assembly	9	4.1	2.47E-11	39.4	3.33E-09
cell-cell adhesion	21	9.5	9.39E-11	6.4	1.12E-08
ribosomal small subunit biogenesis	8	3.6	3.31E-10	41.6	3.57E-08
ribosomal large subunit assembly	8	3.6	3.20E-09	31.7	3.14E-07
translational elongation	6	2.7	1.76E-06	27.7	1.59E-04
virion assembly	5	2.3	9.15E-06	34.6	7.58E-04
ribosomal large subunit biogenesis	6	2.7	1.02E-05	20.0	7.87E-04
MyD88-independent toll-like receptor signaling pathway	5	2.3	1.31E-05	32.0	9.40E-04
regulation of mRNA stability	9	4.1	3.30E-05	7.3	2.22E-03
response to ethanol	9	4.1	3.79E-05	7.1	2.40E-03
DNA damage response, detection of DNA damage	6	2.7	6.52E-05	13.9	3.90E-03
error-prone translesion synthesis	5	2.3	6.71E-05	21.9	3.80E-03
NIK/NF-kappaB signaling	7	3.2	1.36E-04	8.8	7.33E-03

negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	7	3.2	2.05E-04	8.2	1.05E-02
positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition	7	3.2	2.97E-04	7.7	1.45E-02
TRIF-dependent toll-like receptor signaling pathway	5	2.3	3.26E-04	14.8	1.52E-02
regulation of necrotic cell death	4	1.8	3.43E-04	27.7	1.53E-02
intracellular transport of virus	6	2.7	3.51E-04	9.8	1.50E-02
anaphase-promoting complex-dependent catabolic process	7	3.2	3.67E-04	7.4	1.51E-02
regulation of tumor necrosis factor-mediated signaling pathway	5	2.3	4.28E-04	13.9	1.69E-02
regulation of type I interferon production	4	1.8	4.42E-04	25.6	1.69E-02
viral life cycle	5	2.3	4.87E-04	13.4	1.79E-02
maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	5	2.3	5.51E-04	13.0	1.96E-02
MyD88-dependent toll-like receptor signaling pathway	5	2.3	6.21E-04	12.6	2.14E-02
I-kappaB kinase/NF-kappaB signaling	6	2.7	7.49E-04	8.3	2.49E-02
regulation of protein ubiquitination	4	1.8	8.43E-04	20.8	2.72E-02
translesion synthesis	5	2.3	8.70E-04	11.5	2.72E-02
negative regulation of epidermal growth factor receptor signaling pathway	5	2.3	8.70E-04	11.5	2.72E-02
DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	6	2.7	8.70E-04	8.0	2.65E-02
nuclear import	4	1.8	1.21E-03	18.5	3.55E-02
G2/M transition of mitotic cell cycle	8	3.6	1.32E-03	4.9	3.76E-02
error-free translesion synthesis	4	1.8	1.42E-03	17.5	3.95E-02
protein polyubiquitination	9	4.1	1.70E-03	4.1	4.60E-02

Isolating monoclonal cell lines with edits at the *IRF1* variant locus

In order to obtain cell lines to further investigate the *IRF1* reQTL, we used CRISPR/Cas9 genome editing to introduce indels at the variant's genomic locus in HEK293-TLR4 cells. The experimental strategy is exhibited in figure 3-4a. Amplifying and sequencing gDNA from the polyclonal edited population revealed an editing efficiency of 31% NHEJ before sorting and 50% NHEJ after sorting GFP+ cells, the difference presumably explained by increased transfection efficiency. Fifteen plates of monoclonal cell lines from the GFP+ sorted cells were screened for editing. Fifteen of those clones were expanded based on promising results from the monoclonal genotyping sequencing. Because the HEK293 cell line is triploid at this locus, clones are considered heterozygous if they have one or two edited alleles. Homozygous clones have all three alleles edited, and wild type clones show no edited alleles. Upon further inspection, clone 1-A4 was eliminated from further study due to the fact that it has 50% of one deletion and 50% of another, suggesting it has lost an allele. Additionally, clone 2-B4 was eliminated due to long-range PCR producing multiple bands, suggesting large indels at the variant locus (figure 3-4b). The 13 clones selected for functional follow-up include five homozygous reference (wild type) clones (not shown), three heterozygous clones, and five homozygous alternative clones (figure 3-4c).



“heterozygous” clones		
clone	allele status	locus surrounding cut site
3–F12	16 bp del unedited unedited	
3–E1	6 bp del unedited unedited	
2–C10	6 bp del 6 bp del unedited	

“homozygous” clones		
clone	allele status	locus surrounding cut site
2–H1	6 bp del 6 bp del 6 bp del	
1–F12	6 bp del 6 bp del 6 bp del	
1–C9	6 bp del 6 bp del 6 bp del	
1–C8	5 bp del 9 bp del 18 bp del	
1–C1	5 bp del 5 bp del 7 bp del	

Figure 3-4. Eight edited clones with small deletions at the rs17622517 variant locus and five wild type clones were obtained using CRISPR/Cas9 and single cell sorting. (a) Approach to edit and isolate single cell clones. A gRNA which cuts 1 bp from the rs17622517 variant is transfected as a gblock along with a CRISPR/Cas9 and GFP+ plasmid into HEK293-TLR4 cells. GFP+ cells are then sorted into single wells of 96 well plates. Genotyping of the monoclonal cell lines is performed by next-generation amplicon sequencing using locus-specific primers and indexed nextera adapters. (b) Gel electrophoresis image of a 3,496 bp PCR amplicon spanning the cut site to check for larger deletions or insertions in 15 expanded monoclonal cell lines. (c) The deletions in the three “heterozygous” and five “homozygous” edited clones chosen for LPS stimulation and RNA-seq, along with five wild type clones, not pictured.

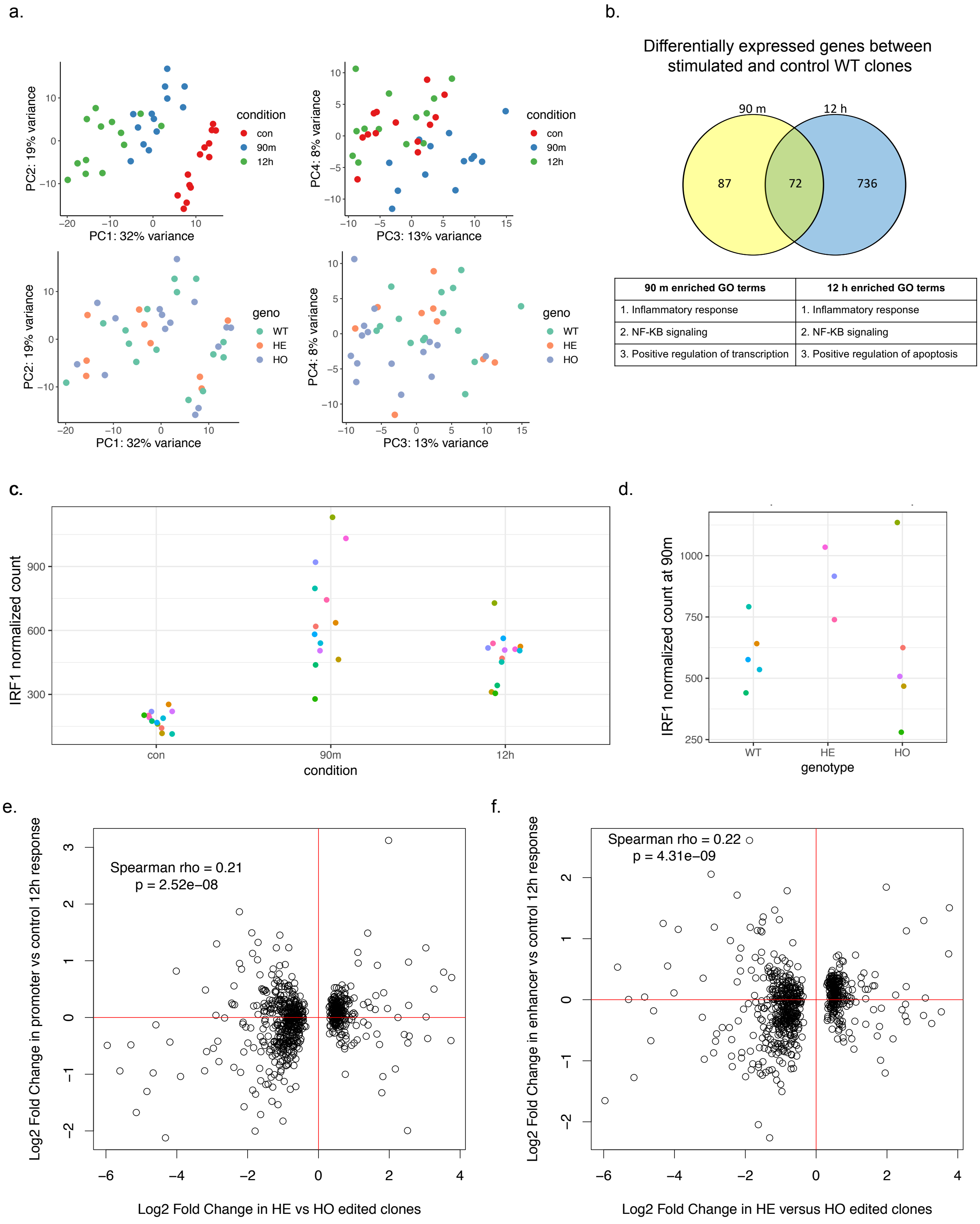
Editing the enhancer locus results in differential expression which correlates with enhancer and promoter CRISPRi perturbation

Each of the 13 clones was exposed to three conditions: untreated, 90 min LPS treatment and 12 hours LPS treatment. The 12 hour timepoint was used instead of the 6 hour timepoint used in the monocyte study because previous experiments in the lab showed that genes involved in LPS response had a longer lag in response time in HEK293-TLR4 cells than in monocytes (data not shown). After treatment, RNA-sequencing was performed on the samples and reads were mapped to the genome and assigned to known genes. The stimulation and sequencing experiment was performed twice and the reads were combined for each sample, resulting in an average of 31.6 million reads per sample. Principle component analysis of normalized gene expression demonstrates that samples separate strongly by treatment effect and slightly by genotype (figure 3-5a). This is as expected, since LPS stimulation should elicit a strong response in these cells and the genotype effect is small and expected to only be active at the 12h timepoint.

Differential expression analysis of the RNA-sequencing samples revealed that many genes are differentially expressed between the 90m or 12h LPS and control samples (figure 3-5b). The DE gene lists for control vs 90m and 12h are enriched for gene ontology (GO) terms involved in inflammatory response and NF-κB signaling, signifying a successful stimulation of the samples. *IRF1* expression is induced upon stimulation at 90m and 12h as compared to control (figure 3-5c), but *IRF1* does not vary significantly between genotypes (Kruskal-Wallis p = 0.17, figure 3-5d). An effect of the edit on expression of genes would manifest as an interaction between genotype and condition, meaning edited clones differ in their transcriptional response to LPS stimulation than wild type clones. However, differential expression (DE) analysis on the interaction between genotype and condition did not result in meaningful significant genes, much like we saw with the

CRISPRi treatment. For the 12-hour timepoint, where we would expect to see the *trans*-eQTL effect, there were 37 DE genes (FDR 5%) between HO and HE clones, 0 genes between HO and WT and 3 DE genes between HE and WT. These DE genes have no overlap with IRF1 targets or *trans*-eGenes from the eQTL study, suggesting that it may be due to noise. This result is not necessarily surprising given the relatively low sample size and modest effects of *trans*-eQTLs. As in the CRISPRi differential expression, we hypothesized that perhaps the interaction effects of gRNA and LPS condition exist, but are subtle and their discovery is limited by power.

In order to determine whether editing the variant locus results in a similar effect as the CRISPRi perturbation of the enhancer or promoter, we looked at the correlation between the differential expression fold changes in the edited cells and CRISPRi 12h differential expression fold changes. The correlation between the promoter interaction fold changes with the monoclonal edited HE vs HO fold changes in DE genes (FDR 0.25) is significant (rho = 0.21, p = 2.52x10^{-8}, figure 3-5e), suggesting that editing the variant has a similar effect on these genes as a knockdown of IRF1 expression. We also see a significant correlation in this same set of genes between the enhancer interaction fold changes and the edited clone fold changes (rho = 0.22, p = 4.31x10^{-9} figure 3-5f), indicating that a similar effect is produced by CRISPRi and genome editing at the enhancer locus.



90 m enriched GO terms	12 h enriched GO terms
1. Inflammatory response	1. Inflammatory response
2. NF-KB signaling	2. NF-KB signaling
3. Positive regulation of transcription	3. Positive regulation of apoptosis

Figure 3-5. RNA-sequencing demonstrates an effect of editing the locus which correlates with the CRISPRi perturbations. (a) PCA plots of normalized RNA-sequencing counts coloring samples by condition and genotype. (b) Shared and unique significantly differentially expressed genes (5% FDR) between 90 min and control and 12 hour and control samples in WT clones along with the three top significantly enriched GO terms for each gene set from DAVID. (c) *IRF1* expression in all clones split by condition and colored by individual clone identity. (d) *IRF1* expression in all clones split by genotype for 90 min samples. (e) Correlation between differential expression fold change between HE and HO edited clones and differential expression fold change between CRISPRi gRNA treatments (promoter vs control) for the same set of genes (HE vs HO DE genes at 0.25 FDR). (f) Correlation between differential expression fold change between HE and HO edited clones and differential expression fold change between CRISPRi gRNA treatments (enhancer vs control) for the same set of genes (HE vs HO DE genes at 0.25 FDR).

IRF1 trans-eQTL effect sizes are correlated with differential expression fold change in edited clones

In order to investigate if and how the effect of the *trans*-eQTL can be captured by genome editing a cell line, we compared the effect size (beta) of association for the *trans*-eQTL eGenes at 6 hours in the monocyte study with the fold changes from the differential expression, testing the interaction between genotype and condition for the same genes. The *trans*-eQTL betas are significantly positively correlated with the fold changes in WT/HE clones (figure 3-6a) and significantly negatively correlated with fold changes in HE/HO clones (figure 3-6b). The correlation between WT/HO fold changes and *trans* betas was not significant. The correlation finding here is strong evidence that the *trans*-eQTL is a true association, which is acting through a condition-specific active enhancer modifying the expression level of the *IRF1* gene.

In order to see if the CRISPRi perturbation has a similar effect a the trans-eQTL associations, we also looked at the correlation of the fold changes in the 12h interaction differential expression with the betas of the IRF1 trans-eGenes from the monocyte study (filtered for FDR 0.25). The correlations between promoter and enhancer interaction fold changes and trans-eGene betas are not significant (figure 3-6c-d), suggesting a stronger similarity between CRISPR editing and the *trans*-eQTL effect than CRISPRi perturbation and the *trans*-eQTL effect.

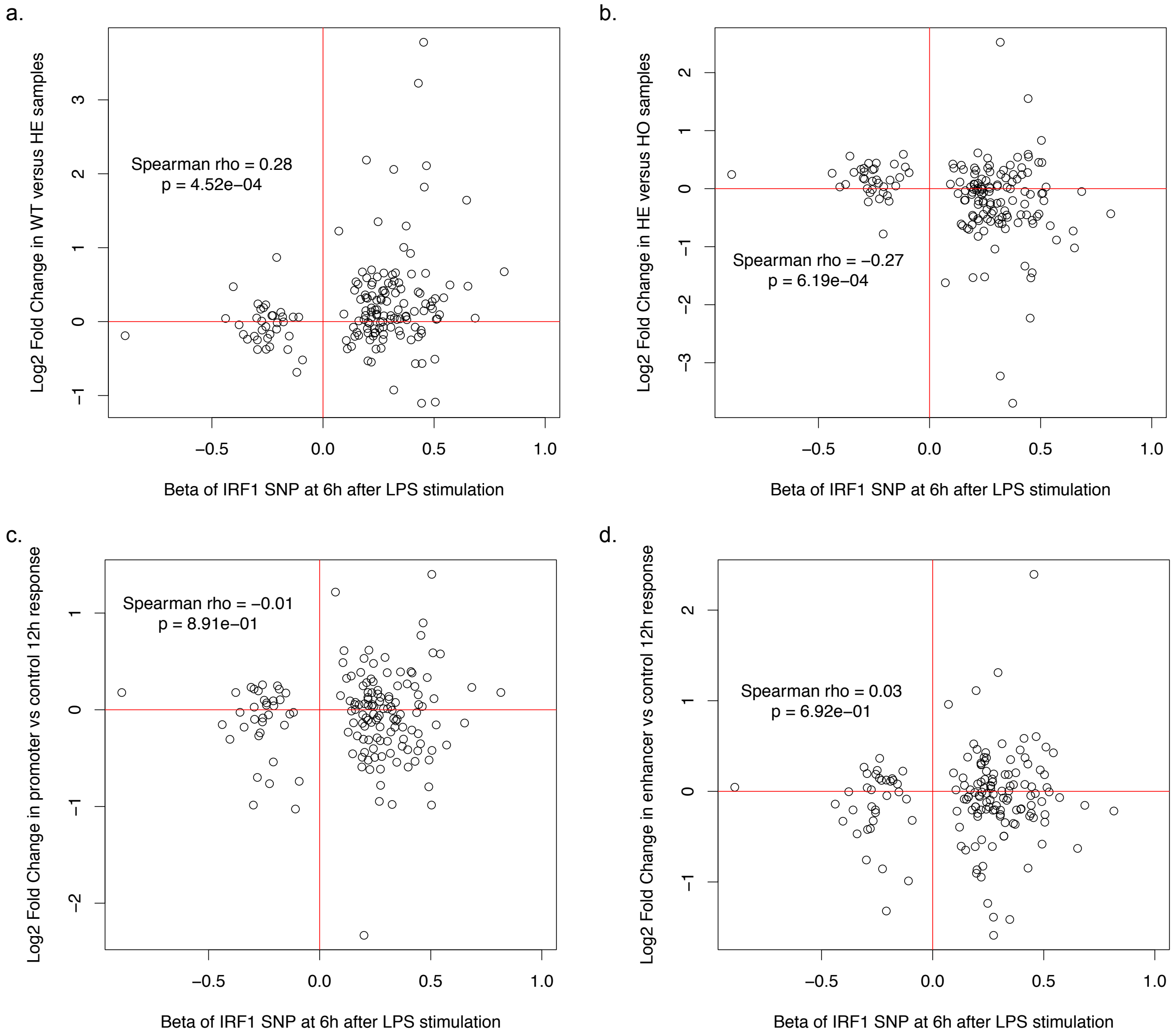


Figure 3-6. RNA-sequencing of the edited clones with and without LPS stimulation demonstrates a mild effect of genotype on the *trans*-eGenes for rs17622517. (a) Correlation between *trans*-reQTL effect sizes from the monocyte study and differential expression fold change between WT and HE clones for the same set of genes (trans-eGenes at 0.5 FDR). (b) Correlation between *trans*-reQTL effect sizes from the monocyte study and differential expression fold change between HE and HO edited clones for the same set of genes (trans-eGenes at 0.5 FDR). (c) Correlation between *trans*-reQTL effect sizes from the monocyte study and differential expression fold change between promoter CRISPRi and control gRNA treatments at

12h for the same set of genes (trans-eGenes at 0.5 FDR). (d) Correlation between *trans*-reQTL effect sizes from the monocyte study and differential expression fold change between enhancer CRISPRi and control gRNA treatments at 12h for the same set of genes (trans-eGenes at 0.5 FDR).

Discussion

In this study, we discovered an immune-response eQTL which is significant not only for one gene *IRF1* in *cis*, but also many other genes in *trans*. *Trans*-eQTLs in general are difficult to detect, in part because of the multiple testing burden of testing each variant against all genes, generally smaller effect sizes, context-specificity and other technical factors (Saha and Battle, 2018; Võsa et al., 2018). We increased the FDR threshold to 0.5 in order to further investigate the *trans* associations, something that is not unusual in *trans*-eQTL studies. The much higher number of *IRF1 trans*-eGenes as compared to the other *cis*-eQTLs as well as the enrichment of *IRF1* targets in those *trans*-eGenes lends strong support for the *trans*-eQTL's validity.

We used a CRISPR-based approach to analyze this trans-eQTL in a cellular model. After establishing that our HEK293-TLR4 cell line has a robust response to LPS, we verified with CRISPRi that silencing of the enhancer affects many genes, demonstrating that the eQTL variant locus is an active enhancer in these cells. Next, we demonstrated that genetic disruption of this eQTL variant locus impacts the expression of *IRF1* targets in a condition-specific manner, and these effects are correlated with the CRISPRi silencing of the enhancer and the effect sizes of the *trans*-eQTLs. These results provide strong experimental support for the *trans*-eQTL.

The CRISPRi experiments to perturb both the IRF1 promoter and the enhancer locus further support the idea that IRF1 plays a role in LPS response. The effect of the enhancer gRNA on many genes demonstrates that it is in fact active in these cells. The differential expression in promoter and enhancer are not significantly correlated with the *trans*-eGenes, which would have further validated the trans-eQTL effect. However, IRF1 is known to affect different genes in many different cellular contexts and it is likely that its effect in this cell line may differ from that in primary monocytes. Nevertheless, the positive correlation between the enhancer CRISPRi differential expression and the monoclonal edited differential expression confirms that by inactivating the enhancer with CRISPRi, we are mimicking the effect of the edited variant. Furthermore, correlation between the promoter CRISPRi differential expression and the monoclonal edited differential expression indicates that the edited variant is acting through affecting transcription of IRF1.

Differential expression analysis between the different genotypes of the monoclonal cell lines under stimulation did not detect the expected effect of editing the variant on expression of the *trans*-eGenes or *IRF1*. One explanation for this result is that we are underpowered to detect subtle changes in gene expression. eQTL studies require hundreds of subjects for this reason: regulatory effects can be small and noise due to smaller sample sizes can obscure true associations. In our case, perhaps the modest number of edited clones was not sufficient to detect the subtle effects of the variant. As we saw with the eliminated clone in figure 3-3b, monoclonal cell lines can have undetected large insertions or deletions which can introduce additional noise into the system. Additionally, we introduced a deletion into the variant locus, rather than introducing the exact variant into the cell line. If the variant somehow increases the activity of the enhancer, it is possible that deleting a portion of the enhancer does not replicate this effect, and that different deletions have different effects. We took this approach instead of introducing the variant with homologous recombination because of the extremely low HDR efficiency when we included a ssDNA template in the transfection. Optimizing the CRISPR and transfection conditions in these cells could potentially increase editing efficiency and allow for isolation of monoclonal cell lines with the precise variant edited.

In both the monoclonal editing and CRISPRi RNA-sequencing experiments, we do not see a perfect correlation of differentially expressed genes with the trans-eQTL effects. One explanation is that the trans-eQTL association was detected in primary patient-derived monocytes, while the CRISPRi silencing and variant editing was performed in a HEK cell line transduced with a hTLR4 receptor. It is not surprising that the cell line does not fully recapitulate the trans-eQTL effect, since it may lack expression of other immune-related genes which may be interacting with IRF1. For example, IRF1 has been found to cooperate with STAT1 (Abou El Hassan et al., 2017) in inducing expression of target genes. Furthermore, it is likely that the trans-eQTL genes are composed not only of direct targets of IRF1, but also downstream targets of IRF1 targets. If these genes are not as inducible in this cell line as in monocytes, these downsteam effects will be dampened. Another drawback to the cell line could be that it lacks expression of a protein which introduces a post-translational modification on IRF1, which affects its transcriptional activity. While research is unclear on the importance of post-translational modifications on IRF1, recent evidence suggests ubiquitination and phosphorylation may affect its activity (Garvin et al., 2019). Perhaps introducing variants into a monocyte cell line, such as THP1 cells, would benefit from a more robust immune response and therefore a larger effect of the enhancer perturbation. However, there are likely to be limitations to using any immortalized cell line. A future avenue of study could be introducing the variant into a mouse line and studying the LPS response in edited and wild type mouse-derived monocytes, as well as cytokine production in the mice in response to LPS.

Our model for the regulatory action of the variant is that it affects the activity of the enhancer in which it resides, likely by changing the affinity for binding of a TF that may be active only under LPS stimulus. This genetic effect on activity of the enhancer in turn affects expression of *IRF1* under stimulation with LPS. The results of this study suggest that the *IRF1* TF plays a role in regulating the transcriptional response of immune cells to LPS stimulation, and that a common genetic variant leads to inter-individual variation in this response. The alternative allele of this variant has additionally been found to be associated with ankylosing spondylitis, Crohn's disease and ulcerative colitis in a GWAS for chronic inflammatory diseases (Ellinghaus et al., 2016). Further support for IRF1 being plausibly involved in these diseases comes from the connection between known ulcerative colitis-related genes and IRF1. For example, ulcerative colitis has also been linked to *IRF1* targets genes *CXCL10* (Shi et al., 2019), IL6 (Wu et al., 2014) and *iNOS* (Wu et al., 2014) and *IRF1*-cooperating transcription factor *STAT1* (Ciorba et al., 2010). Further experiments to explore the link between *IRF1* and these diseases could include observing whether a knockdown of *IRF1* or the enhancer locus in an animal model of ulcerative colitis improves the disease symptoms.

In the *cis*-reQTL study, *IRF1* does not have a significant *cis*-eQTL at baseline. It is only under stimulation with LPS that the regulatory effect of the variant is revealed. Therefore, the *IRF1* locus exemplifies the utility of introducing a cellular perturbation in order to detect variants involved not necessarily in baseline expression of genes, but dynamic expression of genes in different physiological contexts. Additionally, the top variant for this eQTL is found within an intron of another gene, *C5ORF56*. Therefore, when the variant was found as a hit in the inflammatory disease GWAS, it was automatically assigned to this gene, based on proximity. The results of the eQTL association for this variant suggest that the gene of interest might actually be *IRF1*, located 20 Kb away. The variant might be contributing to disease by modifying the expression of IRF1 and its targets and therefore modifying the immune-response in the individual.

This finding further exemplifies the usefulness of integrating eQTL data with GWAS in order to match variant-gene combinations. Although more follow up on the variant and enhancer is needed, this study has identified a genetic variant which modifies individuals’ response to immune stimulation by affecting an enhancer for IRF1 and the downstream regulatory pathways of IRF1 in an immune stimulus -specific manner, suggesting that this is the likely mechanism for the genetic risk for autoimmune disease driven by this locus.

Table 3-3. *Trans*-reQTL associations in monocytes for rs17622517 6 hours after stimulation with LPS (0.5 FDR).

Illumina probe	gene symbol	*trans*-eQTL beta	*trans*-eQTL p-value	*trans*-eQTL FDR
ILMN_1811378	GTPBP1	0.2	3.68E-08	0.001
ILMN_1670000	DCAF6	-0.27	6.67E-08	0.002
ILMN_1756862	APOL3	0.84	2.61E-07	0.008
ILMN_2404512	PSEN2	0.24	2.80E-07	0.008
ILMN_1701613	RARRES3	0.35	3.37E-07	0.01
ILMN_1700671	ETV7	0.28	4.24E-07	0.012
ILMN_1809467	VAMP5	0.47	8.16E-07	0.021
ILMN_1690241	BATF2	0.37	1.03E-06	0.026
ILMN_3307659	SFT2D2	0.38	1.61E-06	0.04
ILMN_2373831	BTN3A3	0.25	1.79E-06	0.043
ILMN_1670305	SERPING1	0.5	2.03E-06	0.047
ILMN_1778599	SP140	0.25	2.09E-06	0.048
ILMN_1768433	CCDC71	0.2	2.41E-06	0.055
ILMN_1669617	GRB10	0.2	3.04E-06	0.068
ILMN_1718558	PARP12	0.41	3.00E-06	0.068
ILMN_2284998	SP100	0.36	3.23E-06	0.072
ILMN_1808148	SMCHD1	0.15	3.73E-06	0.081
ILMN_2233783	CD38	0.82	4.08E-06	0.086
ILMN_1764380	GLTP	-0.34	4.78E-06	0.096
ILMN_1783843	MIIP	0.24	4.82E-06	0.096
ILMN_2395236	CHEK2	0.19	5.27E-06	0.103
ILMN_1668378	SFT2D2	0.18	5.38E-06	0.105
ILMN_2373177	PANK2	0.24	5.44E-06	0.105
ILMN_1771921	HSCB	0.23	6.39E-06	0.117
ILMN_3237462	IDO2	0.23	6.87E-06	0.12
ILMN_1805201	PML	0.18	6.86E-06	0.12
ILMN_1690921	STAT2	0.46	7.11E-06	0.121
ILMN_1753758	IL27	0.86	7.41E-06	0.124
ILMN_1703263	SP140	0.34	7.88E-06	0.129
ILMN_1749372	GGT5	0.24	8.00E-06	0.13
ILMN_2296950	APOBEC3F	0.22	8.25E-06	0.132
ILMN_1776602	ANG	0.2	8.55E-06	0.135
ILMN_1710726	APOBEC3F	0.24	8.76E-06	0.137
ILMN_1662964	PRMT3	-0.27	8.85E-06	0.138
ILMN_3204136	LOC100132707	0.1	9.04E-06	0.14
ILMN_1687533	SEMA4D	0.51	9.26E-06	0.142
ILMN_1757845	SPIRE1	-0.26	9.42E-06	0.143
ILMN_3238326	RNF144A	0.15	9.45E-06	0.143
ILMN_1780831	SLC6A12	0.32	9.63E-06	0.144
ILMN_1731299	PML	0.11	1.01E-05	0.151
ILMN_1782487	GBP1	0.69	1.06E-05	0.153
ILMN_1691393	DNPEP	0.19	1.06E-05	0.153
ILMN_2148785	GBP1	0.65	1.05E-05	0.153
ILMN_1727271	WARS	0.51	1.07E-05	0.154
ILMN_1683178	JAK2	0.44	1.13E-05	0.161

ILMN_1797191	KIAA0040	0.2	1.20E-05	0.166
ILMN_2183510	MANF	0.29	1.23E-05	0.167
ILMN_2337655	WARS	0.43	1.24E-05	0.167
ILMN_1728224	OGFR	0.35	1.26E-05	0.168
ILMN_1788017	HSH2D	0.27	1.28E-05	0.171
ILMN_1742929	HESX1	0.42	1.30E-05	0.173
ILMN_1662026	BTK	0.26	1.34E-05	0.176
ILMN_1659960	NUP62	0.45	1.34E-05	0.176
ILMN_1692168	UBE2Z	0.27	1.37E-05	0.178
ILMN_1726769	CNDP2	0.34	1.44E-05	0.184
ILMN_1678766	DYNLT1	0.45	1.44E-05	0.184
ILMN_1665865	IGFBP4	0.36	1.47E-05	0.187
ILMN_1701114	GBP1	0.7	1.52E-05	0.19
ILMN_1790472	SLC25A28	0.48	1.53E-05	0.191
ILMN_1704477	COX5A	-0.29	1.60E-05	0.198
ILMN_1786612	PSME2	0.35	1.71E-05	0.207
ILMN_2106725	NCF1B	0.26	1.72E-05	0.207
ILMN_1751079	TAP1	0.38	1.72E-05	0.208
ILMN_1721411	PARP10	0.46	1.77E-05	0.212
ILMN_1731044	NDUFC2	0.26	1.89E-05	0.217
ILMN_2325338	APOL2	0.27	1.90E-05	0.217
ILMN_1665682	IL15RA	0.49	1.87E-05	0.217
ILMN_1784320	ELMO1	0.32	1.89E-05	0.217
ILMN_1813455	SP110	0.43	2.02E-05	0.219
ILMN_1670572	IDO2	0.33	2.03E-05	0.219
ILMN_1769143	KCNE2	0.12	2.01E-05	0.219
ILMN_2355168	MGST1	-0.46	1.97E-05	0.219
ILMN_1740572	TCN2	0.2	1.97E-05	0.219
ILMN_1806017	PSME1	0.23	2.15E-05	0.229
ILMN_1697409	TNFRSF14	0.27	2.18E-05	0.231
ILMN_2349061	IRF7	0.52	2.26E-05	0.234
ILMN_1771385	GBP4	0.77	2.29E-05	0.237
ILMN_1725700	MOV10	0.44	2.32E-05	0.238
ILMN_1713285	NAPA	0.24	2.34E-05	0.239
ILMN_1738704	TRIM26	0.25	2.58E-05	0.254
ILMN_2289093	RNF213	0.49	2.54E-05	0.254
ILMN_2115752	MEFV	0.41	2.56E-05	0.254
ILMN_2379718	RAB24	0.35	2.73E-05	0.259
ILMN_2092333	GPR141	0.39	2.85E-05	0.264
ILMN_2058782	IFI27	0.81	2.93E-05	0.268
ILMN_1739274	PDHB	-0.25	3.10E-05	0.275
ILMN_1713561	C20orf103	0.54	3.11E-05	0.275
ILMN_1753745	HDDC2	-0.3	3.19E-05	0.276
ILMN_1696654	IFIT5	0.37	3.24E-05	0.277
ILMN_1769520	UBE2L6	0.57	3.24E-05	0.277
ILMN_1658759	PEX19	-0.26	3.24E-05	0.277
ILMN_1705241	TDRD7	0.44	3.35E-05	0.28
ILMN_1811823	MED25	0.19	3.35E-05	0.28
ILMN_3237165	LOC100128164	0.13	3.34E-05	0.28

ILMN_2344079	ZGPAT	0.14	3.49E-05	0.288
ILMN_1772824	WNT5B	0.39	3.57E-05	0.29
ILMN_1733176	LIMS1	-0.44	3.57E-05	0.29
ILMN_3246953	FTSJD2	0.31	3.64E-05	0.294
ILMN_2104696	ERICH1	0.27	3.63E-05	0.294
ILMN_2198376	PSMA4	0.22	3.70E-05	0.296
ILMN_1729115	UBE2S	0.22	3.89E-05	0.304
ILMN_1776723	PHF11	0.34	3.91E-05	0.304
ILMN_1795704	KIAA0232	-0.09	3.95E-05	0.306
ILMN_1792305	ZNF318	-0.38	3.97E-05	0.307
ILMN_1687495	SLC37A1	0.16	4.04E-05	0.31
ILMN_1812926	ANTXR2	0.51	4.19E-05	0.314
ILMN_1678422	DHX58	0.65	4.17E-05	0.314
ILMN_1673649	HYOU1	0.34	4.19E-05	0.314
ILMN_1731001	ERICH1	0.25	4.26E-05	0.318
ILMN_1780756	RBM23	0.16	4.30E-05	0.321
ILMN_1651346	TICAM2	0.32	4.41E-05	0.326
ILMN_1802151	OSBPL5	0.22	4.45E-05	0.326
ILMN_1809086	XRN1	0.43	4.44E-05	0.326
ILMN_2393544	PRMT2	0.27	4.53E-05	0.329
ILMN_1652525	FAM125B	0.4	4.82E-05	0.34
ILMN_1803652	C9orf91	0.3	4.84E-05	0.34
ILMN_1710740	C2	0.13	4.86E-05	0.34
ILMN_1765547	IRF2	0.27	4.88E-05	0.341
ILMN_1903568	CR625988	-0.32	4.93E-05	0.341
ILMN_1789095	BMPR2	0.31	4.96E-05	0.341
ILMN_1718303	PVRL2	0.33	4.98E-05	0.341
ILMN_2311826	USP6NL	0.23	5.19E-05	0.346
ILMN_1684789	CCDC101	0.17	5.20E-05	0.346
ILMN_1753547	STAT5A	0.49	5.13E-05	0.346
ILMN_1811171	GPR132	0.44	5.20E-05	0.346
ILMN_2376108	PSMB9	0.55	5.13E-05	0.346
ILMN_1795991	C22orf28	0.25	5.11E-05	0.346
ILMN_2388466	TIA1	0.21	5.12E-05	0.346
ILMN_3268914	CLEC2D	-0.25	5.39E-05	0.354
ILMN_2112988	NCF1C	0.58	5.45E-05	0.355
ILMN_2365465	XBP1	0.26	5.43E-05	0.355
ILMN_2167416	MR1	0.24	5.62E-05	0.361
ILMN_2339006	KIAA0564	-0.3	5.66E-05	0.361
ILMN_1802106	APOBEC3G	0.43	5.68E-05	0.361
ILMN_2415157	ARID5A	0.22	5.81E-05	0.364
ILMN_2103362	ARHGAP27	0.18	5.97E-05	0.367
ILMN_1701455	FBXO6	0.37	6.00E-05	0.368
ILMN_1745356	CXCL9	0.55	6.11E-05	0.371
ILMN_1750400	C19orf66	0.42	6.17E-05	0.371
ILMN_2344373	MVP	0.29	6.15E-05	0.371
ILMN_1719392	FH	-0.24	6.31E-05	0.376
ILMN_1723414	HACL1	-0.21	6.35E-05	0.376
ILMN_1789955	PNRC1	0.27	6.38E-05	0.376

ILMN_1654488	KDM6A	0.28	6.43E-05	0.377
ILMN_1761820	EDARADD	0.13	6.46E-05	0.377
ILMN_1695917	C5orf15	0.27	6.54E-05	0.379
ILMN_1665428	GSDMD	0.44	6.78E-05	0.386
ILMN_2166524	CCNYL1	0.26	6.76E-05	0.386
ILMN_1729973	ZC3HAV1	0.68	6.87E-05	0.389
ILMN_1745397	OAS3	0.82	6.86E-05	0.389
ILMN_1683678	SPATS2L	0.52	6.96E-05	0.391
ILMN_1777660	RNF144A	0.16	7.11E-05	0.392
ILMN_2262044	PARP10	0.5	7.02E-05	0.392
ILMN_1860051	C20656	0.33	7.30E-05	0.398
ILMN_1787680	SELS	0.18	7.30E-05	0.398
ILMN_3199438	X69637	0.23	7.30E-05	0.398
ILMN_1801307	TNFSF10	0.82	7.29E-05	0.398
ILMN_1706326	MRPL33	-0.16	7.45E-05	0.402
ILMN_1774287	CFB	0.99	7.57E-05	0.404
ILMN_1801766	CCDC109B	0.4	7.56E-05	0.404
ILMN_1807114	UNC93B1	0.27	7.63E-05	0.404
ILMN_1694070	FAM114A1	0.11	7.76E-05	0.405
ILMN_1760727	ANG	0.18	7.79E-05	0.405
ILMN_1751330	RBCK1	0.33	7.92E-05	0.405
ILMN_1695432	TPST2	-0.28	7.92E-05	0.405
ILMN_2300186	DYNLL1	-0.21	7.94E-05	0.406
ILMN_1691567	GNPDA2	-0.13	8.13E-05	0.407
ILMN_1781374	TUFT1	0.26	8.14E-05	0.407
ILMN_1716272	KBTBD8	-0.36	8.23E-05	0.409
ILMN_1683026	PSMB10	0.23	8.31E-05	0.41
ILMN_1710844	PARP10	0.44	8.68E-05	0.414
ILMN_1728349	TMEM63B	0.19	8.50E-05	0.414
ILMN_1674063	OAS2	0.57	8.74E-05	0.414
ILMN_1794470	ANKFY1	0.24	8.73E-05	0.414
ILMN_3243928	DDX60L	0.45	8.42E-05	0.414
ILMN_1731181	TEX2	-0.23	8.80E-05	0.415
ILMN_1728073	DENND1A	0.16	8.91E-05	0.416
ILMN_1897741	CR610863	-0.28	8.97E-05	0.418
ILMN_1769129	CCL19	0.75	1.00E-04	0.44
ILMN_1688526	ARL5A	-0.23	1.01E-04	0.44
ILMN_3219806	UNC93B1	0.42	1.01E-04	0.44
ILMN_1797001	DDX58	0.44	1.03E-04	0.441
ILMN_1750401	C17orf62	0.27	1.03E-04	0.441
ILMN_2362581	FNDC3A	0.33	1.03E-04	0.441
ILMN_1759250	TAP2	0.43	1.02E-04	0.441
ILMN_1808661	TOMM5	-0.24	1.04E-04	0.441
ILMN_1678054	TRIM21	0.31	1.05E-04	0.442
ILMN_2360784	RRBP1	0.24	1.07E-04	0.442
ILMN_1767934	PCSK5	-0.28	1.07E-04	0.442
ILMN_1653711	FZD2	0.32	1.08E-04	0.442
ILMN_1807044	UBAC1	-0.19	1.07E-04	0.442
ILMN_1678140	HEATR8-TTC4	-0.15	1.08E-04	0.442

ILMN_1779324	GZMA	0.14	1.05E-04	0.442
ILMN_2046896	ESRRA	-0.21	1.06E-04	0.442
ILMN_1795227	DYNLL1	-0.19	1.09E-04	0.446
ILMN_1682098	PSMA4	0.28	1.10E-04	0.446
ILMN_2343010	BOLA3	-0.26	1.10E-04	0.446
ILMN_1662795	CA2	-0.89	1.10E-04	0.447
ILMN_3239785	CHEK2	0.25	1.10E-04	0.447
ILMN_1801938	NHS	0.07	1.11E-04	0.447
ILMN_2329679	TPST2	-0.3	1.12E-04	0.45
ILMN_2235851	NEURL3	0.48	1.12E-04	0.45
ILMN_1720083	EHD4	0.21	1.13E-04	0.45
ILMN_2404665	TRIM5	0.23	1.14E-04	0.452
ILMN_1776777	ADAR	0.31	1.16E-04	0.453
ILMN_1676555	TTC26	0.2	1.17E-04	0.456
ILMN_1695058	SLC38A5	0.13	1.19E-04	0.457
ILMN_2359627	BCL2L11	0.22	1.19E-04	0.457
ILMN_2373763	CASP7	0.41	1.20E-04	0.457
ILMN_1793012	C7orf44	0.14	1.20E-04	0.458
ILMN_2406313	RBCK1	0.28	1.21E-04	0.46
ILMN_2045729	WDR12	-0.31	1.22E-04	0.461
ILMN_1718734	MLLT6	0.24	1.23E-04	0.463
ILMN_1765851	TRADD	0.28	1.24E-04	0.466
ILMN_2129877	PARP11	0.22	1.25E-04	0.468
ILMN_1750079	PURB	-0.29	1.26E-04	0.47
ILMN_1760490	ACVR1	-0.4	1.27E-04	0.47
ILMN_1745374	IFI35	0.54	1.27E-04	0.47
ILMN_2312275	SRP54	0.14	1.28E-04	0.47
ILMN_1739840	LRRC8A	-0.18	1.31E-04	0.472
ILMN_3234831	RPL17	-0.12	1.31E-04	0.472
ILMN_1688621	C9orf80	-0.18	1.33E-04	0.473
ILMN_1678862	FUT11	0.09	1.34E-04	0.476
ILMN_1727315	DENND1A	0.22	1.35E-04	0.477
ILMN_1750051	FAM123B	0.19	1.36E-04	0.478
ILMN_1763207	BATF3	0.5	1.36E-04	0.478
ILMN_1724181	IL15	0.46	1.37E-04	0.479
ILMN_1804738	MEFV	0.28	1.38E-04	0.481
ILMN_1664646	NSUN6	0.1	1.38E-04	0.481
ILMN_2400935	TAGLN	0.07	1.43E-04	0.489
ILMN_2065783	EXOC2	0.22	1.44E-04	0.491
ILMN_1670532	GMCL1	-0.11	1.48E-04	0.496

Chapter 4: Conclusions and future perspectives

In this body of work, we set out to experimentally investigate common regulatory variants discovered in population-based eQTL studies and rare regulatory variants. The motivation for this goal stems from the fact that that the vast majority of GWAS variants are regulatory (Maurano et al., 2012). eQTL signals which overlap GWAS signals can help in the interpretation of these variants by pointing to a mechanism and gene of action (Battle et al., 2014; Grundberg et al., 2012; GTEx Consortium et al., 2017; Lappalainen et al., 2013; Nicolae et al., 2010). If these signals are to be understood, identifying the causal variants at eQTLs and the consequences of these regulatory variants through experimental validation is essential.

In chapter 2, we established a method to validate regulatory variants that are found within the transcript. While previous studies have tested rare variants' regulatory effect (Li et al., 2017), and saturation mutagenesis of all variants in particular exons (Findlay et al., 2014; 2018), a scalable method for testing eQTL variants within the transcript has not previously been developed. Additionally, our study is unique in that we establish a control variant distribution with which variants genome-wide can be compared. We showed that matched control variants which introduce a non-stop codon at the same locus as the variant of interest are not ideal controls. Some of these variants have unpredictable effects on gene expression levels and thus result in an artificially large control distribution. Instead, we find that a distribution of synonymous variants which are not significantly associated with eQTLs is a more appropriate control distribution. By using this control distribution, we take into account the background noise involved in the assay and are therefore better able to confidently identify causal regulatory variants.

A reliable assay which not only identifies causal regulatory variants, but also captures the magnitude and direction of effect within the native genomic context has also not existed until now. While MPRAs are scalable assays to identify variants which affect expression, they have been found to have low directional concordance with population data (Tewhey et al., 2016; van Arensbergen et al., 2019), likely due to the removal of the variants from their genomic context by introducing them into a reporter vector. It is notable that in this study we did maintain the genomic context, allowing us to detect the specific and directional effects of regulatory variants in the genome. Additionally, MPRAs are not able to interrogate variants which act by affecting the stability of transcripts.

We demonstrated that the polyclonal assay works particularly well for variants which introduce premature stop codons into the transcript. Of the fourteen total stop-gained variants (both common GTEx stop-gained and disease gene stop-gained) which were expected to trigger NMD, eleven had significant effects on transcript abundance after editing, all in the expected direction. Additionally, none of the four disease gene stop-gained variants which were not expected to trigger NMD were significant, demonstrating sensitivity and specificity of the assay. The reason for the strong success with the stop-gained variants may be because there is a clear cell-type-independent mechanism through which the variants act. NMD is a universal pathway present in all cell types, whereas other mechanisms through which eQTL variants regulate might be cell-type specific. Additionally, NMD effects of rare variants tend to be larger than the relatively small effects of eQTL variants.

When we applied the polyclonal assay to eQTL variants, we found five of the thirteen variants to be significant. Importantly, all five of these variants' effects were in the same direction as we saw in GTEx, demonstrating the assay's ability to capture directional effects of regulatory variants. There are a few possible reasons for why the remaining eQTL variants were not detected to a significant level. Perhaps the most likely reason for this shortfall is that the fine mapping fell short of identifying the true causal variants at the loci. With a scarcity of validated experimental data in the field (a fact that motivated this study), it is hard to assess fine-mapping approaches' ability to distinguish causal variants from those in linkage disequilibrium. When we looked at the p-value distributions of the eQTL signals, we observed that many of the variants are in high linkage disequilibrium with neighboring variants, which also have highly significant p-values. These neighboring significant variants could conceivably be causal and were mis-identified by fine-mapping. Thus far, fine-mapping approaches have only been tested for efficacy on simulated data (Brown et al., 2017; Hormozdiari et al., 2014; Wen et al., 2016), meaning we have little insight into how effective they are at identifying causal variants at real eQTLs. In the future, the polyclonal assay could be utilized to distinguish between significant variants at a locus and thus test the efficacy of fine-mapping approaches.

However, the non-significant eQTL variants whose effect hovers around zero seem to be truly not causal in this cell line. By investigating the range of effects both across tissues and across individuals, we showed that some of these variants might be context-specific and not active in this cellular context. Additionally, the possibility of eQTLs reflecting combined effects of multiple variants in a haplotype could explain why a single variant does not replicate the effect seen in the population.

Another likely explanation for why we don't detect every eQTL as significant is that the assay is not sensitive enough to detect the very small effects of weak eQTLs. This phenomenon can be observed in the three eQTL variants, chr20_3213332_G_A_b38, chr17_81294933_T_C_b38 and chr9_133361131_C_T_b38, which are not called as significant but have similar effects to those seen in GTEx. Technical improvements to homologous recombination efficiency, which is a major area of research (Aird et al., 2018; Gutschner et al., 2016; Song et al., 2016), could reduce noise in the assay and make it able to detect these smaller effects in the future.

The success of the polyclonal assay when applied to stop-gained variants in disease-associated genes demonstrates the applicability of the assay to confirming the causality of potential disease-causing variants. The relationship between disease manifestation and position of the premature stop codon in the transcript in *GLI3* (Furniss et al., 2007) and *ROR2* (Ben-Shachar et al., 2009) promotes the idea that understanding whether a variant results in NMD or truncated protein can provide diagnostic insight into the disease. While the 55 bp NMD cutoff provides a guideline for whether a variant triggers NMD or not, this prediction is far from perfect (Rivas et al., 2015). Therefore, the polyclonal assay could feasibly be applied to improving diagnostic or prognostic information for patients with rare variants of unknown consequence, such as variants falling right at the NMD border.

In some cases, there have been conflicting reports on the mechanism of some disease-causing stop-gained variants, and this assay could assist in resolving those cases. We showed the capacity to do such by demonstrating that the ClinVar variant *ROR2* Arg442Ter triggers NMD. This variant was found in a patient with Robinow Syndrome and shown experimentally to produce truncated protein (Schwarzer et al., 2009). However, the study used an overexpression vector to express the two alleles of the variant in cell culture, and thus was perhaps overwhelming the NMD machinery and thus producing truncated protein. It appears that in the native genomic context, as in our assay, the variant triggers NMD and results in an allele-specific decrease in gene expression level. Our result is more consistent with the clinical manifestation of the disease: both parents of the patient are unaffected, suggesting the disease is not caused by a dominant negative effect of truncated protein. Similarly, we demonstrated in the assay that ClinVar variant GLI3 Arg792Ter, identified in patients with Grieg cephalopolysyndactyly, triggers NMD. An early study hypothesized that the variant acted through production of truncated protein (Kalff-Suske et al., 1999). Another study suggested the variant may be triggering NMD, but their data from allele specific expression in a patient line did not support it (Johnston et al., 2005). Finally, a study demonstrated allele-specific depletion of the mutation in a patient-derived fibroblast line (Furniss et al., 2007), a result which agrees with our findings. The polyclonal assay avoids the need to isolate patient-derived cell lines and is thus more scalable and appropriate for testing many variants with appropriate controls.

Furthermore, we see an application of this assay to narrowing down causal variants in studies where one may have dozens of possible causal variants, such as in whole exome sequencing studies. In such a case, the monoclonal cell line isolation approach is too labor-intensive for this number of variants, and has the potential to be confounded by large on- and off-target mutations, but a higher throughput method like saturation mutagenesis, currently applicable only to individual genes, is laborious and lacks the ability to assay variants across the genome. This polyclonal assay is the ideal throughput for identifying causal variants from a list of a few dozen candidate variants discovered from a genetic study. Although we did not take this approach in this study, it would be feasible to perform the polyclonal assay on a number of potential regulatory variants, sequencing mRNA and gDNA from the polyclonal culture, and then take the same polyclonal culture and sort monoclonal cell lines for only the variants which demonstrate allele-specific regulatory activity. In this approach, the polyclonal assay narrows down the pool of variants to a reasonable number for in-depth follow up with functional assays, protein quantification, and other assays. The straight-forward nature of the polyclonal assay makes it easily adoptable in any lab with tissue culture facilities and access to a sequencing instrument.

The assay need not be limited to SNPs associated with gene expression levels. It can easily be applied to assessing the effect of indels, by introducing the indel into the homologous template instead of a SNP and comparing the presence of the indel in the cDNA versus gDNA. Additionally, we see a clear application of the assay to testing whether or not transcript variants affect splicing. If applied to splicing, the only modifications to the assay would be to look for splicing changes in the amplified region between the reference and alternative alleles.

In Chapter 3, we sought to experimentally validate a variant which is associated with *IRF1* in *cis* and many genes in *trans* under immune stimulation. *Trans*-eQTLs have been more difficult to discover than *cis*-eQTLs due to multiple testing, false positives and negatives attributed to technical confounders and generally smaller effect sizes (Saha and Battle, 2018; Võsa et al., 2018). However, there is evidence that *trans*-eQTLs, which are enriched in enhancers and tend to be more tissue-specific (GTEx Consortium et al., 2017), and are even more informative for understanding the genetic basis of complex disease (Aguet et al., 2019; Westra et al., 2013). Unlike *cis*-eQTLs, *trans*-eQTLs have the potential to influence broad cellular networks by regulating the expression levels of many genes (Brynedal et al., 2017), and thus can illuminate pathway-level effects of genetic variants, which is important for understanding potential disease mechanisms beyond the local effects in *cis*. While there are a few examples of experimental validation of *cis*-eQTLs (Gupta et al., 2017; Soldner et al., 2016; Zhu et al., 2016), there has been almost no validation of *trans*-eQTLs. Therefore, detecting and validating *trans*-eQTLs is of the utmost importance.

Variants within close proximity to *IRF1* have been found through GWAS to be associated with immune-related diseases such as Crohn’s Disease (Franke et al., 2010), eczema (Kichaev et al., 2019) and asthma (Demenais et al., 2018; Ferreira et al., 2017). Additionally, variants in IRF1 and signaling disruption of IRF1 have been implicated in multiple sclerosis (Fortunato et al., 2008; Ren et al., 2011), further linking IRF1 to autoimmune disease. Our work provided strong evidence of the functional mechanisms of this disease-associated locus. The further mechanism of how the IRF1 pathway affects autoimmune disease risk is yet unknown, but since IRF1 induces expression of pro-inflammatory cytokines which can induce an inflammatory response in a broad range of cell types (Kröger et al., 2002), this is the likely mechanism. Future studies to analyze the variant's effect on cytokine production in an immune cell line or mouse model could further explore the physiological and cellular effects of the variant.

IRF1 is known to be induced by TLR signaling, and in turn induces expression of type-I interferons (Miyamoto et al., 1988). However, its role in TLR4 signaling in response to LPS is not well characterized beyond the finding that IRF1 is induced in monocytes in response to LPS (Kim-Hellmuth et al., 2017) and that *irf1* knockout mice have increased survival after LPS infection as compared to controls (Pan et al., 2013). Our finding that an enhancer variant alters expression of IRF1 and a multitude of downstream genes in response to LPS further bolsters the idea that IRF1 plays a role in LPS response in the innate immune system.

The correlation between our edited cell lines and the *trans*-eQTL associations under stimulation helps to validate the *trans*-eQTL association. The eQTL variant can be classified as a master regulatory variant, due to its impact on a number of genes involved in immune response. Individuals with the variant exhibit increased expression levels of *IRF1* and its targets, i.e. an increased immune response to stimulus. We discovered the variant in the context of LPS stimulation, but the variant could possibly modify *IRF1* and its targets in any case where *IRF1* is upregulated via activation of this enhancer. The association of the variant in a GWAS for inflammatory traits (Ellinghaus et al., 2016) is consistent with this idea, since inflammatory disorders are characterized by an overactive inflammatory response. The *trans*-eQTL could feasibly contribute to inflammatory disease by increasing the expression of *IRF1* and its targets in response to LPS stimulation.

With the *IRF1* locus, we discovered a variant which is not active under baseline but affects *IRF1* expression under LPS stimulation. This context-dependence lends information as to how the variant might influence cellular networks in a physiological context. Immune-response eQTLs have been found to be more highly enriched for autoimmune GWAS loci than baseline eQTLs (Kim-Hellmuth et al., 2017). This variant might be widely important for immune response in humans and would not have been discovered under baseline conditions, emphasizing the importance of looking at how regulatory variants impact the transcriptome in a variety of cellular contexts.

Our attempts at validating regulatory variants also demonstrate the challenges involved. Isolating monoclonal cell lines, the approach utilized in chapter 3, is time- and resource-consuming. Additionally, unexpected genetic and epigenetic effects in monoclonal cell lines can create undetected noise. In this case, we could not utilize the polyclonal assay to validate the *cis* effect because the top variant is located in an enhancer, not in the gene body, and it would not have given us information about the *trans* effects of the variant. In differential expression analysis between clones with deletions at the eQTL variant locus and wildtype clones, as well as in the CRISPRi-treated cells, the *cis*-regulatory signal was not clear, and we failed to detect the *trans*-eGenes as significantly differentially expressed genes. It may be that the number of clones of each genotype needed to detect the *trans*-eQTL signal is much higher than the number isolated in this study, and in similar studies (Gupta et al., 2017; Soldner et al., 2016; Zhu et al., 2016). Logically, it seems as though you would need many fewer edited cell lines than individuals in eQTL studies because you are introducing the mutation into a constant genomic background. However, the results from ours and similar studies suggest that the number of clones needed for detecting transcriptome effects has been underestimated. In our analysis, we were able to capture the effects of CRISRPi and CRISPR perturbations by leveraging the power of analysis of the entire transcriptome that allows detection of even small effects.

In addition to undetected off-target and on-target mutations in these clones (Kosicki et al., 2018), transformed cell lines are known to have a multitude of chromosomal rearrangements, which may vary between individual cells in the cell line (Boone et al., 2014). For future studies utilizing the edited monoclonal cell line approach to validate regulatory effects of variants, it is imperative to isolate many clones of each genotype. Additionally, more thorough genotyping of the targeted locus in wild type and edited clones to detect large insertions or deletions at the targeted locus should be considered. This thorough genotyping can be performed using approaches such as long-range PCR, as was used in chapter 3 to eliminate one of the clones, or genomic qPCR which can be used to quantify the copy number of the genomic region surrounding the variant (D'haene et al., 2010). Furthermore, the proper choice of control is important. If a cell line which has not undergone editing is used as a control, there is a risk of confounding the results by detecting off-target or clone-specific effects in the edited clones instead of the effect of the specific variant. Therefore, control cell lines should be chosen from the same set of monoclonal cell lines as the edited cell lines, as we did in this study. Since there is generally no shortage of clones which do not harbor the desired mutation, isolating a number of these control cell lines is not a challenge, although their analysis by RNA-sequencing adds to the cost of these studies.

Another consideration to be taken into account when doing these types of validation experiments is that studying genetic effects in cell lines will never perfectly capture the effect of a variant that is observed in a population. All cell lines have limitations and the genomic rearrangements that occur in immortalized cell lines (Boone et al., 2014) no doubt can affect the genome and transcriptome of the cells. However, in order to find causality at a population-based association signal, experimental validation of the regulatory variant is essential. Additionally, eQTL studies, especially trans-eQTL, which are based on correlation in the population, are subject to false positives and negatives. Therefore, association studies and experimental validation studies provide complementary information on the effects of genetic variants.

Over the past fifteen years, researchers have discovered tens of thousands of GWAS associations to an extensive array of human traits. However, identifying a genetic signal is just the beginning of the story. The causal variant, the gene of action, and the mechanism of action of the association are often unclear. Therefore, our ability to identify and prove causality of regulatory variants in the population is of the utmost importance to the field of genomics. This elucidation is crucial for understanding the mechanism of variants that cause disease, and therefore discovering therapeutic targets. Despite the drawbacks associated with the different methods applied to functional genomics, further optimization and development in this field is necessary. This development is needed in order to expand our understanding of how the genetic variation between us humans defines the combination of traits that makes each of us unique.

References

Abdollahi, A., Lord, K.A., Hoffman-Liebermann, B., and Liebermann, D.A. (1991). Interferon regulatory factor 1 is a myeloid differentiation primary response gene induced by interleukin 6 and leukemia inhibitory factor: role in growth inhibition. Cell Growth Differ. *2*, 401–407.

Abou El Hassan, M., Huang, K., Eswara, M.B.K., Xu, Z., Yu, T., Aubry, A., Ni, Z., Livne-Bar, I., Sangwan, M., Ahmad, M., et al. (2017). Properties of STAT1 and IRF1 enhancers and the influence of SNPs. BMC Mol. Biol. *18*, 6–19.

Aguet, F., Barbeira, A.N., Bonazzola, R., Brown, A., Castel, S.E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., et al. (2019). The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv *7*, 1860–32.

Aird, E.J., Lovendahl, K.N., St Martin, A., Harris, R.S., and Gordon, W.R. (2018). Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template. Commun Biol *1*, 54.

Arbab, M., Srinivasan, S., Hashimoto, T., Geijsen, N., and Sherwood, R.I. (2015). Cloning-free CRISPR. Stem Cell Reports *5*, 908–917.

Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. *24*, 14–24.

Ben-Shachar, S., Khajavi, M., Withers, M.A., Shaw, C.A., van Bokhoven, H., Brunner, H.G., and Lupski, J.R. (2009). Dominant versus recessive traits conveyed by allelic mutations - to what extent is nonsense-mediated decay involved? Clin. Genet. *75*, 394–400.

Boone, M., Meuris, L., Lemmens, I., Van Roy, N., Soete, A., Reumers, J., Moisse, M., Plaisance, S.E.P., Drmanac, R., Chen, J., et al. (2014). Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. Nat Commun *5*, 1–12.

Brandt, M., and Lappalainen, T. (2017). SnapShot: Discovering Genetic Regulatory Variants by QTL Analysis. Cell *171*, 980–980.e981.

Brown, A.A., Viñuela, A., Delaneau, O., Spector, T.D., Small, K.S., and Dermitzakis, E.T. (2017). Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. Nat Genet *49*, 1747–1751.

Brynedal, B., Choi, J., Raj, T., Bjornson, R., Stranger, B.E., Neale, B.M., Voight, B.F., and Cotsapas, C. (2017). Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. Am. J. Hum. Genet. *100*, 581–591.

Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell *167*, 1398–1414.e24.

Chu, V.T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., and Kühn, R. (2015). Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat Biotechnol *33*, 543–548.

Ciorba, M.A., Bettonville, E.E., McDonald, K.G., Metz, R., Prendergast, G.C., Newberry, R.D., and Stenson, W.F. (2010). Induction of IDO-1 by immunostimulatory DNA limits severity of experimental colitis. J. Immunol. *184*, 3907–3916.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. Science *339*, 819–823.

Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med *9*, eaal5209.

D'haene, B., Vandesompele, J., and Hellemans, J. (2010). Accurate and objective copy number profiling using real-time quantitative PCR. Methods *50*, 262–270.

Demenais, F., Margaritte-Jeannin, P., Barnes, K.C., Cookson, W.O.C., Altmüller, J., Ang, W., Barr, R.G., Beaty, T.H., Becker, A.B., Beilby, J., et al. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat Genet *50*, 42–53.

Ellinghaus, D., Jostins, L., Spain, S.L., Cortes, A., Bethune, J., Han, B., Park, Y.R., Raychaudhuri, S., Pouget, J.G., Hübenthal, M., et al. (2016). Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet *48*, 510–518.

Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., et al. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science *343*, 1246949–1246949.

Ferreira, M.A., Vonk, J.M., Baurecht, H., Marenholz, I., Tian, C., Hoffman, J.D., Helmer, Q., Tillander, A., Ullemar, V., van Dongen, J., et al. (2017). Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nat Genet *49*, 1752–1757.

Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. Nature *513*, 120–123.

Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature *562*, 217–222.

Fortunato, G., Calcagno, G., Bresciamorra, V., Salvatore, E., Filla, A., Capone, S., Liguori, R., Borelli, S., Gentile, I., Borrelli, F., et al. (2008). Multiple sclerosis and hepatitis C virus infection are associated with single nucleotide polymorphisms in interferon pathway genes. J. Interferon Cytokine Res. *28*, 141–152.

Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet *42*, 1118–1125.

Furniss, D., Critchley, P., Giele, H., and Wilkie, A.O.M. (2007). Nonsense-mediated decay and the molecular pathogenesis of mutations in SALL1 and GLI3. Am. J. Med. Genet. A *143A*, 3150–3160.

Garvin, A.J., Khalaf, A.H.A., Rettino, A., Xicluna, J., Butler, L., Morris, J.R., Heery, D.M., and Clarke, N.M. (2019). GSK3β-SCFFBXW7α mediated phosphorylation and ubiquitination of IRF1 are required for its transcription-dependent turnover. Nucleic Acids Research *47*, 4476–4494.

Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., et al. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell *176*, 1516.

Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., and Liu, D.R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature *551*, 464–471.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell *154*, 442–451.

Grundberg, E., Small, K.S., Hedman, Å.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.-P., Meduri, E., Barrett, A., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet *44*, 1084–1089.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Gupta, R.M., Hadaya, J., Trehan, A., Zekavat, S.M., Roselli, C., Klarin, D., Emdin, C.A., Hilvering, C.R.E., Bianchi, V., Mueller, C., et al. (2017). A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. Cell *170*, 522–533.e15.

Gutschner, T., Haemmerle, M., Genovese, G., Draetta, G.F., and Chin, L. (2016). Post-translational Regulation of Cas9 during G1 Enhances Homology-Directed Repair. CellReports *14*, 1555–1566.

Harada, H., Takahashi, E., Itoh, S., Harada, K., Hori, T.A., and Taniguchi, T. (1994). Structure and regulation of the human interferon regulatory factor 1 (IRF-1) and IRF-2 genes: implications for a gene network in the interferon system. Mol. Cell. Biol. *14*, 1500–1509.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

Heigwer, F., Kerr, G., and Boutros, M. (2014). E-CRISP: fast CRISPR target site identification. Nature Publishing Group *11*, 122–123.

Holbrook, J.A., Neu-Yilik, G., Hentze, M.W., and Kulozik, A.E. (2004). Nonsense-mediated decay approaches the clinic. Nat Genet *36*, 801–808.

Honda, K., and Taniguchi, T. (2006). IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors. Nat. Rev. Immunol. *6*, 644–658.

Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508.

Hug, N., Longman, D., and Cáceres, J.F. (2016). Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Research *44*, 1483–1495.

Inoue, F., and Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. Genomics *106*, 159–164.

Iwanaszko, M., and Kimmel, M. (2015). NF-κB and IRF pathways: cross-regulation on target genes promoter level. BMC Genomics *16*, 307.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821.

Johnston, J.J., Olivos-Glander, I., Killoran, C., Elson, E., Turner, J.T., Peters, K.F., Abbott, M.H., Aughton, D.J., Aylsworth, A.S., Bamshad, M.J., et al. (2005). Molecular and clinical analyses of Greig cephalopolysyndactyly and Pallister-Hall syndromes: robust phenotype prediction from the type and position of GLI3 mutations. The American Journal of Human Genetics *76*, 609–622.

Kalff-Suske, M., Wild, A., Topp, J., Wessling, M., Jacobsen, E.M., Bornholdt, D., Engel, H., Heuer, H., Aalfs, C.M., Ausems, M.G., et al. (1999). Point mutations throughout the GLI3 gene cause Greig cephalopolysyndactyly syndrome. Hum. Mol. Genet. *8*, 1769–1777.

Kamijo, R., Harada, H., Matsuyama, T., Bosland, M., Gerecitano, J., Shapiro, D., Le, J., Koh, S.I., Kimura, T., and Green, S.J. (1994). Requirement for transcription factor IRF-1 in NO synthase induction in macrophages. Science *263*, 1612–1615.

Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am. J. Hum. Genet. *104*, 65–75.

Kim, S., Becker, J., Bechheim, M., Kaiser, V., Noursadeghi, M., Fricker, N., Beier, E., Klaschik, S., Boor, P., Hess, T., et al. (2014). Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes. Nat Commun *5*, 5236.

Kim-Hellmuth, S., Bechheim, M., Pütz, B., Mohammadi, P., Nédélec, Y., Giangreco, N., Becker, J., Kaiser, V., Fricker, N., Beier, E., et al. (2017). Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. Nat Commun *8*, 266.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310–315.

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424.

Kosicki, M., Tomberg, K., and Bradley, A. (2018). Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nat Biotechnol *36*, 765–771.

Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat Commun *8*, 15824.

Kröger, A., Köster, M., Schroeder, K., Hauser, H., and Mueller, P.P. (2002). Activities of IRF-1. J. Interferon Cytokine Res. *22*, 5–14.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research *46*, D1062–D1067.

Langlais, D., Barreiro, L.B., and Gros, P. (2016). The macrophage IRF8/IRF1 regulome is required for protection against infections and is associated with chronic inflammation. J. Exp. Med. *213*, 585–603.

Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Lappalainen, T., Scott, A.J., Brandt, M., and Hall, I.M. (2019). Genomic Analysis in the Age of Human Genome Sequencing. Cell *177*, 70–84.

Lee, M.N., Ye, C., Villani, A.-C., Raj, T., Li, W., Eisenhaure, T.M., Imboywa, S.H., Chipendo, P.I., Ran, F.A., Slowikowski, K., et al. (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science *343*, 1246980–1246980.

Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J., et al. (2017). The impact of rare variation on gene expression across tissues. Nature *550*, 239–243.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Liu, J., Cao, S., Herman, L.M., and Ma, X. (2003). Differential regulation of interleukin (IL)-12 p35 and p40 gene expression and interferon (IFN)-gamma-primed IL-12 production by IFN regulatory factor 1. J. Exp. Med. *198*, 1265–1276.

Liu, T., Zhang, L., Joo, D., and Sun, S.-C. (2017). NF-κB signaling in inflammation. Signal Transduct Target Ther *2*, 17023.

Lu, Y.-C., Yeh, W.-C., and Ohashi, P.S. (2008). LPS/TLR4 signal transduction pathway. Cytokine *42*, 145–151.

Maruyama, T., Dougan, S.K., Truttmann, M.C., Bilate, A.M., Ingram, J.R., and Ploegh, H.L. (2015). Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. Nat Biotechnol *33*, 538–542.

Matsuyama, T., Kimura, T., Kitagawa, M., Pfeffer, K., Kawakami, T., Watanabe, N., Kündig, T.M., Amakawa, R., Kishihara, K., and Wakeham, A. (1993). Targeted disruption of IRF-1 or IRF-2 results in abnormal type I IFN gene induction and aberrant lymphocyte development. Cell *75*, 83–97.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122–14.

Miller, J.N., and Pearce, D.A. (2014). Nonsense-mediated decay in genetic disease: friend or foe? Mutat Res Rev Mutat Res *762*, 52–64.

Miyamoto, M., Fujita, T., Kimura, Y., Maruyama, M., Harada, H., Sudo, Y., Miyata, T., and Taniguchi, T. (1988). Regulated expression of a gene encoding a nuclear factor, IRF-1, that specifically binds to IFN-beta gene regulatory elements. Cell *54*, 903–913.

Mohammadi, P., Castel, S.E., Brown, A.A., and Lappalainen, T. (2017). Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. Genome Res. *27*, 1872–1884.

Nagy, E., and Maquat, L.E. (1998). A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem. Sci. *23*, 198–199.

Negishi, H., Fujita, Y., Yanai, H., Sakaguchi, S., Ouyang, X., Shinohara, M., Takayanagi, H., Ohba, Y., Taniguchi, T., and Honda, K. (2006). Evidence for licensing of IFN-gamma-induced IFN regulatory factor 1 transcription factor by MyD88 in Toll-like receptor-dependent gene induction program. Proc. Natl. Acad. Sci. U.S.a. *103*, 15136–15141.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet *6*, e1000888–10.

Ogasawara, K., Hida, S., Azimi, N., Tagaya, Y., Sato, T., Yokochi-Fukuda, T., Waldmann, T.A., Taniguchi, T., and Taki, S. (1998). Requirement for IRF-1 in the microenvironment supporting development of natural killer cells. Nature *391*, 700–703.

Oikonomou, P., Goodarzi, H., and Tavazoie, S. (2014). Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts. CellReports *7*, 281–292.

Pan, P.-H., Cardinal, J., Li, M.-L., Hu, C.-P., and Tsung, A. (2013). Interferon regulatory factor-1 mediates the release of high mobility group box-1 in endotoxemia in mice. Chin. Med. J. *126*, 918–924.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc *9*, 171–181.

Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.

Ren, Z., Wang, Y., Liebenson, D., Liggett, T., Goswami, R., Stefoski, D., and Balabanov, R. (2011). IRF-1 signaling in central nervous system glial cells regulates inflammatory demyelination. J. Neuroimmunol. *233*, 147–159.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. (Nature Publishing Group), pp. 405–424.

Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., Tsang, E.K., Karczewski, K.J., Maller, J.B., Kukurba, K.R., DeLuca, D.S., Fromer, M., et al. (2015). Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science *348*, 666–669.

Rosadini, C.V., and Kagan, J.C. (2017). Early innate immune responses to bacterial LPS. Curr. Opin. Immunol. *44*, 14–19.

Saha, A., and Battle, A. (2018). False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. F1000Res *7*, 1860.

Santiago-Algarra, D., Dao, L.T.M., Pradel, L., España, A., and Spicuglia, S. (2017). Recent advances in high-throughput approaches to dissect enhancer function. F1000Res *6*, 939.

Schwabe, G.C., Tinschert, S., Buschow, C., Meinecke, P., Wolff, G., Gillessen-Kaesbach, G., Oldridge, M., Wilkie, A.O., Kömec, R., and Mundlos, S. (2000). Distinct mutations in the receptor tyrosine kinase gene ROR2 cause brachydactyly type B. The American Journal of Human Genetics *67*, 822–831.

Schwarzer, W., Witte, F., Rajab, A., Mundlos, S., and Stricker, S. (2009). A gradient of ROR2 protein stability and membrane localization confers brachydactyly type B or Robinow syndrome phenotypes. Hum. Mol. Genet. *18*, 4013–4021.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelson, T., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84–87.

Shi, W., Zou, R., Yang, M., Mai, L., Ren, J., Wen, J., Liu, Z., and Lai, R. (2019). Analysis of Genes Involved in Ulcerative Colitis Activity and Tumorigenesis Through Systematic Mining of Gene Co-expression Networks. Front Physiol *10*, 662.

Soldner, F., Stelzer, Y., Shivalila, C.S., Abraham, B.J., Latourelle, J.C., Barrasa, M.I., Goldmann, J., Myers, R.H., Young, R.A., and Jaenisch, R. (2016). Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. Nature *533*, 95–99.

Song, J., Yang, D., Xu, J., Zhu, T., Chen, Y.E., and Zhang, J. (2016). RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. Nat Commun *7*, 10548–7.

Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., Balzereit, D., Lehrach, H., and Yaspo, M.-L. (2014). Influence of RNA extraction methods and library selection schemes on RNA-seq data. BMC Genomics *15*, 675–13.

Takeda, K., and Akira, S. (2005). Toll-like receptors in innate immunity. Int. Immunol. *17*, 1–14.

Tamura, T., Yanai, H., Savitsky, D., and Taniguchi, T. (2008). The IRF Family Transcription Factors in Immunity and Oncogenesis. Annu. Rev. Immunol. *26*, 535–584.

Tewhey, R., Kotliar, D., Park, D.S., Liu, B., Winnicki, S., Reilly, S.K., Andersen, K.G., Mikkelsen, T.S., Lander, E.S., Schaffner, S.F., et al. (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell *165*, 1519–1529.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Turner, M.D., Nedjai, B., Hurst, T., and Pennington, D.J. (2014). Cytokines and chemokines: At the crossroads of cell signalling and inflammatory disease. Biochim. Biophys. Acta *1843*, 2563–2582.

van Arensbergen, J., Pagie, L., FitzPatrick, V.D., de Haas, M., Baltissen, M.P., Comoglio, F., van der Weide, R.H., Teunissen, H., Võsa, U., Franke, L., et al. (2019). High-throughput identification of human SNPs affecting regulatory element activity. Nat Genet *526*, 68.

Vergoulis, T., Vlachos, I.S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., Gerangelos, S., Koziris, N., Dalamagas, T., and Hatzigeorgiou, A.G. (2012). TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Research *40*, D222–D229.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. 1–57.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Wellcome Trust Case Control Consortium, Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M.M., Auton, A., Myers, S., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet *44*, 1294–1301.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research *42*, D1001–D1006.

Wen, X., Lee, Y., Luca, F., and Pique-Regi, R. (2016). Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. Am. J. Hum. Genet. *98*, 1114–1129.

Wen, X., Luca, F., and Pique-Regi, R. (2015). Cross-population joint analysis of eQTLs: fine mapping and functional annotation. PLoS Genet *11*, e1005176.

Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature Publishing Group *45*, 1238–1243.

Wu, X.-F., Ouyang, Z.-J., Feng, L.-L., Chen, G., Guo, W.-J., Shen, Y., Wu, X.-D., Sun, Y., and Xu, Q. (2014). Suppression of NF-κB signaling and NLRP3 inflammasome activation in macrophages is responsible for the amelioration of experimental murine colitis by the natural compound fraxinellone. Toxicol. Appl. Pharmacol. *281*, 146–156.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806–1254806.

Yang, Y.-C.T., Di, C., Hu, B., Zhou, M., Liu, Y., Song, N., Li, Y., Umetsu, J., and Lu, Z.J. (2015). CLIPdb: a CLIP-seq database for protein-RNA interactions. BMC Genomics *16*, 51.

Yao, C., Joehanes, R., Johnson, A.D., Huan, T., Liu, C., Freedman, J.E., Munson, P.J., Hill, D.E., Vidal, M., and Levy, D. (2017). Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. Am. J. Hum. Genet. *100*, 571–580.

Yeo, N.C., Chavez, A., Lance-Byrne, A., Chan, Y., Menn, D., Milanova, D., Kuo, C.-C., Guo, X., Sharma, S., Tung, A., et al. (2018). An enhanced CRISPR repressor for targeted mammalian gene regulation. Nature Publishing Group *15*, 611–616.

Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T., and Flicek, P.R. (2015). The ensembl regulatory build. Genome Biol. *16*, 56.

Zhao, G.-N., Jiang, D.-S., and Li, H. (2015). Interferon regulatory factors: at the crossroads of immunity, metabolism, and disease. Biochim. Biophys. Acta *1852*, 365–378.

Zhu, D.-L., Chen, X.-F., Hu, W.-X., Dong, S.-S., Lu, B.-J., Rong, Y., Chen, Y.-X., Chen, H., Thynn, H.N., Wang, N.-N., et al. (2018). Multiple Functional Variants at 13q14 Risk Locus for Osteoporosis Regulate RANKL Expression Through Long-Range Super-Enhancer. J. Bone Miner. Res. *33*, 1335–1346.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet *48*, 481–487.