FINE-MAPPING OF HUMAN GENETIC REGULATORY VARIANTS

A Dissertation
Presented to
The Academic Faculty

by

Ruoyu Tian

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biological Sciences

Georgia Institute of Technology
August 2020

FINE-MAPPING OF HUMAN GENETIC REGULATORY VARIANTS

Approved by:

Dr. Greg Gibson, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Melissa Kemp
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Ciaran M Lee
APC Microbiome Ireland
University College Cork

Dr. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. James Dahlman
School of Biomedical Engineering
Georgia Institute of Technology

Date Approved: May, 15, 2020

Dedicated to my mother and family

ACKNOWLEDGEMENTS

My four-year Ph.D. journey at Georgia Institute of Technology has been filled with joy, enthusiasm, hard works and challenges. I would like to thank all my mentors, family and friends whose thoughtful help and encouragement always cheers me up in the completion of my Ph.D. First of all, I feel grateful to have Dr. Greg Gibson to be my advisor. His curiosity and passion for science inspires my interest in human genetics. He is a lifelong learner which motivates me to explore the unknown. His guidance always keeps me on the right direction, and his encouragement and support ensure my accomplishment of research.

Secondly, I want to thank my collaborators, Dr. Gang Bao, Dr. Ciaran Lee and Dr. Yidan Pan for their insightful ideas and experimental support, which helped a lot for my eQTL fine-mapping project. I also appreciate the suggestions from my committee members, Dr. Ciaran Lee, Dr. James Dahlman, Dr. Melissa Kemp and Dr. King Jordan. They provide valuable guidance and ideas on research design and edits of my thesis.

I would also thank lab members in Gibson's lab, especially Dr. Dalia A. Gulick, Dr. Urko Marigorta, Dr. Biao Zeng, Dr. Swetha Garimalla, Khalid Alhuminidi, Angela Mo, Meixue Duan, Sini Nagpal and Maggie Brown. We discuss scientific questions, share pipelines and update daily fun facts. We are colleagues, but also friends. The friendship from Georgia Tech is my precious treasure.

Last but not least, I would thank my great family for unconditional love and support. I would express my deepest gratitude to my mom, who raises me up and teaches me how to be kind, respectful, humble, independent and successful. I feel grateful to be your daughter.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

3’-LTR	3’-long terminal repeats
7-AAD	7-Aminoactinomycin D
AMD	Age-related macular degeneration
ANOVA	ANalysis Of VAriance
ATCC	American Type Culture Collection
CADD	Combined Annotation Dependent Depletion
CAGE	Consortium for the Architecture of Gene Expression
CNV	Copy Number Variation
COSMID	CRISPR Off-target Sites with Mismatches, Insertions, and Deletions
CPM	Counts per million reads
CRISPR/Cas9	Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-associated protein 9
CRISPRa	CRISPR activation
CRISPRi	CRISPR interference
CROP-seq	CRISPR droplet sequencing
dbSNP	The Single Nucleotide Polymorphism Database
dCas9	deficient Cas9
eGene	expression Gene
EP	Evolutionary Probability
eQTL	expression Quantitative Trait Locus
FACS	Fluorescence-activated cell sorting
FBS	Fetal bovine serum
FDR	False Discovery Rate
FHS	Framingham Heart Study

GATK	The Genome Analysis Toolkit
GFP	Green fluorescent protein
GO	Gene ontology
gRNA	guide RNA
GTEx	Genotype-Tissue Expression project
GWAS	Genome-Wide Association Study
HDR	Homology-directed repair
hg19	human genome, version 19
hg38	human genome, version 38
IBD	Inflammatory bowel disease
IMR	Interpersonal motor resonance
indel	Insertion or Deletion
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MERA	Multiplexed Editing Regulatory Assay
Mosaic-seq	Mosaic single-cell analysis by indexed CRISPR sequencing
MPRA	Massively parallel reporter assays
ncRVIS	noncoding Residual Variation Intolerance Score
NHEJ	Nonhomologous end joining
OMIM	Online Mendelian Inheritance in Man
ORF	Open reading frame
PAM	Protospacer Adjacent Motif
PBMC	Peripheral blood mononuclear cells
PBS	Phosphate buffered saline
PCA	Principal Component Analysis

PCR	Polymerase Chain Reaction
pLI	the probability of being loss-of-function intolerant
PTV	Protein-truncating variants
PVCA	Principal Variance Component Analysis
qRT-PCR	quantitative real-time polymerase chain reaction
RA	Retinoic acid
raQTL	Reporter assay QTL
RefSeq	Reference Sequence
RIVER	RNA-informed variant effect on regulation
rRNA	Ribosomal ribonucleic acid
RVIS	Residual Variation Intolerance Score
rWGS	rapid Whole Genome Sequencing
scRNA-seq	single-cell RNA sequencing
sdu	standard deviation unit
sgRNA	single guide RNA
SNP	Single Nucleotide Polymorphism
SuRE	Survey of regulatory elements
TMM	Trimmed mean of M-value
TSS	Transcription Start Site
UMI	Unique Molecular Identifier
urWGS	ultra-rapid Whole Genome Sequencing,
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WTCCC	Wellcome Trust Case Control Consortium

SUMMARY

The majority of GWAS (Genome-Wide Association Study) identified common genetic variants map to regulatory regions of gene, and are likely to influence disease risk by affecting gene expression. One of the most important challenges is to experimentally fine-map causal regulatory variants that typically lie in credible intervals of 100 or more variants. Another large proportion of genetic variants, rare variants, are expected to have large effects causing disease in individual, but are not detectable in GWAS. Herein, I provide both experimental and computational approaches for fine-mapping common and rare genetic variants accounting for medium and large effect on population or individual. First, I describe a single cell clone-based strategy for targeted single-nucleotide polymorphism (SNP) evaluation wherein microindels are introduced by CRISPR/Cas9. Multiple constraints, including the variability in mutability, clonal genotype and expense, render this approach infeasible for fine-mapping 10%-20% moderate effect size expression SNPs (eSNPs), which is also validated in a simulation study. Subsequently, I switch to a moderate-throughput parallel screening tool that characterizes multiplexed CRISPR/Cas9 perturbed transcriptomes by single-cell RNA-seq, called “expression CROP-seq”. Two causal SNPs, rs2251039 and rs35675666, are identified that significantly alter the expression of *CISD1* and *PARK7*, respectively. The sites overlap with chromatin accessibility peaks and are risk loci of inflammatory bowel disease. Expression CROP-seq reduces the variability identified in previous method and is powerful to screen genetic regulatory variants within credible intervals. Finally, to extend its application to rare variants, I develop a novel gene categorization system according to gene intolerance to promoter polymorphism and depletion of rare regulatory variants with GTEx v8 data. 49

GTEx tissues are clustered into functional groups with gene features. It supports the use of tissue-gene genomic annotation for prioritization of GWAS tagged risk loci. In summary, this work comprehensively describes and evaluates two CRISPR/Cas9-based eSNP screening systems. The use of rare regulatory variants in gene classification with tissue information demonstrates its potential in rare disease diagnoses. Both researches inevitably contribute to the genetic interpretation of human complex disease and personalized medicine in post-GWAS era.

CHAPTER 1. INTRODUCTION

1.1 Genome-wide association study (GWAS)

Since completion of the human genome project in 2003[1], millions of SNPs have been found throughout the human genome. With the implementation of SNP genotyping, scientific questions concerning the extent of correlation between SNPs and common diseases or quantitative traits have been addressed. GWAS, namely genome wide study of the association of genotypes with disease status or traits, was first reported in 2005. The first GWAS study was conducted with 96 cases and 50 controls using 116,204 genotyped SNPs, and found an intronic and common variant in the *CFH* gene strongly associated with age-related macular degeneration (AMD), which increased likelihood by 7.4 in individuals homozygous for the risk allele[2]. One year later, two follow-up studies[3,4] found another risk locus in *HTRA1* for AMD, and a promoter region SNP rs11200638 was identified as the causal variant.

In general, GWAS detects common variants with small odds ratios (from 1.2 to 2), thousands of which each explain of the order of 0.1% of common disease risk[5]. The most well-known GWAS study was published in 2007 from the Wellcome Trust Case Control Consortium, WTCCC, which examined 14,000 cases and 3,000 controls for 7 common diseases, and found 58 risk loci[6]. Subsequently, the publication of GWA studies has been increasing dramatically. As of February 2020, 172,351 associations have been collated in the GWAS catalog (https://www.ebi.ac.uk/gwas/).

Figure 1.1 SNP-trait associations with p-value ≤ 5.0×10^{-8}, published in the GWAS Catalog (as of Feb, 2020).

Investments are now being made in post-GWAS functional characterization. First, fine-mapping. GWAS mostly discovers tagging SNPs only, and the identity of the causal SNP remains unknown. A possible reason is the linkage disequilibrium (LD) of genome. Linkage disequilibrium is the non-random alleles association between loci in the same chromosome or at different chromosome. And different population have distinct LD structures. The LD structure is a confounding factor for identifying causal variant in GWAS, but it also helps to efficiently design a small fraction of markers genotyped without actually obtaining whole genome sequences. Targeted resequencing the region surrounding tagging SNPs makes it possible to refine the list of candidate causal SNPs. For example, deep next generation resequencing of 56 associated risk loci of inflammatory bowel disease (IBD) identified additional new, rare and probably functional variants in 8 genes[7]. Second, integration of functional regulatory data. Most of the GWAS-identified tagging SNPs are located in regulatory regions, such as enhancer and promoters. Functional genomics data integration,

including DNase sensitivity, histone modification and epigenetics, is an efficient approach to prioritize causal variants and also provide functional interpretation of GWAS identified SNPs. Third, expression quantitative trait loci, eQTL. These are genomic loci associated with gene expression. Since GWAS tagging SNPs sometimes overlap with eQTL signals, studying eQTL is a straightforward way to study disease and phenotype association. Moreover, another advantage of eQTL study is that it requires smaller sample sizes to achieve sufficient power. eQTLs tend to explain more expression variation than GWAS SNPs, because their effect size is much higher. Furthermore, eQTL studies can be easily performed in cell lines, primary cells and tissues. eQTL will be discussed further in the following section.

1.2 eQTL

An eQTL is a locus that explains a significant proportion of the variation in the transcript abundance of a gene. eQTLs are classified as *cis*-eQTLs and *trans*-eQTLs. *cis*-eQTLs are typically defined as being located within 1Mb from either side of transcription start sites (TSS), and *trans*-eQTLs are located farther than 5Mb from the TSS, or on another chromosome[8]. Discovery of *cis*-eQTL has a much higher yield than *trans*-eQTL because they tend to explain considerably more variation[9], most likely because they directly regulate expression of the nearby gene[10].

For human eQTL mapping, the gene expression data is mostly from peripheral blood, since it is the most accessible tissue. However, some eQTLs are cell-type specific, tissue-specific and context-dependent, so discovering eQTL only from blood is not sufficient to uncover their relationship to disease. Dimas et.al.[11] conducted *cis*-eQTL analysis on 75 individuals for three cell types, primary fibroblasts, lymphoblastoid cell lines and T cells, and found that 79.5% of the eQTLs were cell-type specific. This early study implied that only 20%

of the eQTL were shared either by two or three cell types, but subsequent work has gradually increased that percentage. Several other research groups have purified various subsets of peripheral blood cells. For example, a comparison of naïve T CD4+ cells and monocytes using RNA sequencing, showed that disease and trait-associated *cis*-eQTLs display more cell specificity than on average: 46% were monocyte-specific, 29% were T cell-specific and 25% were shared[12]. On the other hand, CD4+ and CD8+ T-cells share the vast majority of eQTL[13], with some functionally important exceptions, and enrichment studies are likely to highlight which cell types mediate which diseases.

Initiated in 2013, the Genotype-Tissue Expression (GTEx) project has provided insights into the correlation between genotype and tissue-specific gene expression from autopsies of recently deceased individuals[14]. The most recent GTEx v8 data reports 17,382 RNA-sequencing samples from 54 tissues of 948 post-mortem donors, 838 out of which also have genotypes available, mostly from whole genome sequencing[15]. There are 18,262 protein-coding eGenes (94.7% of all eGenes) discovered with at least one eSNP in at least one tissue and 4,278,636 genetic *cis*-eQTL. The tissue-specific eQTLs are either highly tissue-specific or highly shared across tissues.

1.3 CRISPR/Cas9 genome editing

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/Cas is an adaptive immune system in bacteria and archaea[16] that has been bioengineered into a genome editing tool. Jennifer Doudna and colleagues first showed that it is possible for the endonuclease cas9 to cleave targeted genomic DNA in vitro when guided to the site by a guide RNA (sgRNA)[17]. Around about the same time, Feng Zhang and colleagues first reported the application of a CRISPR/Cas system in human cell lines and mice cell lines for efficient and

precise editing of the genome[18]. This RNA-guided genome editing tool has since been widely used in bacteria[19], yeast[20], zebrafish[21], mice[22] and plants[23]. The mechanism of this technology is that Cas9 nuclease introduces a double-stranded break in the targeted site under the guidance of the sgRNA. Then, two potential DNA repair pathways, either nonhomologous end joining (NHEJ) or homology-directed repair (HDR), are engaged in the repair process. In the absence of a repair template, NHEJ is activated, which introduces unpredictable patterns of insertions and deletions near the edited site. This reaction can reach high editing efficiency up to 20%-60%[24]. With the introduction of template, HDR is activated to precisely replace the target site with the exogenous template DNA, but the editing efficiency is only 0.5%-20%.

Off-target effects, where cleavage is observed at a similar sequence elsewhere in the genome, are common in human cell lines. The CRISPR/Cas9 system can be highly active even where fives mismatches occur in the off-target site[25]. This is not random and can be predicted to some extent in silico[26-28]. The research group of Gang Bao at Rice University have developed a web interface software COSMID (CRISPR Off-target Sites with Mismatches, Insertions, and Deletions), which predicts the potential off-target sites[26]. Based on input guide RNA sequence, COSMID identifies potential off-target sites through the genome with specified number of mismatched bases and insertions or deletions.

1.4 Experimental approaches for screening regulatory elements

1.4.1 Massively parallel reporter assays (MPRA)

Massively parallel reporter assays couple short segments of potentially regulatory DNA to barcodes which are transcribed following transfection into cells or animals. An invariant promoter-ORF (open reading frame) segment is inserted between the screened

variants and barcode. Sequencing approaches allow identification of under- or over-represented barcodes indicating differential expression due for example to polymorphisms (Figure 1.2).

MPRA has been applied to investigate the transcriptional regulation rules in yeast[29]. A fluorescence reporter based MPRA was used to systematically survey 6,500 yeast promoters' for regulatory activity, thereby documenting the effects on gene expression of binding-site location, number, orientation and affinity. Soon after this study, the application on human enhancer screening was reported[30]. This group tested 2,104 wild type enhancers and 3,314 mutated enhancers with disrupted motifs in K562 and HepG2 cell lines. They found that enhancer activity is cell-type specific and that predicted activator motifs and evolutionary conserved motifs can be used to predict enhancer activity.

The application of MPRA for mapping genetic variants down to single nucleotide level has been reported. The screening scale ranges from tens to hundreds and up to millions of variants. Tewhey et. al[31], first screened 79,000 single nucleotide variants (single-nucleotide and small insertion/deletion of both alleles) in two lymphoblastoid cell lines, NA12878 and NA19239, and HepG2 cells. They then used a refined screening library with 7,500 variants replicating the positive signals from the first trail. 852 variants showed differential expression between alleles, most of which were clustered near the lead eQTL and GWAS SNP, in regions of high so-called linkage disequilibrium. The limitation of this approach is that it can only detect regulatory activity leading to increased expression since baseline expression is typically low, but it can reach a resolution of 3 positive signals in one eQTL credible interval. MPRA alone is not an efficient tool with which to identify the

causal variants of disease or eQTL. CRISPR/Cas9 based genome editing experiments in cells and animals are need for further functional validation.

Another large scale regulatory screening was conducted by modified MPRA with 5.9 million SNPs, 57% of which are common variants, in K562 and HepG2 cell lines, called the survey of regulatory elements (SuRE)[32]. The identified variants that caused differential expression are called reporter assay QTL (raQTL), and approximately 20,000 were identified in each tested cell line. Most of raQTLs are enriched in promoter, enhancer and DNA hypersensitive sites, and they are likely to change expression by altering transcription factor binding affinity. The identified raQTLs were particularly useful when integrated with functional data alongside GWAS signals to prioritize disease risk variants.

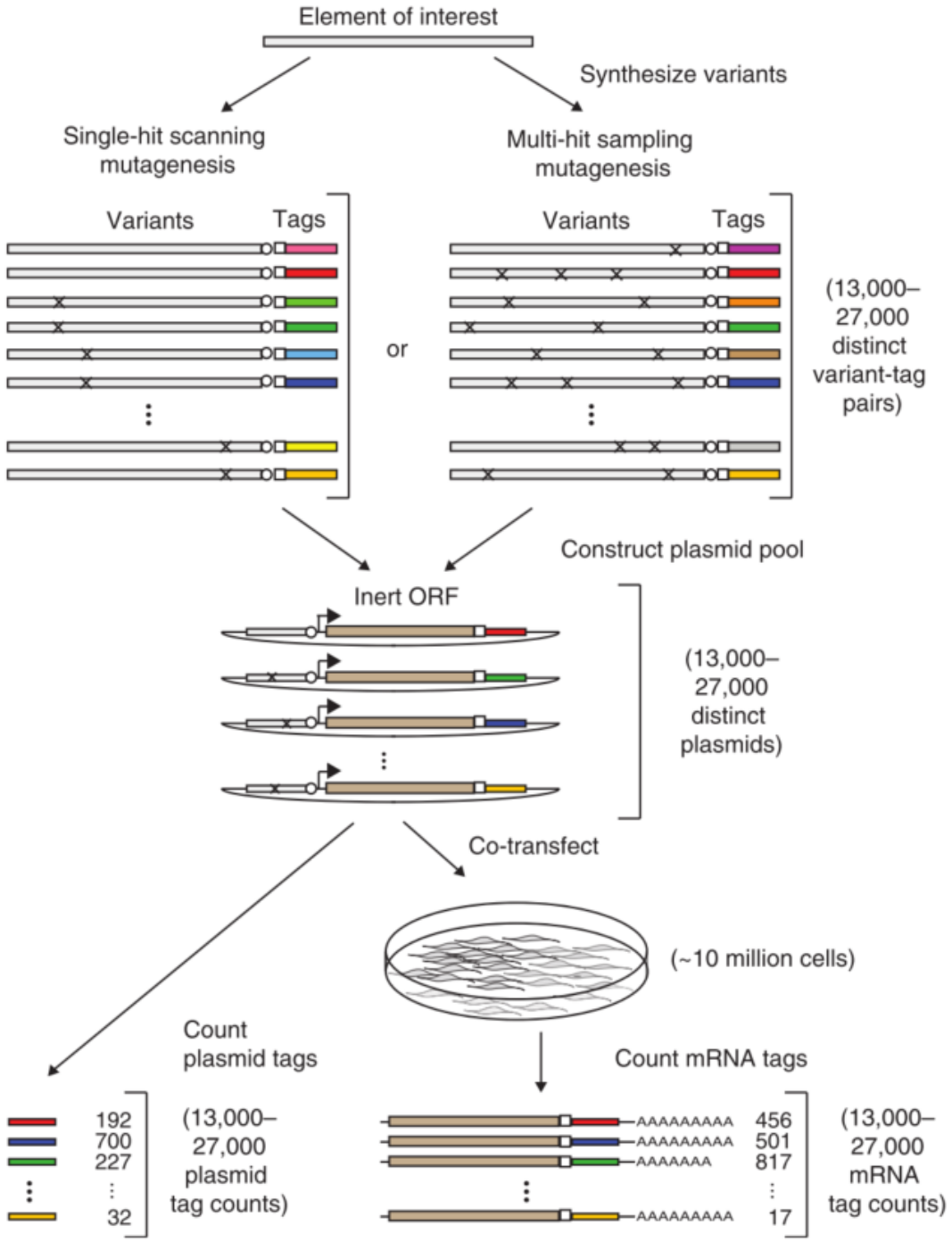

Figure 1.2 Schematic of MPRA[33]. Open reading frame (ORF) is inserted into elements of interest and coupled barcodes. The construct library is transfected into a population of

cultured cells. The ratio of mRNA counts over plasmid counts quantified by deep sequencing can be used to infer the element regulatory activity.

1.4.2 CRISPR/Cas9 based bulk cell screening

There are two major approaches to screening of regulatory variants with CRISPR/Cas9 genome editing or CRISPRi functional genome inhibition. One is sequencing of barcodes of libraries of guide RNA from bulk cellular populations, assaying for enrichment or depletion of guide RNAs that target elements required for driving transcription. Another one is parallel perturbation of regulatory variants with a pool of guide RNAs in a library, then quantifying the target gene expression at single-cell resolution by single-cell RNA-seq. I will discuss the second strategy in the next subsection.

Feng Zhang group[34] CRISPRed comprehensively across 100kb regions 5' and 3' of the *NF1*, *NF2* and *CUL3* genes, selecting for loss-of-mutations that resulted in vemurafenib resistance in a BRAF-mutated human melanoma cell line A375. Each cell had been mutagenized with one guide RNA. The sgRNA count differential measured by deep sequencing between cells treated with and without vemurafenib was used as the bulk output measurement of which guides led to resistance. A similar approach named multiplexed editing regulatory assay (MERA) was used to tile thousands of mutations across 40 kb of the *cis*-regulatory region of four stem cell-specific genes[35]. Furthermore, regulatory elements screening is not limited to the single gene-level, since it has also been expanded to the genomic wide scale by massively parallel tiling of binding sites. Such screening of transcription factor binding sites for two genes has been reported in two different studies[36,37]: one of transcription factor p53 and ERα binding sites, and the other of FOXA1 and CTCF binding sites, measuring gRNA counts in a breast cancer cell line.

Recent CRISPRi and CRISPRa pooled inhibition and activation screening assays utilize modified Cas9 proteins that bind to but do not cut the target site. They have enabled high-throughput screening of genomic elements influencing transcription[38]. Although this study perturbed 5,920 human candidate enhancers with dCas9, each cell had on average 20 multiplex perturbations, increasing statistical power but potentially increasing off-target effects. Adapting the eQTL analysis framework, they described 664 *cis*-enhancer localization events, but this approach does not fine-map regulatory polymorphisms.

1.4.3 CRISPR/Cas9 based single-cell RNA sequencing

Five publications to date have described related methods using single cell RNAseq (scRNA-seq) as the read-out of pools of cells genetically perturbed via lentiviral CRISPR. The methods are known as Perturb-seq[39,40], CRISP-seq[41], CROP-seq[42], and Mosaic-seq[43].

The primary idea is that each individual cell is paired with a lentiviral vector which is assumed to have disrupted the target of the gRNA it carries. The lentiviral vector is a third-generation lentiviral vector that contains two expression cassettes: one is an RNA polymerase II-driven guide barcode (GBC), the other is an RNA polymerase III driven-gRNA expression cassette. Each gRNA is labeled with a guide barcode. During the step of single cell cDNA generation, each cell is encapsulated with beads as in regular scRNA-seq so that each cell is labeled with a cell barcode, and each transcript is labeled with a unique molecular identifier, or UMI. All of the transcripts are pooled together during the sequencing stage, and can be demultiplexed computationally by the identification of the three levels of barcodes, which are also sequenced. In CROP-seq[42] (Figure 1.3), the lenti-construct incorporates the gRNA cassette into the 3'-LTR region. During lentiviral

integration, the gRNA cassette is copied so that both the gRNA and barcode can be transcribed into small RNAs, one of which receives a polyA tail and can then be sequenced. In this way, the gRNA is sequenced without adding a guide barcode. Guide RNA pools have been reported targeting genes or enhancer regions, but there are no reports using guide RNA to target specific variants in non-coding regions within eQTL credible intervals.

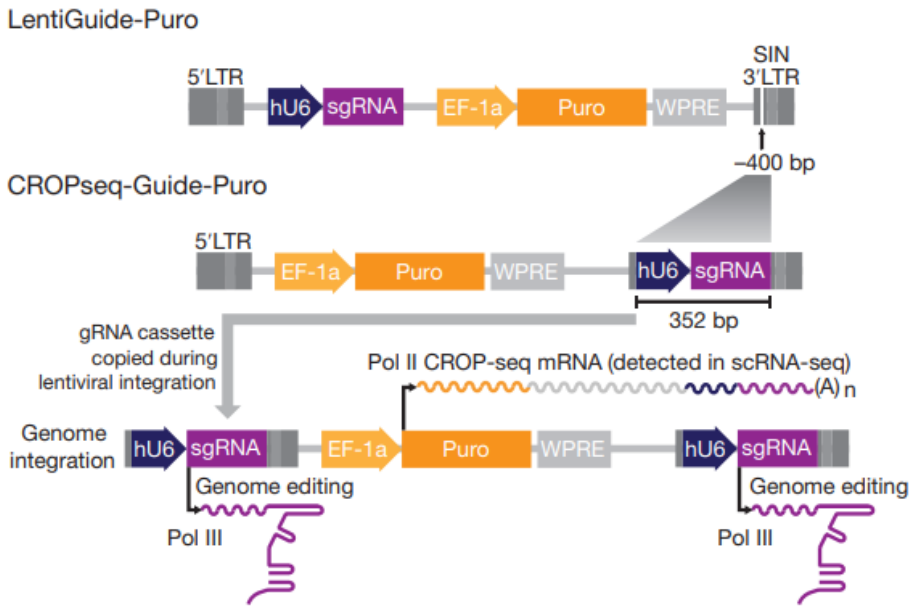

Figure 1.3 CROP-seq lentiviral construct[42]

1.5 Rare variants

Geneticists aim to explain the etiology of diseases in terms of identified effects of individual genetic variants, which include single-nucleotide polymorphism (SNP) and copy number variation. And each variant can be measured by risk allele frequency in the population and the effect size is typically described as an odds ratio. The product of the risk allele frequency and effect size determines the amount of variance explained in the population, and hence the statistical power to identify the causality of the variant. The higher the risk allele

frequency, or effect size, the higher the statistical power. Most interest and studies focus on the identification of associations for genetic variants between the diagonal dotted lines in Figure 1.4, namely common variants with small effect size, rare variants with large effect size and intermediate alleles with moderate effect size. GWAS as discussed in section 1.1 is powerful to identify the association between common variants with common diseases or traits. Most GWAS-identified risk loci only explain a small proportion of the heritability. For example, 18 identified risk loci for type two diabetes only account for 6% of the heritability[44]; 20% of the total heritability is explained by 32 identified risk loci associated with Crohn's disease. A major challenge for the field is to reduce GWAS intervals to individual causal variants as this will illuminate mechanisms, and improve the development of polygenic risk scores.

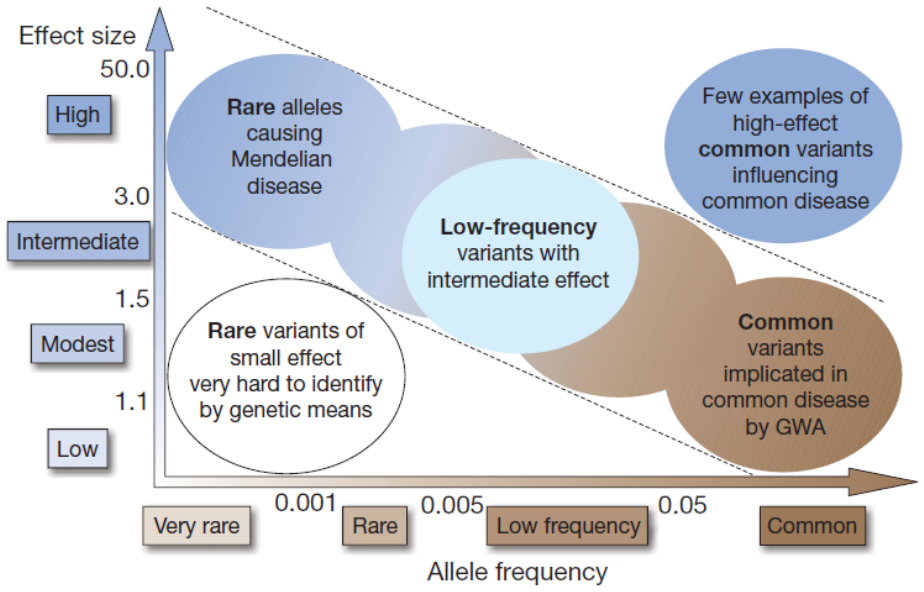

Figure 1.4 Genetic variants classified by allele frequency and effect size. The most investigated variants are in the parallel diagonal lines.

Rare mutations also make a major contribution to disease risk for individual people. However, rare variants cannot be detected in GWAS, due to the limited sample size of the

allele. For example, if we want to detect a rare variant with allele frequency 10^{-5} and with effect size one standard deviation unit, a sample size of more than one million would be required[45]. Such large sample sizes are hard to achieve. Furthermore, current GWAS genotyping arrays do not capture rare variants. Targeted sequencing, whole genome sequencing and family-based studies will help to uncover causal rare variants. Another hypothesis is called "synthetic association"[46]. When rare variants are in linkage disequilibrium with a GWAS tagging SNP, a "synthetic association" between the common variant and the phenotype can arise, even though it is not causal. In this case, the effect size is usually underestimated, and the fact that different rare variants can be responsible in different cases is missed. Simulation studies shows that synthetic association is inevitable, and real cases such as hearing loss show how rare variants can be synthetically associated with common GWAS identified variants[46].

Indeed, some rare variants with very large effect sizes are thought to cause diseases. There are 350 million people worldwide who suffer from rare diseases, and about 7,000 diseases are defined as rare in the United States. Most of the rare diseases are genetic, caused by a mutation in single gene, and thus are present at birth. 30% of children with rare diseases die before the age of five. Even though each single disease affects a small number of people, collectively the burden on individual patients, families and socioeconomics is tremendous. To raise public awareness for rare disease, the last Friday of February is observed as rare disease day. In addition, there are a large number, perhaps the majority, of rare variants with odds ratios at least 10 that nevertheless have low penetrance. A ten-fold increase in risk for a disease present in 1 percent of the population still only affects 10% of carriers. Mapping rare causal variants of this type, as well as Mendelian ones, is a major objective of contemporary genetics.

With the advent of next generation sequencing, the identification of causal genes, and description of mutation patterns affecting them, has progressed remarkably. Table 1.1[47] summarizes eight rare neuropsychiatric conditions caused by rare copy number variation.

Table 1.1 Disease association with CNVs (modified from Manolio et. al., 2009[47])

Disease	Locus	Type of CNV	Size (kb)	Population frequency	Case frequency	Effect size (OR)
Autism/IMR	16p11.2	*De novo* deletion	600	1×10^{-4}	1%	100
Autism	16p11.2	Rare duplication	600	3×10^{-4}	0.50%	16
Schizophrenia	1q21.1	Rare deletion	1,400	2×10^{-4}	0.30%	15
IMR	1q21.1	Rare deletion	1,400	2×10^{-4}	0.47%	NA
Schizophrenia	15q13.3	Rare deletion	1,600	2×10^{-4}	0.20%	12
Epilepsy	15q13.3	Rare deletion	1,600	2×10^{-4}	1.0%	NA
IMR	15q13.3	Rare deletion	1,600	2×10^{-4}	0.30%	NA
Schizophrenia	22q11.2	Rare deletion	3,000	2.5×10^{-4}	1%	40

Apart from conventional diagnosis, molecular diagnosis has been developed with the implementation of next generation sequencing based approaches, whole-exome sequencing, whole-genome sequencing and targeted sequencing. Diagnostic yield for rare diseases is typically around 40% and it is important to achieve a higher diagnostic yield in infants. Whole exome sequencing diagnostic yield can be improved from 27% to 40% of children with developmental disorders[48]. Rapid whole genome sequencing, rWGS, improved diagnostic sensitivity to 43%, compared with 10% for standard genetic testing[49]. Ultra-rapid whole genome sequencing, urWGS, yields 46% diagnostic rate in infants with diseases of unknown etiology[50]. As a complementary molecular diagnostic tool, RNA sequencing shows success in

diseases that whole-exome sequencing fails, or improves diagnostics. For example, RNA-seq identifies splice-altering variants in both exonic and deep intronic regions, resulting in an overall diagnostic rate of 35% in a cohort of 50 patients with previously genetically undiagnosed rare muscle disorders[51]. Another study performed RNA sequencing of whole blood, which is the most easily accessible tissue. The cohort compromised 94 individuals with undiagnosed rare diseases spanning 16 disease categories. By analyzing gene expression levels and abnormal splicing, and combining these measures with variant identification, RNA-seq yielded a 7.5% diagnostic rate and 16.7% diagnostic improvement[52].

1.6 Thesis structure

1.6.1 Specific aims: comprehensive identification of causal disease associated risk variants

Over the past decade, tens of thousands disease/phenotype associated genetic risk loci have been identified by genome-wide association studies, but follow-up functional validation lags far behind. The search for causal regulatory elements has become an important challenge for precise identification of causal risk loci. It is aided by the facts that most risk variants are located in noncoding regions, gene expression is a simple phenotype that is easy to measure in the lab, and regulatory variants often explain more variation than variants associated with visible traits. In this thesis, my primary goal was to develop a CRISPR/Cas9 based screening method to identify causal regulatory single nucleotides using RNA sequencing as the output measurement. I primarily focused on genes and SNPs that are also associated with autoimmune diseases since they are likely to be active in immune cell types sampled in blood or represented by human cell lines. Another objective was to work toward extending the strategy from common variants to rare variants, by

developing a rare variant pathogenic probability score. Such score should help to prioritize rare variants for experimental screening and facilitate personalized genome medicine.

1.6.2 Aim 1: Establish a single-cell clone CRISPR/Cas9 based eQTL screening method

Fine mapping refines statistical confidence intervals containing 100 and more SNPs to causal variants by integrating functional annotation. Two challenges are that it is hard to pinpoint one causal variant due to complex haplotype structure, and at least one third of all loci harbor multiple independent regulatory association signals. Therefore, experimental assays are needed to discriminate the effective SNP from the confounding remainder. In pursuit of this aim, in Chapter 2 I describe an NHEJ-based CRISPR/Cas9 single-cell clone assay for introducing microindels covering SNPs affecting *CISD1*, *SDCCAG3*, *NFXL1* and *AMFR* expression. First, I characterized the transcriptomes of single-cell derived clones and neutrophil-, monocyte-differentiated HL60/S4 cells by RNA sequencing, quantifying biological variation among single-cell clones before and after induction of differentiation. Second, I tested the feasibility of the single-cell clone strategy by obtaining single-cell CRISPRed clones with introduced indels. Finally, I used the observed variance components to perform a statistical power computation under different simulated scenarios. After careful discussion of four major constraints, I concluded that this method is not recommended.

1.6.3 Aim 2: Establish massively parallel eQTL screening with single-cell RNA-seq as readout

Subsequently, Aim 2 described in Chapter 3 switches the strategy to a single-cell RNA-seq based CRISPR/Cas9 pooling strategy, called expression CROP-seq. Single-cell

RNA-seq robustly sequences thousands of cells in parallel, which reduces technical variation and captures individual cell variability at the same time. Infecting cells with gRNA library accomplishes moderate- to high-throughput genetic screening. Together these methods resolved the problems from aim 1, leading to report of a feasible method for moderate-throughput credible interval fine mapping I aimed to screen 57 eSNPs along with 10 control, where each SNP was targeted by one gRNA. Approximately, 20 SNPs each were from credible intervals of *DAP*, *CISD1* and *PARK7,* which have colocalized eQTL and IBD GWAS signals. I characterized the transcriptional profiles of thousands of single gRNA assigned cells in two biological replicates. The experiment was design to assign about 100 cells to each gRNA, and gRNAs were evenly distributed among engineered cells, except for those targeting two essential positive control genes. Two SNPs, rs2251039 and rs35675666, significantly reduced expression of the target genes *CISD1* and *PARK7*, respectivelyin both replicates. Since ATAC-seq and DNase-seq peaks overlap with the two SNPs, and individual targeting confirmed the results, the proposed moderate-thought method is suitable to be used in screening of genetic variants with at least moderate effect size within credible intervals.

1.6.4 Aim 3: Classification of the tolerance of promoter regions with the burden of rare variants across tissues

Rare variants are abundant in individuals. From an evolutionary perspective, large-effect disease risk alleles should be rare under the pressure of purifying selection. A corollary is that ultra rare alleles often have large effects causing disease in individual. Thus, it is of high interest to extend the application of expression CROP-seq to rare regulatory variant screening. In Chapter 4, I describe an approach that comprehensively

assigns rare variants a probability score that prioritizes which rare alleles are more likely to be pathogenic in a specific tissue. I tested whether the enrichment of rare variants in extreme expression individuals is common across 49 tissues with GTEx v8 data, and whether those rare alleles are depleted in genes that are intolerant to promoter region mutations. These results indicate that combining the relative promoter polymorphism and regulatory rare alleles accounting for large proportion of the variation in aberrant expression can be used for a gene categorization system.

In summary, the scope of this thesis is to provide both experimental and computational approaches to fine-map common and rare genetic variants accounting for medium and large effect to the population or individual. This reduces to massively parallel screening of regulatory SNPs within credible intervals and to pinpointing regulatory rare alleles with evolutionary intolerance of promoter region mutations across tissues.

CHAPTER 2. SINGLE CLONE CRISPR/CAS9 MUTAGENESIS TO FINE-MAP REGULATORY INTERVALS

ABSTRACT: The majority of genetic variants affecting complex traits map to regulatory regions of genes, and typically lie in credible intervals of 100 or more SNPs. Fine mapping of causal variant(s) at a locus depends on assays that are able to discriminate effects of polymorphisms or mutations on gene expression. Here we evaluate a moderate-throughput CRISPR/Cas9 mutagenesis approach based on replicated measurement of transcript abundance in single cell clones, by deleting candidate regulatory SNPs affecting 4 genes known to be affected by large-effect eQTL in leukocytes and using Fluidigm qRT-PCR to monitor gene expression in HL60 pro-myeloid human cells. We conclude that there are multiple constraints that render the approach generally infeasible for fine mapping. These include non-targetability of many regulatory SNPs, clonal variability of single cell derivatives, and expense. Power calculations based on the measured variance attributable to major sources of experimental error indicate that typical eQTL explaining 10% of the variation in expression of a gene would usually require at least 8 biological replicates of each clone. The above is published in *Genes*. This work was performed in collaboration with the group of Dr. Gang Bao, supervised by Dr. Ciaran Lee and with wet lab experiments performed by Dr. Yidan Pan at Rice University. My contribution was mostly the statistical analysis and writing of the paper, cell culture experiments and RNA-seq library preparation performed by me at Georgia Tech, along with input into experimental design.

2.1 Introduction

Genome-wide association studies (GWAS) over the past decade have been highly successful in identifying tens of thousands of loci influencing disease risk[45,53,54], but the fine mapping of causal variants has failed to keep pace. Exhaustive studies of Crohn's disease and Type 2 diabetes associations, for example, indicate that the average credible interval size for hundreds of loci remains over 100 SNPs, and fewer than 15% of the loci have been reduced to a single high-confidence causal polymorphism[55,56]. This gap in knowledge impedes both the understanding of the biological functions of risk loci, and the progress in clinical genetic risk assessment. There are three main challenges to fine mapping. First, the haplotype structure of the human genome ensures that multiple SNPs lie in high linkage disequilibrium (LD) with the peak association signal, so that it is rarely possible to promote one variant as causal on statistical evidence alone. Second, it is now clear that at least one third of loci harbor multiple independent associations, most with overlapping credible intervals[55-57]. Third, the majority of the risk loci are located in non-coding regions of genes[58,59] where they exert their function through regulation of gene expression. Tools for predicting the function of such causal variants generally have low predictive value[60,61].

Moderate-to-high throughput methods are needed to prioritize likely causal variants by experimentally monitoring their effects on gene expression[62]. Two broad classes of approaches have been described: massively parallel reporter assays and genome editing. Massively parallel reporter assays couple short segments of potentially regulatory DNA to guide barcodes which are transcribed following transfection into cells or animals. Sequencing approaches allow identification of under- or over-represented barcodes

indicating differential expression due for example to polymorphisms. Genome editing approaches now most commonly use CRISPR/Cas9 to introduce short insertions, deletions and substitutions into targetable regions across the whole genome. RNA sequencing or other functional readouts such as fluorescence of a reporter gene, can be used to monitor the impact of specific variants. Recent CRISPRi and CRISPRa pooled screening assays utilize 52 catalytically dead/inactivated Cas9 enzymes (dCas9) that bind to but do not cut the target site. These 53 modified Cas9s have their endonuclease activity removed, but they are still able to bind to the target 54 sites where they contribute to inhibition or activation of gene expression via fused effector domains 55 such as KRAB (CRISPRi) and VP64 (CRISPRa). They have enabled high-throughput screening of genomic elements influencing transcription[38] and cellular phenotypes[39-42] with single-cell transcriptome readout. However, the majority of these strategies screen regulatory intervals rather than individual SNPs, so are not appropriate for fine mapping causal variants.

Here we evaluate the feasibility of gene-centric single cell clonal analysis, focusing on a handful of genes known to influence the risk of inflammatory bowel disease (IBD) through modulation of gene expression in immune cells. Specifically, we chose to examine four genes with evidence for two independent *cis*-eQTL intervals each, as well as GWAS-significant associations with IBD. The CDGSH iron sulfur domain 1, *CISD1*, and serologically defined colon cancer antigen 3, *SDCCAG3*, genes are associated with both ulcerative colitis and Crohn's disease[63,64]. The Autocrine Motility Factor Receptor, *AMFR*, encodes a glycosylated transmembrane receptor that is also an E3 ubiquitin ligase, knockdown of which in the acute monocytic leukemia cell line, THP-1, induces cell cycle

arrest and apoptosis, indicating a critical role for *AMFR* in cell proliferation[65]. *NFXL1* is one of the most up-regulated genes in IL-4 induced macrophages[66].

We used an experimental strategy for targeted SNP evaluation wherein microdeletions targeting candidate eSNPs are introduced by CRISPR/Cas9, and then isolated as single-cell clones on a uniform genetic background. Although homology-directed repair (HDR) would provide more precise evaluation of allelic replacement, the low efficiency relative to non-homologous end joining (NHEJ) and expectation that indels may have larger effects led us to use NHEJ in these experiments I chose the HL60 cell line, a pro-myelocytic lineage, which can be induced to undergo differentiation toward neutrophil- or monocyte-like fate, allowing evaluation of SNP effects in different cell types. Given the challenges in demonstrating conclusively the impacts of a single causal variant, we discuss sources of experimental variance encountered with this strategy, including batch, clonal and differentiation effects, and use these to derive realistic power estimates for dissection of causal variants. Comparing these estimates with empirically defined eQTL effect sizes, we conclude that this approach is generally incapable of resolving most regulatory associations to single causal variants.

2.2 Materials and methods

2.2.1 eGenes, candidate eSNPs and control SNP selection

The eGenes *CISD1* and *SDCCAG3* were chosen due to the colocalization of eQTL signals and 85 associations with inflammatory bowel disease[67]. *NFXL1* and *AMFR* were included as they are essential for myeloid cell differentiation. Candidate eSNPs were selected from one of at least two independent eQTL credible intervals at each locus

identified in a multiple eQTL study using stepwise conditional regression[57] in two large peripheral blood microarray datasets, the Consortium for the Architecture of Gene Expression (CAGE)[9] and Framingham Heart Study (FHS). They were also confirmed to be eQTL in monocytes[68]. It remains possible that they are not actually active in HL60 cells or their derivatives, and our experiments should be interpreted with this in mind. We also evaluated each SNP in the credible interval with Combined Annotation Dependent Depletion (CADD) score[69] and evolutionary probability (EP)[70]. In each credible interval, we chose the SNP with the lowest p-value, named as "Top SNP", SNPs with low evolutionary probabilities (EP) of the minor allele and (or) high CADD scores, named as "Both" and "High CADD", respectively (Table 2.1). We also picked SNPs as negative controls with no eQTL signals and in linkage equilibrium with the top SNP, named as "Control". Conditional eQTL profiles can be visualized using our eQTL Hub shiny browser at http://bloodqtlshiny.biosci.gatech.edu/.

Table 2.1 Guide RNAs and target SNPs. Guide RNAs and target SNPs. Each guide RNA targets on the "SNP", which is within a credible set of "gene". The effect size (z-score) of each SNP from the eQTLGen browser[67]. "Top SNP" is the SNP with lowest p-value in the credible set. Several criteria was used to predict the likelihood of candidate SNPs: "High CADD" is the SNP with high CADD (Combined Annotation Dependent Depletion) score that has high level of deleteriousness of its variants, including Indel variants; "Top" is the SNP with strongest signal of eQTL-mapping; "Both" is the SNP with both high CADD score and low evolutionary probabilities (EP) of the minor allele; "Control" is the negative control SNP in high linkage disequilibrium (LD) with the top SNP but low CADD and normal EP.

gRNA	Gene	Top SNP	SNP	z-score	Type	Genome location	Coding region
RG14	*SDCCAG3*	rs10870171	rs3812594	-34.60	High CADD	Exon of *SEC16A*	Yes
RG16	*CISD1*	rs4397793	rs4397793	-23.84	Top	Intron of *TFAM*	No
RG17	*CISD1*	rs4397793	rs648138	-70.54	Control	Intergenic of *TFAM*	No

RG19	*CISD1*	rs2590375	rs2590363	-100.37	Both	Intron of *IPMK*	No
RG20	*CISD1*	rs2590375	rs1416763	-100.27	Both	Intron of *CISD1*	No
RG26	*NFXL1*	rs116521751	rs321622	-63.35	Both	Intron of *NIPAL1*	No
RG34	*AMFR*	rs2550303	rs8060037	-14.09	Top	Intron of *NUDT21*	No

2.2.2 SNP-targeting and gRNA screening design

The chromosomal position of each candidate SNP in reference genome hg19 was obtained from dbSNP[71] by searching their RSID. The sequences flanking the targeted SNP were fetched from NCBI Reference Sequence (RefSeq), providing a gRNA screening window[72]. In each window, all the 19-base sequences followed by the correct *S. pyogenes* Cas9 Protospacer Adjacent Motif (PAM) sequence (NGG) were collected as candidate gRNAs. gRNAs with GC rate over 80% or less than 10% were filtered out to assure better cutting performance, and only the gRNAs with distance of cut site to targeted SNP not more than 10 nucleotides were selected for off-target effect analysis. The in silico predictions of their off-target effects were tested using COSMID[73]. The online tool is available through https://crispr.bme.gatech.edu/.

2.2.3 Single cell clone generation

HL60 (ATCC, Manassas, VA, USA, CCL-240) and HL60/S4 (ATCC, Manassas, VA, USA, CRL-3306) cells were grown in suspension at 2×10^5 to 1×10^6 cell/ml in RPMI-1640 with 10% FBS, 2 mM L-glutamine and 100 μg/ml normocin. After culturing for 18 hours to 24 hours, cells were pelleted at 200 g for 3 min. Used media was collected and filtered to obtain conditioned media. Bulk cells suspensions were serial diluted on a

96-well plate with conditioned media to facilitate cell growth. Statistically, there were wells that only had single cell. Alternatively, some single cell clones were generated by sorting bulk cells by flow cytometry on a BD FacsAria Fusion with 100-micron nozzle at 37°C, and seeded onto each well of a 96-well plate with the same conditioned media.

2.2.4 Myeloid lineage differentiation

Differentiation of cells into neutrophils was achieved by culturing with 1 μM retinoic acid, RA [74]. Cells were seeded 18 hours before treatment at 2×10^5 cell/ml. HL60 cells were treated for 4 days and HL60/S4 were treated for 2 days. During differentiation, cell density and viability were checked every 24 hours to maintain 2×10^5 to 1×10^6 cell/ml cell density. Additional culture media with RA was added if needed. Cells treated with the same volume of ethanol were used as negative control.

Differentiation of cells into monocytes was achieved by culturing with 100nM α1,25-dihydroxyvitamin D3(D3) dissolved in ethanol[75]. Cells were seeded at 1.5×10^5 cell/ml at least 18 hours before treatment. Both HL60 cells and HL60/S4 were treated for 3 days. During differentiation, alive cell density was checked and normalized every 24 hours to maintain 2.5×10^5/ml cell density. Additional culture media with D3 was added if required. Cells treated with the same volume of ethanol were used as negative controls.

2.2.5 Flow cytometry

After collection, cells were washed with PBS twice at room temperature. Cells under neutrophil differentiation were then incubated with 7-Aminoactinomycin D (7-AAD) (ThermoFisher Scientific, Waltham, MA, USA, cat. no. A1310) and PE-conjugated

mouse anti-Human CD11b (clone ICRF44) (BD Biosciences, San Jose, CA, USA, cat. no. 557321) or PE-conjugated isotype control mouse mAb (clone: MOPC-21) (Biolegend, San Diego, CA, USA, cat. no. 400112) for 40 min at 4°C in the dark. Samples were analyzed by BD FacsAria Fusion with 100 micron nozzle at 4°C. Cells under monocyte differentiation were incubated with V450 Mouse Anti-Human CD14 (BD Biosciences, San Jose, CA, USA, cat. no. 560349) and APC Mouse Anti-Human CD71 (BD, cat. no. 551374), or V450 Mouse IgG2b (BD, cat. no. 560374) and APC Mouse IgG1 (BD Biosciences, San Jose, CA, USA, cat. no. 555751) for isotype control. Samples were analyzed by BD FACSMelody at 4°C. All data were analyzed with FlowJo software v10.6.1 downloaded from https://www.flowjo.com/.

2.2.6 Immunofluorescence

After collection, cells were washed with PBS twice at room temperature. Then, cells were incubated with Hoechst-33342 (ThermoFisher, Waltham, MA, USA, cat. no. H3570) for 10 to 15 min at 37°C in the dark. 10 µl of the cell suspension was used to make a slide, which was sealed with clear nail polish. UV excitation and microscopic imaging was done on an Olympus IX73 inverted microscope system.

2.2.7 RNA isolation

Cells were grown in suspension at 2×10^5 to 1×10^6 cell/ml in RPMI-1640 with 10% FBS, 2 mM L-glutamine and 100 µg/ml normocin. Cells were seeded at 2×10^5 cell/ml 18 hours to 24 hours before extraction. Each clone had two biological replicates, except bulk HL60/S4. One million cells from each sample were collected by centrifuging at 300 g for 5 minutes. Total RNA was isolated and purified by RNeasy Plus Mini Kit (Qiagen, ,

Hilden, Germany, cat. nos. 74134 and 74136). Quality control of RNA samples were assessed with a Bioanalyzer 2100 instrument (Agilent, Santa Clara CA).

2.2.8 Bulk RNA-Seq and differential gene expression analysis

cDNA library preparation for single cell clones was done using Illumina TruSeq Stranded Sample Preparation, Low Sample (LS) Protocol. Sequencing was performed on an Illumina HiSeq 2500 at Georgia Tech, generating 100 bp paired-end libraries with an average of 51.8 million paired reads per sample. Library preparation for differentiated cells was done using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, MA, USA, cat. no. E7760S). Sequencing was performed on Illumina NextSeq, high output, generating 75 bp paired-end libraries with an average of 36 million paired reads per sample. The gene expression data is available at the Gene Expression Omnibus (GEO) under the accession code GSE135507.

RNA-Seq quality control was initiated with Trim Galore, which was used to trim the 13bp Illumina standard adapter ('AGATCGGAAGAGC') by default, after which quality control was reported by FastQC. Reads were mapped to hg38 human reference genome by STAR[76], and on average the mapped reads were 90% of total reads. Aligned sequencing reads were counted with intersection-strict mode in HTSeq[77] to get read counts for each gene. Scale factors of each sample were computed using the trimmed mean of M-value (TMM) algorithm in the R package, edge R[78]. Raw read counts were normalized by scale factors and then transformed into log2 counts per million reads (CPM). Genes were kept if expressed in at least three samples. 11,746 genes were kept in single cell clone RNA-Seq, while 13,485 genes were kept in differentiated cell RNA-Seq.

Differential gene expression analysis was conducted in edgeR with generalized linear models to contrast effects of each treatment group. Pairwise comparisons between control and neutrophil derivative, control and monocyte derivative, as well as within each clone of each type of cell were performed. Likelihood ratio tests was assessed to obtain lists of differentially expressed genes and following Benjamini-Hochberg false discovery rate correction.

Gene ontology analysis was performed using ToppFun[79]. By uploading a list of differentially expressed genes (FDR<0.001) from the differential gene expression analysis into the website, functional enrichment features were listed, including pathways, Gene Ontology (GO) terms and phenotypes. Gene ontology analysis was also performed by enrichR[80,81] with four sets of differentially expressed genes (FDR<0.001) uniquely in HL60 monocyte (968 genes), HL60/S4 monocytes (521 genes), HL60 neutrophils (1,462 genes) and HL60/S4 neutrophils (2,275 genes).

Principal Component Analysis (PCA) was done performed on 17 single cell clone samples, and 47 differentiated cell samples by "prcomp" function in R, with default settings. Principal Variance Component Analysis (PVCA) was performed in JMP Genomics 8 (SAS Institute, Cary, NC, USA), which sums the weighted proportions of each variance component associated with covariates of interest in order to estimate the overall contribution of biological and technical factors to the gene expression variation. Plots were plotted with R package, ggplot2.

2.2.9 Variant calling

Variants were called by GATK[82,83] best practice RNA-seq short variant discovery (SNPs and Indels). Raw RNA-seq reads was mapped to hg19 by STAR[76]. "SplitNCigarReads" was used to split reads that span introns and hard clip mismatching overhangs. Variants were called by "HaplotypeCaller" with default settings. Due to the high false positive rate of calling variants from RNA-seq data, the "VariantFiltration" function was used to filter potential false positive calls. Clusters of at least three SNPs within a window of 35 bases were excluded and calls with read depth lower than 50 were filtered. Moreover, the variant calls were only included if they were consistent in the two biological replicates of the same clone, and only exonic polymorphisms were counted.

2.2.10 Fluidigm qRT-PCR

Fluidigm Real-Time qPCR was conducted on a 48 × 48 nanoscale microfluidic chip with 48 EvaGreen probes targeting transcripts of the CRISPR targeted genes, as well as a representative set of lymphoid and myeloid cell marker genes[84], and housekeeping genes. The 48 array samples included single-cell clone CRISPR edited HL60/S4 from two batches and experimental controls. 2,304 qRT-PCR assays with 30 amplification cycles were conducted in parallel according to manufacturer's protocol. Average Ct value was computed at the exponential phase of each PCR amplification reaction. Since large Ct values correspond, counter-intuitively, to low expression, modified expression values were computed as the Ct values subtracted from 30 (the maximum number of PCR cycles) and the negative outputs were set as 0. This results in a range from null to 30 where each increment in theory represents a doubling of initial transcript abundance. To clean up the data, samples with more than 40 unexpressed genes and probes expressed in less than 5 samples were removed. Processed expression data and sample phenotypic

information are provided in online Tables S1 and S2 (https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter2_supp_tables), respectively. We note that numerous studies have established the high sensitivity of Fluidigm relative to standard qRT-PCR[85-87], and that all expression levels were in the normal range of detection and not subject to drop-out seen with very low abundance transcripts.

2.2.11 Plasmid construction

The SpyCas9 expressing plasmid pX330-U6-Chimeric_BB-CBh-hSpCas9[18] (Addgene plasmid #42230) was a gift from Dr. Feng Zhang. The pX330 vector was digested by BbsI. For each designed gRNA sequence, a pair of annealed oligos was cloned into the vector before the gRNA scaffold and after the U6 promoter. All clones were validated by Sanger sequencing (Eurofins Genomics, Louisville, KY, USA).

2.2.12 CRISPR-edited single cell clone generation

A total of 2×10^5 HL60/S4 clone 3 cells and 1 μg of pX330 plasmid per nucleofection reaction (program CA-137, solution SF) were electroporated using the Lonza Nucleofector 4-D based on the manufacturer's protocol. 1 μg of pmaxGFP™ Vector per nucleofection reaction was co-transfected as the reporter. The cells were cultured at 37°C for 72 hours after nucleofection, and the GFP positive cells were sorted individually by BD FACSMelody to make single-cell clones following standard protocols. Post-sorting, cells were grown for a week before harvesting and DNA extraction. DNA was extracted using Quick-DNA Miniprep Plus Kit (Zymo Research, , Irvine, CA, USA, cat. no. D3024) following the manufacturer's protocol. For each target

locus, a PCR product was amplified from the genomic DNA of cells modified by CRISPR/Cas9 and analyzed by Sanger Sequencing (Eurofins Genomics, Louisville, KY, USA). The genotypes of clones selected in this study are shown in online Table S3a (https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter2_supp_tables), and the number of clones screened, and mutations observed per clone are shown in online Table S3b (https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter2_supp_tables).

2.2.13 Power simulation studies

Power analysis was performed using the Mixed Model Power expression utility in JMP Genomics (SAS Institute, Cary, NC, USA). We created a design file with duplicates of 10 gRNAs, and designated one guide as the causal variant. Additional random effect options for representing Batch effects (distributing the guides across into two batches of 5), and Clone effects (where the causal variant was represented by two different clones) allowed modeling of the impact of these additional sources of variance. We assessed power at alpha = 0.05, 0.01, and 0.001 for effect sizes of the causal variant in increments of 0.1 standard deviation units (sdu) between 0 and 2, assuming experiments with 2, 4, 8 or 16 replicates of each guide. Batch and Clone effects were assumed to be 0.1 or 0.2 sdu. For an additional analysis, three of the guides were assumed to affect gene expression, modeling the situation where multiple linked variants account for an eQTL effect.

2.3 Results

2.3.1 Effect of Clonal Variability on Gene Expression in HL60 Cells

Since genetic screens are best performed in uniform genetic backgrounds under conditions where environmental variation can be carefully controlled, we started by evaluating the magnitude of effect of biological and technical factors on gene expression in HL60 cells. HL60 is a pro-myeloid cell line derived from a person with acute promyelocytic leukemia[88,89]. It is known to be homozygous for a *TP53* deletion and a *CDKN2A* premature stop codon, and heterozygous for an *NRAS* missense substitution. The main factors of interest were (i) batch effects, (ii) HL60 sub-type, (iii) clonal heterogeneity, and (iv) differentiation status. A derivative known as HL60/S4 has been isolated which is reported to more efficiently differentiate into myeloid derivatives such as neutrophils and macrophages[90]. Given the almost 40 years in culture, we reasoned that point mutations that are likely to affect overall gene expression may have accumulated, and to control for this isolated 3 single cell clones (labelled 1 through 3) of HL60, and 4 single cell clones (labelled a through d) of HL60/S4. Differences in growth rates among clones and relative to the bulk parental line were noted.

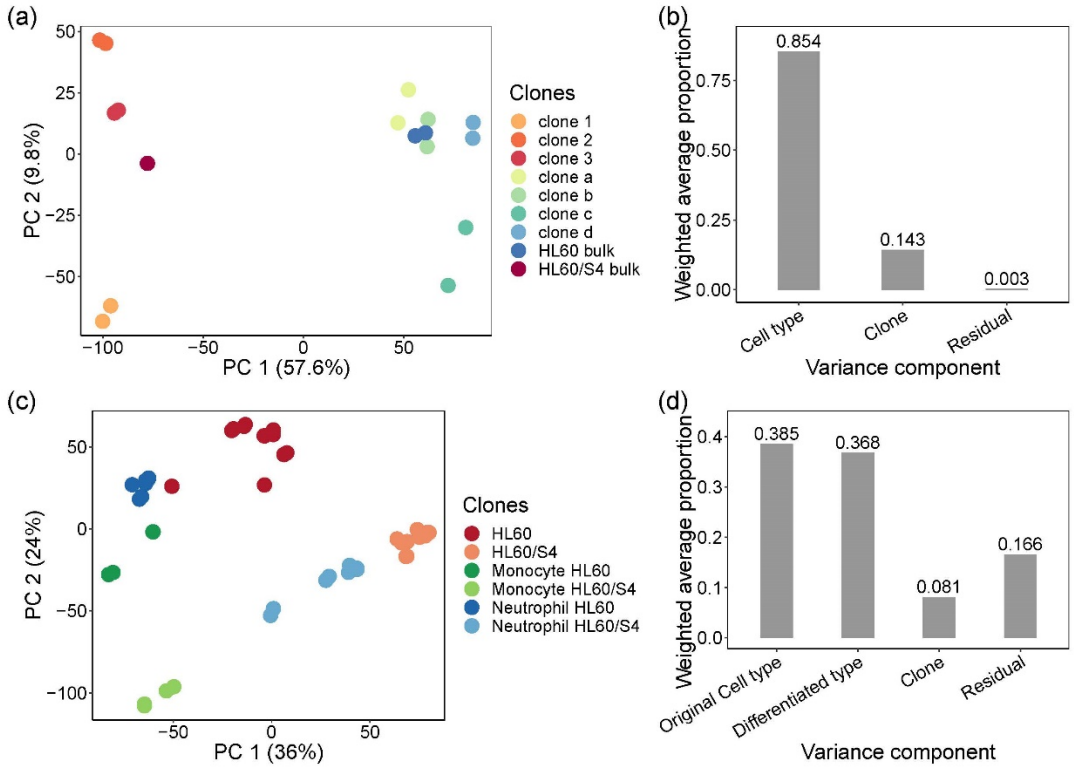

Figure 2.1 Heterogeneity of gene expression in single cell clones and myeloid lineage differentiated clones. (a) Principal component analysis (PCA) of bulk RNA sequencing of parental single cell clones and bulk cells. PCA was performed on normalized log2 CPM count expression matrix of 17 samples from HL60 and HL60/S4 generated single cell clones. Each dot represents 17 samples, two biological replicates for each clone and bulk, except for HL60/S4 bulk. Samples are colored by clones: warms color dots are samples from HL60/S4 cell lines, while cold color dots are samples from HL60 cell lines. PC1 separates samples by cell type, explaining 57.6% of the total variation. PC2 separate samples by clones, representing 9.8% of the total variation. (b) Principal variance component analysis shows the weighted average proportion of each variance component, cell type (85.4%), clone (14.3%) and residual (0.3%), all of which explain variance captured by the first five principal components (86.8% of total variance). Majority of the total expression variance of single cell clones is explained by cell type and clone variance components. (c) Principal component analysis of bulk RNA sequencing of myeloid lineage differentiated clones, performed by normalized log2 counts per million (CPM) expression matrix. Each dot represents 47 samples from differentiated monocytes and neutrophils and undifferentiated control cells, two biological replicates for each stimulation on each clone. Clone d is excluded due to sequencing error. Samples are colored by cell type and differentiation lineages: monocytes are green, neutrophils are blue and control cells are red. To distinguish the original cell type of each sample, HL60 cells are dark colors and HL60/S4 cells are light colors. (d) Principal variance component analysis shows the weighted average proportion of each variance component, original cell type (38.5%), differentiated type (36.8%), clone (8.1%) and residual (16.6%), all of which explain variance captured by the first five principal components (83.9% of total

variance). The 16.6% of unexplained variance may from the variance of biological replicates and culture differences between two labs.

Clonal variability in gene expression was monitored by bulk RNA-seq of two batches for each of the 7 single cell clones and 2 parental lines. Figure 2.1a plots the first two principal components (PC) of expression of 11,746 expressed genes detected with an average depth of over 50 million paired-end 100 bp reads per sample. PC1 separates the two HL60 sub-types unambiguously, and 85% of the variance attributable to the first 5 PC (86.8% of total variance) is between HL60 and HL60/S4 cells. Individual clones separate along PC2 with relatively little separation between replicates, with the parental lines taking intermediate values. Just 14% of the variance is among clones, but residual replicate effects account for less than 1% of it (Figure 2.1b). These results confirm that single cell clones are likely genetically differentiated, implying that as far as possible CRISPR/Cas9 editing should be performed on a purified clone.

The extent of genetic differentiation of single cell clones was evaluated by calling genotypes directly from the RNA-seq data. Given that false positive calls are elevated due to errors induced by the reverse transcriptase during cDNA preparation, and that allele-specific expression causes SNP ratios not observed in genomic DNA sequence data, we applied variant hard filtering in GATK. Clusters of at least three SNPs within a window of 35 bases were excluded, the variant calls were only included if they were consistent in the two biological replicates of the same clone, and only exonic polymorphisms were counted. On average, each of the HL60 single cell clones differed from the bulk consensus sequence at 103 of the 7482 single nucleotide variants (1.38%) passing our hard filters. A little over fifty percent more divergence, 166 of 7104 SNVs (2.34%) were uniquely observed in

HL60/S4 pairwise clonal comparisons with the bulk HL60/S4 consensus. Furthermore, approximately 3% of the total SNVs were different in the comparison of bulk HL60/S4 and HL60 lines and their derivatives, indicating that there is considerable genetic variability both between the two lines and in single cell clones. Similar findings were reported[91] in an analysis of somatic mutation accumulation in a cancer cell line.

Next, we asked how consistent chemical-induced differentiation is across clones. Each of the single cell clones, with the exception of HL60/S4 clone d, was treated with 1 μM retinoic acid for 4 days (HL60) or 2 days (HL60/S4) in order to generate neutrophil-like cells, or with 100nM α1,25-dihydroxyvitamin D3 for 3 days in order to generate monocyte-like cells. Figure 2.2 shows characteristics of the cells stained with Hoechst to monitor changes in morphology of the nucleus, 7-AAD to monitor cell viability, and CD11b, a neutrophil marker. Growth conditions were chosen to optimize the balance of cell differentiation and viability, which also varied among clones. As previously reported[90], HL60/S4 cells more readily differentiated toward neutrophil fate than did HL60 cells. Figure 2.3 confirms initiation of CD14 expression, as well as loss of CD71, both markers of monocyte fate, to similar degrees in both bulk HL60 and HL60/S4, though variation among clones of HL60 was also seen (online Table S4a,b, https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter2_supp_tables), including variability of cell surface marker expression at baseline.

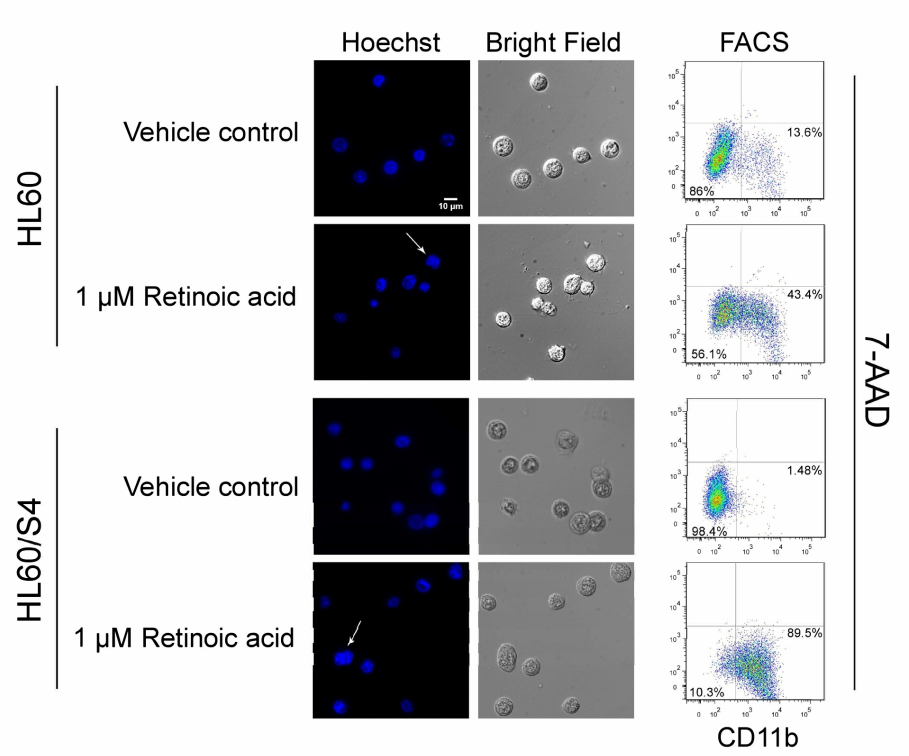

Figure 2.2 Characteristics of neutrophils by immunofluorescence and flow cytometry analysis. Left two panels show the Hoechst staining of multilobular nuclei of differentiated neutrophils. A representative figure of one clone of each cell type, HL60 and HL60/S4. The cell pointed by the arrow represents two lobes of the nucleus. The right panel shows flow cytometry analysis of CD11b (cell surface marker for neutrophil) expression and 7-AAD staining to stain dead cells. The gated populations with cell proportion show CD11b-/7-AAD+, live undifferentiated cells (bottom left quadrant) and CD11b+/7-AAD-, live differentiated cells (bottom right quadrant).

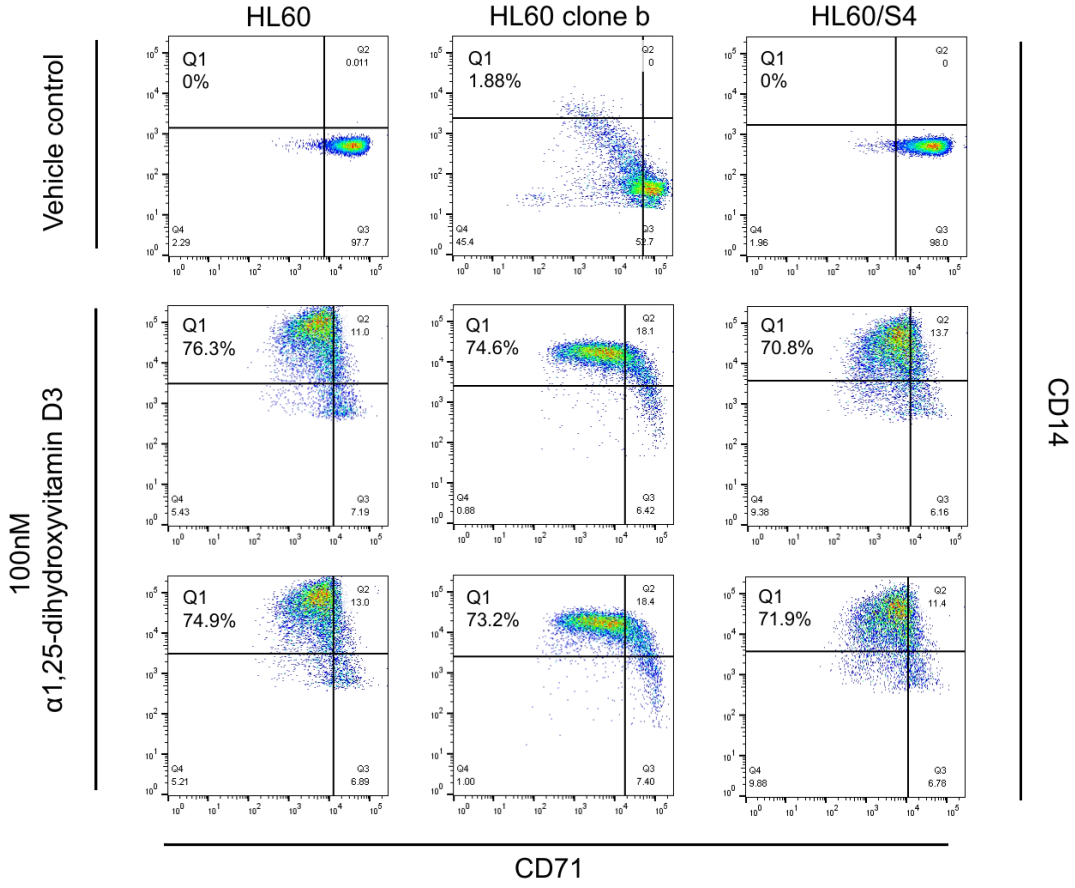

Figure 2.3 Characteristics of monocytes by flow cytometry analysis. The flow cytometry analysis of CD71 and CD14 expression in control and monocyte differentiated HL60 bulk, HL60 clone b and HL60/S4 bulk cells. The gated population (Q1) with cell proportion shows monocytes (CD71-/CD14+).

As with the untreated clones, gene expression was again observed to vary substantially between the two sub-types and among clones, with a generally uniform response to treatment and relatively small differences between replicates (Figure 2.1c). In a joint analysis, HL60/S4 cells tend to have more positive values of PC1 and negative values of PC2 than HL60, and the overall cell-type accounts for 38.5% of the variance captured by the first five PC (83.9% of total variance). Neutrophils occupy an intermediate position between monocytes and undifferentiated cells along both PC axes, and cell fate captures 36.8% of the variance. At baseline, HL60/S4 cells appear to be more divergent

from the derived neutrophil-like and, especially, monocyte-like cells than are HL60 from their derivatives. Clonal differences remain significantly higher than replicate effects.

In total, 5,885 and 3,319 genes (FDR<0.0001) were identified that were differentially expressed before and after monocyte and neutrophil lineage differentiation across all clones of two cell types, HL60 and HL60/S4, respectively.

After differentiation, HL60/S4-derived monocyte cells were more transcriptionally divergent from their parental cells than were HL60-derived monocytes: 7,381 monocytic differentially expressed genes were detected in HL60/S4, compared with 4,167 genes in HL60. *B2M*, a neutrophil-specific differentiation marker, is one of the 4,167 genes that were differentially expressed in the neutrophil-derived clone a, clone b, and HL60 bulk cells. There were 5,079 differentially expressed genes in the monocyte derivatives of HL60, including the transcription factors *CEBPE* specifically in clone c derivatives, and *PU.1* in clone b derivatives. Similar gene markers were also documented in a time course of myeloid differentiation[92], although we observed a higher number of differentially expressed genes at the terminal differentiated stage of monocytes than neutrophils, whereas the opposite pattern was found at 6 hour post-differentiation[92].

Differences in the degree of inter-clonal differentiation were also detected (Figure 2.4). For the monocyte derivatives, 1,781 genes were differentially expressed relative to undifferentiated cells in all of the clones of the two cell types, and these were enriched in cell cycle, neutrophil degranulation, and rRNA processing pathways. On the other hand, 968 genes were uniquely differentially expressed in the HL60 clonal comparisons, also showing enrichment for neutrophil degranulation and innate immune system pathways.

Gene ontology (GO) and pathway analysis was performed by Toppfun, and the significant GO terms and pathways (Bonferroni corrected p-value < 0.00001) for these 968 genes were listed in Figure 2.5. Similarly, for neutrophil lineage differentiation, 413 differentially expressed genes were shared by HL60 and HL60/S4, enriched for neutrophil degranulation, innate immune system activity, interleukin-10 signaling, chemokine signaling and cytokine signaling pathways. There were 1,462 and 2,275 clonal specific differentially expressed genes in HL60 clones and HL60/S4 clones, respectively, engaging pathways involved in cell cycle and mitochondrial function, translation and rRNA processing were also enriched. Significant GO terms and pathways (Bonferroni corrected p-value < 0.00001) for HL60 and HL60/S4 were shown in Figure 2.6 and 2.7, respectively. Gene ontology enrichment analysis of uniquely differentially expressed genes was also performed using the gene set enrichment tool Enrichr[80,81], with results summarized in Figure 2.8.

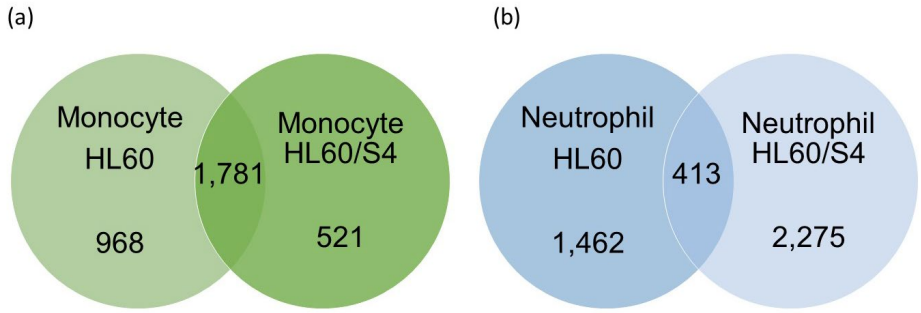

Figure 2.4 Venn diagram of differential expressed genes. The Venn diagrams show the number of significant inter-clonal differentially expressed genes (FDR<0.0001) in monocyte lineage (a) and neutrophil lineage (b) differentiation in HL60 and HL60/S4.

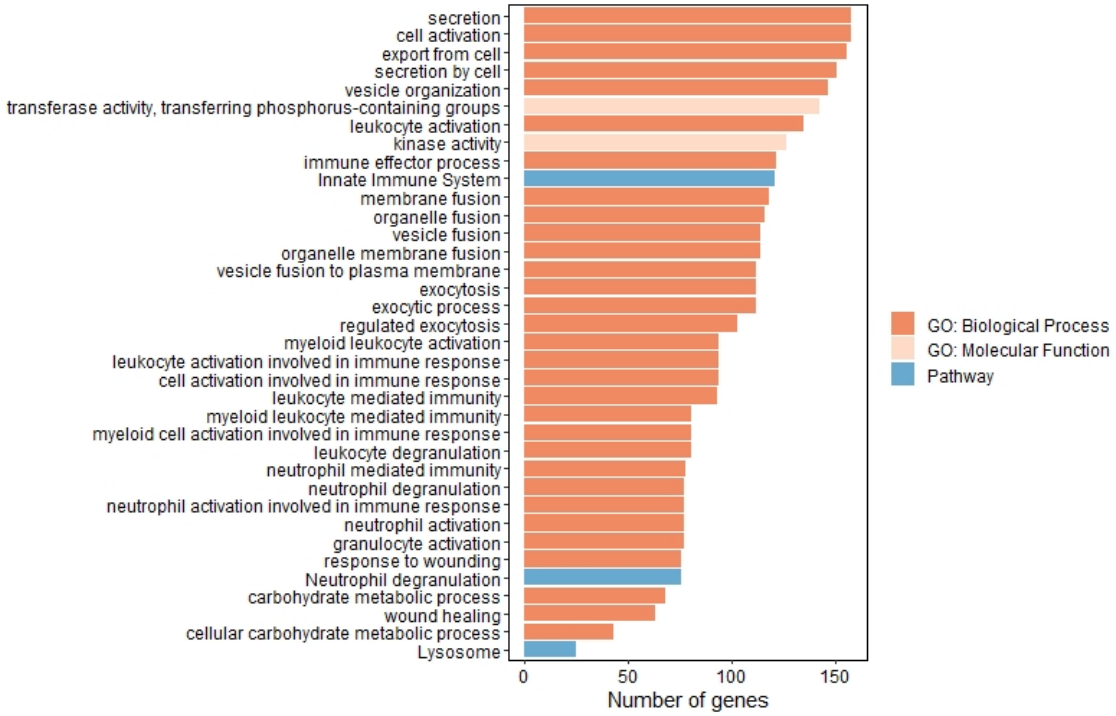

Figure 2.5 Gene ontology and pathway analysis of differential expression genes in HL60 monocyte derivatives. Uniquely differentially expressed genes in HL60 monocytes (968) were used as input to perform gene ontology and pathway analysis by Toppfun. Gene ontology terms involving in biological process and molecular function, and pathways are categorized and listed on the vertical x-axis. And the number of DE genes involved in each term is shown on the horizontal y-axis. P-value is corrected by Bonferroni correction with 0.00001 as cut-off.

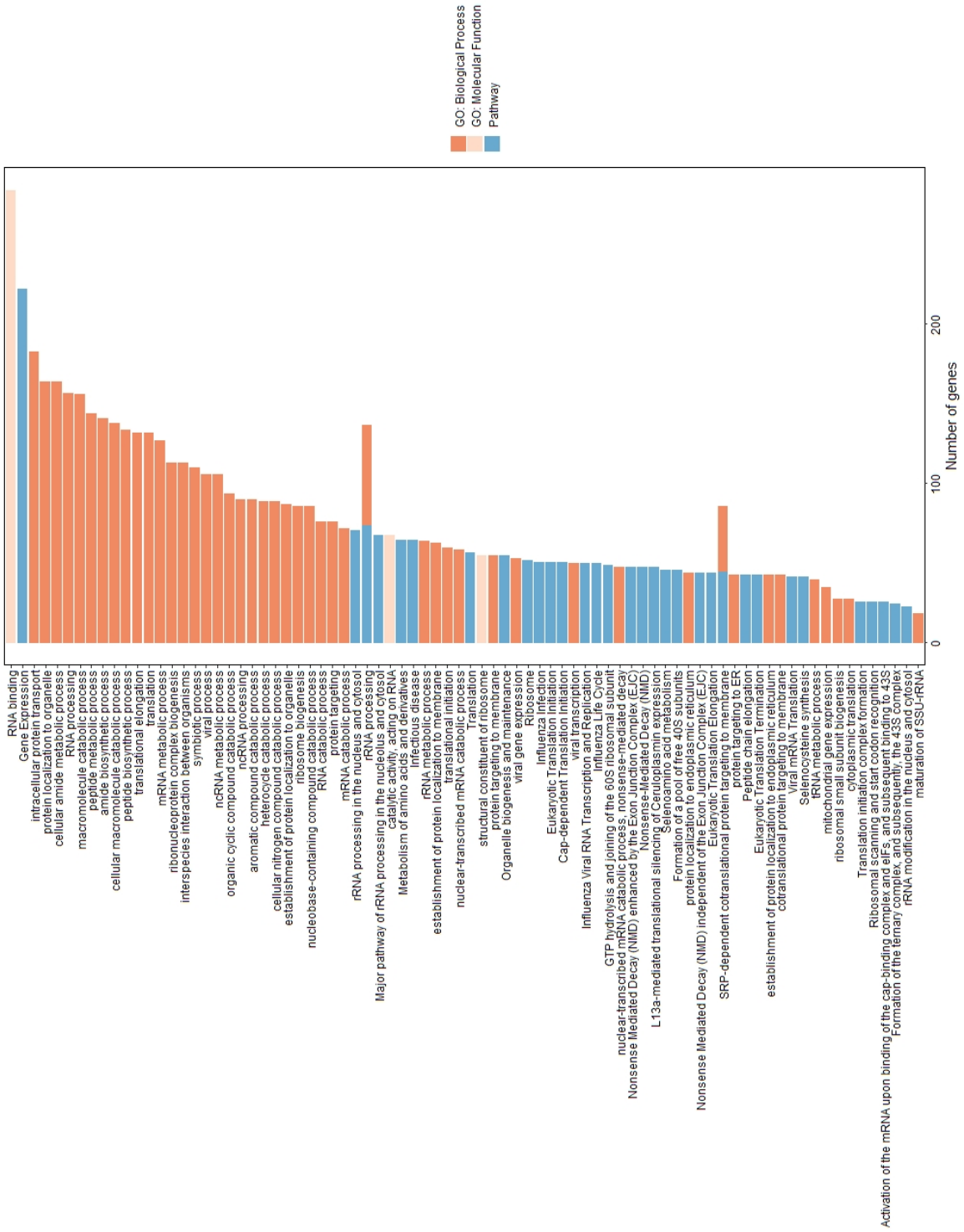

Figure 2.6 Gene ontology and pathway analysis of differential expression genes in HL60 neutrophil derivatives. Uniquely differentially expressed genes in HL60 neutrophil (1,462) were used as input to perform gene ontology and pathway analysis by Toppfun. Gene ontology terms involving in biological process and molecular function,

and pathways are categorized and listed on the vertical x-axis. And the number of DE genes involved in each term is shown on the horizontal y-axis. P-value is corrected by Bonferroni correction with 0.00001 as cut-off.

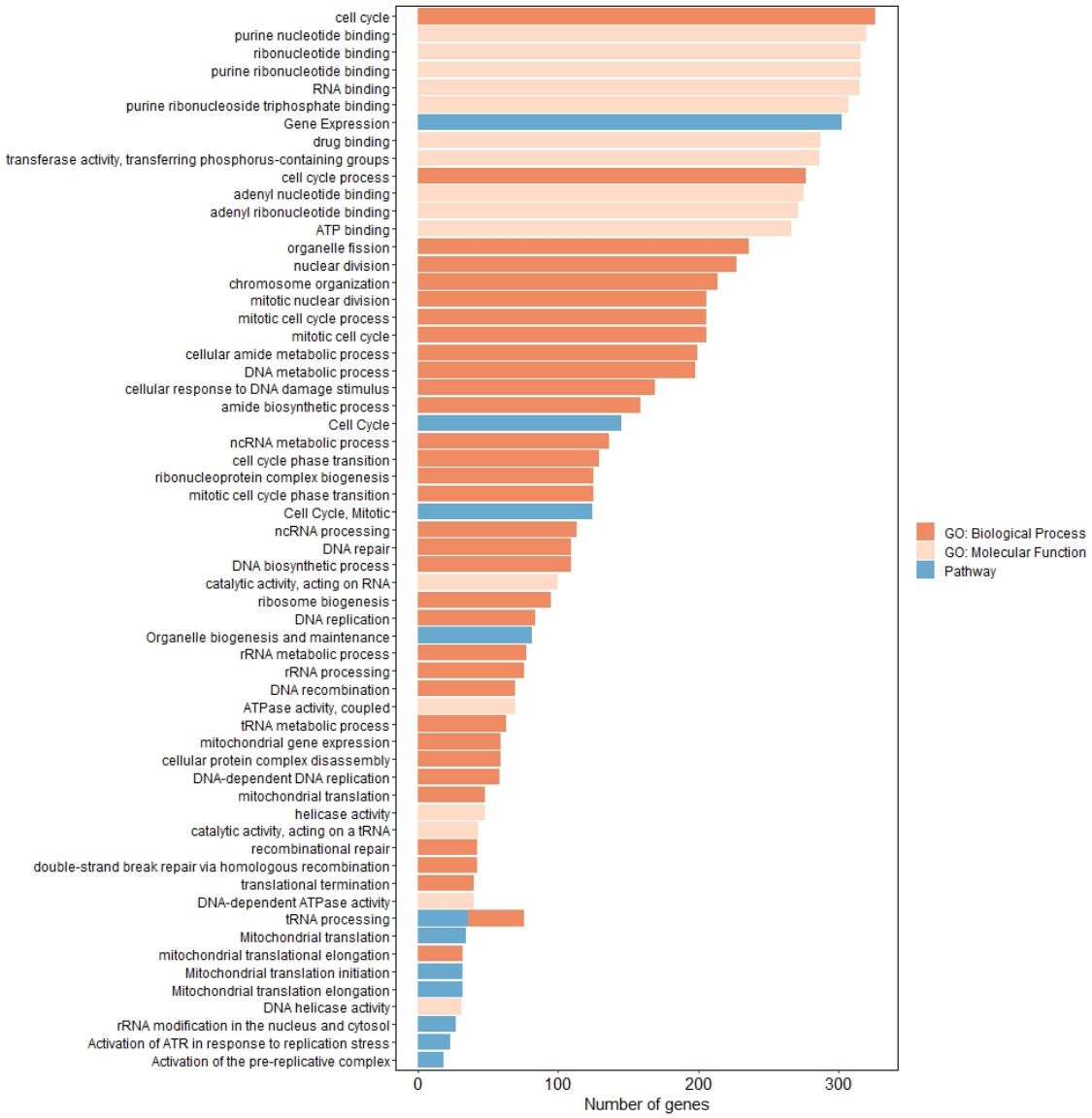

Figure 2.7 Gene ontology and pathway analysis of differential expression genes in HL60/S neutrophil derivatives. Uniquely differentially expressed genes in HL60/S4 neutrophil (2,275) were used as input to perform gene ontology and pathway analysis by Toppfun. Gene ontology terms involving in biological process and molecular function, and pathways are categorized and listed on the vertical x-axis. And the number of DE

genes involved in each term is shown on the horizontal y-axis. P-value is corrected by Bonferroni correction with 0.00001 as cut-off.

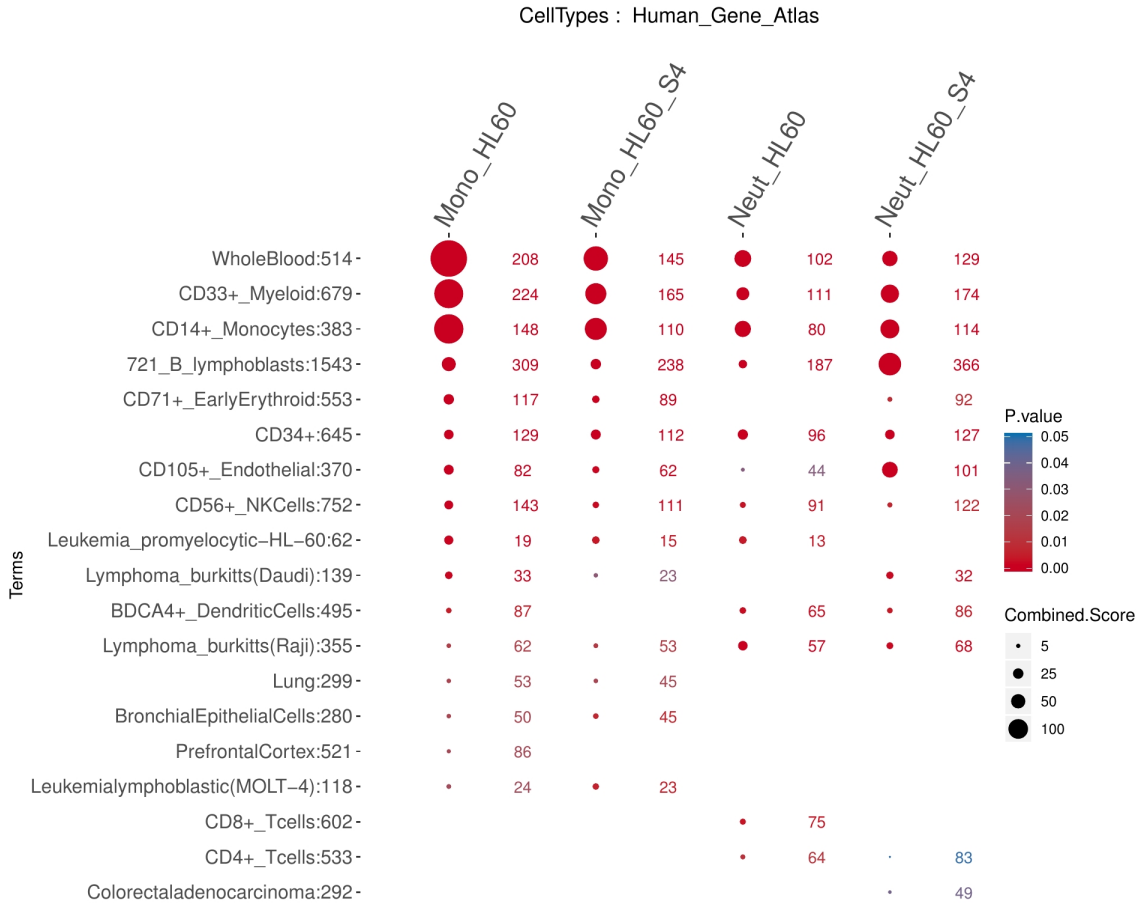

Figure 2.8 Gene ontology enrichment analysis of differential expression genes in HL60 and HL60/S4 monocyte and neutrophil derivatives. Uniquely differentially expressed genes in HL60 monocytes (968), HL60/S4 monocytes (521), HL60 neutrophils (1,462), HL60/S4 neutrophils (2,275) were used as input to perform gene ontology enrichment analysis by Enrichr. The total number of genes in each cell type term is shown on the y-axis and the number of DE genes enriched in each term is shown next to each dot. p-value and combined score (log(p-value) × z-score) are coded by the color and size of dots, respectively.

Taken together these results imply that single cell clones differ in basal gene expression, and although they respond similarly to treatment with retinoic acid or vitamin D3, clonal differences need to be accounted for when evaluating the effect of CRISPR/Cas9 mutagenesis of regulatory regions of target genes.

2.3.2 Isolation and evaluation of CRISPR/edited single cell clones

We selected 7 SNPs in 4 genes for our initial evaluation of the effect of non-homologous end-joining (NHEJ)-based CRISPR mutagenesis in HL60/S4 clone 3 as a uniform genetic background. *SDCCAG3*, *NFXL1*, and *AMFR* were each targeted for a single peak eQTL SNP detected by whole blood gene expression, whereas *CISD1* was targeted with 4 SNPs in one credible eQTL interval. Potential off-target sites analysis of each gRNA with up to two mismatches are provided in online Table S5 (https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter2_supp_tables). With genome wide bioinformatic screening, none of potential off-target sites were located in coding regions and the gRNAs have no extra perfect match other than the designed target site. Bulk transfection efficiency was 24.8% based on percentage of cells expressing GFP signal. GFP positive cells were considered capable to uptake plasmid vectors and were single cell sorted to enrich the edited cells. Of all expanded GFP positive single cell clones, 23 out of 166 had obtained Indels, 8 of which had removed the target SNP at both allelic copies, while the remainder affected sequences immediately adjacent to the target SNP or only had SNP removal in one allele.

RNA-seq would be prohibitively expensive for comparing gene expression on the scale of dozens of multiply replicated clones, so we next evaluated the potential of high

throughput nanoscale quantitative RT-PCR to detect subtle differences in transcript abundance. A 48×48 Fluidigm chip was designed, facilitating measurement of 48 genes (including the four targets, housekeeping controls, and various markers of expression in diverse immune cell type) in 48 samples. The HL60/S4 parental cell line and eight clones were chosen for profiling, one for each guide RNA, and each was grown in duplicate in suspension for 18-24hrs, with half the sample frozen down for storage, and the other half used for RNA preparation from fresh cells.

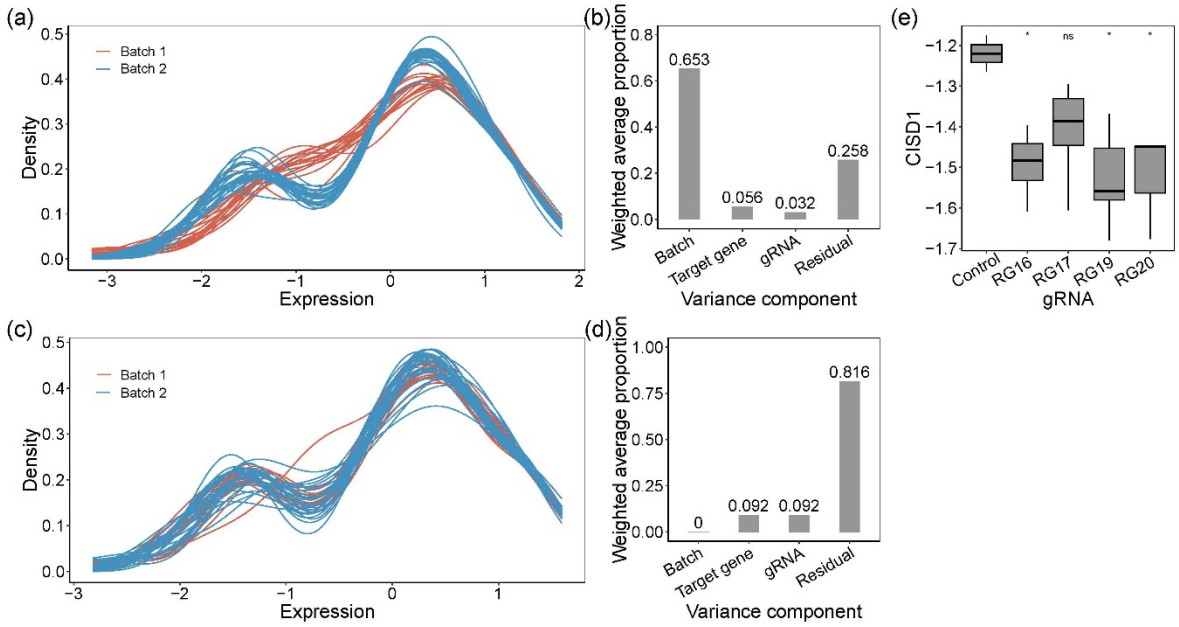

Figure 2.9 Quantification of gene expression by Fluidigm qRT-PCR and analysis of the variance components. Kernel density plot of standardized gene expression from each sample, color coded by batches, before (a) and after (c) removing batch effect. Before (b) and after (d) batch effect correction, principal variance component analysis showed the weighted average proportion of each variance component: batch 65.3%, 0%, respectively; target gene 5.6%, 9.2%, respectively; gRNA 3.2%, 9.2%, respectively; and residual 25.8%, 81.6%, respectively. All of the components explained variance captured by the first five principal components (99.1% and 99.1% of total variance, respectively). (e) Expression of CISD1. Pairwise t-test has done to test the difference between CRISPR/Cas9 edited samples (RG16, RG17, RG19 and RG20) and negative control. RG16, RG19 and RG20 were significantly different from the negative control. * denotes p-value<0.05; ns, not significant.

For ease of interpretation, we subtracted the Ct value for each measurement from the number of PCR cycles, 30, resulting in expression values where high values correspond to high expression. Figure 2.9a shows that this results in a bimodal distribution of gene expression measures, with the smaller peak representing low-abundance transcripts. There was a major difference in the profiles of the frozen and fresh cells, accounting for almost two thirds of the variance explained by the first five PCs (99.1%) (Figure 2.9b). To correct for this batch effect, we used Combat, which also standardizes the data to a mean of zero

and standard deviation of one (Figure 2.9c). On this scale, most of the variance is now among samples, whereas 9% of first five PCs (99.1%) distinguishes clones by which gene was targeted, and 9% is due to differences among gRNAs for *CISD1* (Figure 2.9d). This implies either that single gene knockouts affect the expression of a substantial number of other genes in each clone, or that there is substantial variability among clones that by chance correlates with the nature of the guide RNA. We also observed that normalized *CISD1* expression was lower in cells edited by each of the four gRNAs targeting *CISD1* than in the untreated control parental cell line (Figure 2.9e). Clone RG17 affected a control SNP in high LD with the peak eQTL but with low CADD score[69,93] and high evolutionary probability[70] of the alternate allele, and was the only clone not significantly different from the parental line. However, since it is unlikely that each of the other three sites causally influence gene expression, this result serves as a further caution that the process of transfection with CRISPR reagents itself may influence cell growth and gene activity.

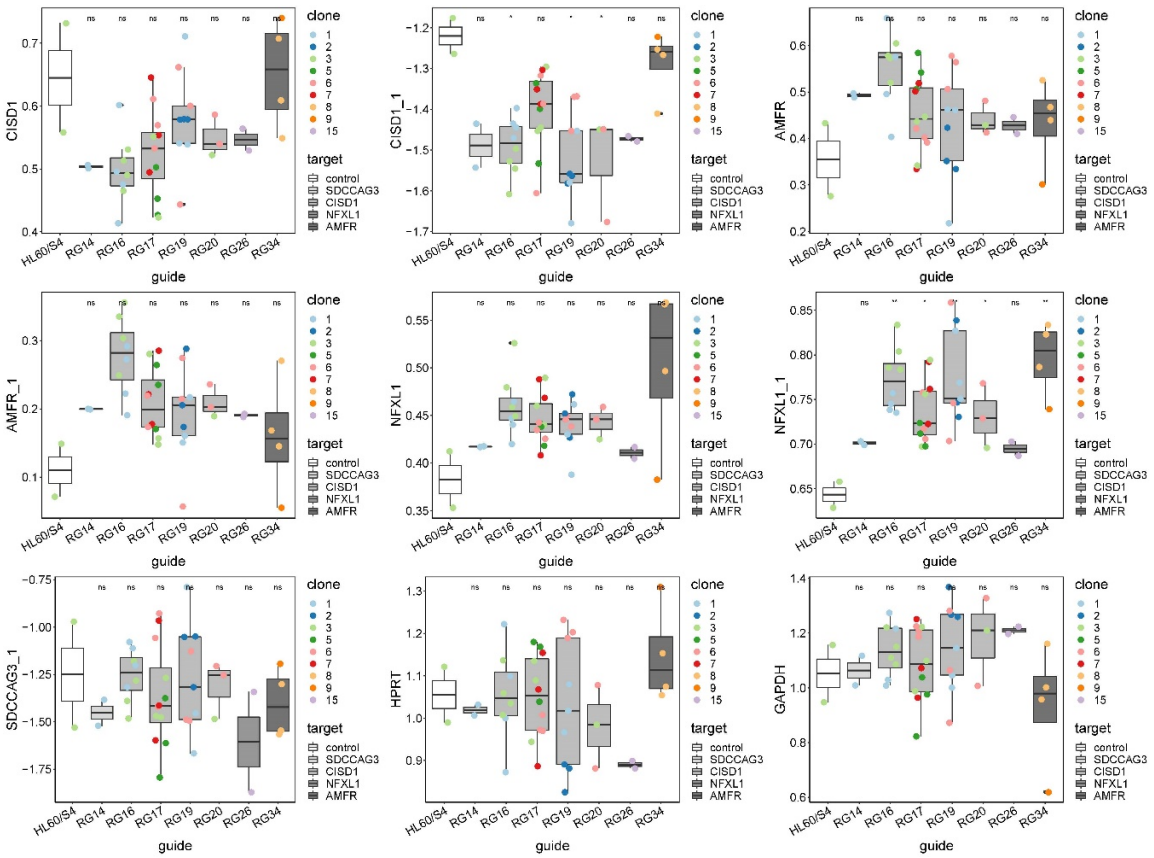

Figure 2.10 Quantification of all targeted gene expression in all CRISPR/Cas9 edited single cell clones by Fluidigm qRT-PCR (Table S4, S5). *HPRT* and *GAPDH* are housekeeping controls. Single cell clones were grouped by guide RNA and the expression of seven probes was shown as boxplot across all clones within each guide RNA group. Clones with the same genotype in each guide RNA group were colored coded. Pairwise t-test has done to test the difference between CRISPR/Cas9 edited clones and HL60/S4 negative control. * denotes p-value<0.05; ns, not significant.

Similarly, inconsistent results were obtained for the other three genes, as summarized in Figure 2.10. Each panel shows box-and-whisker plots for each of the 7 guide RNAs and control HL60/S4 cells, with the mean and interquartile range of 9 single cell clones measured with two different PCR probes for three of the genes and one for *SDCCAG3*. In no case is the expression the most extreme for the guide RNA corresponding to the linked gene. For example, *AMFR* expression was highest in cells carrying a mutation in the RG16 guide disrupting a candidate regulatory site in *CISD1*, whereas *AMFR*

expression itself was on average the closest to expression in the control cells. Disregarding the control, there were also no cases where the appropriate guide RNA was significantly different from the remaining guides. These results imply either that the selected SNPs are not causal, or that the effect sizes of causal variants are too small relative to the observed experimental variability to detect differential expression.

2.3.3 Simulation studies to establish power of Fluidigm-based single cell regulatory assessment

We used these results to guide our design and interpretation of power calculations for experiments designed to determine the effect of single regulatory site disruption. Our baseline scenario assumes targeting of 10 polymorphisms in a single credible interval in which a single eQTL is assumed to account for at least 10% of the variance in transcript abundance at the locus. Such an eQTL corresponds to a difference of approximately 1 standard deviation unit (sdu) in a quantitative assay such as Fluidigm qRT-PCR or RNA-seq. Given that most single cell CRISPR/edited clones are heterozygous, it also corresponds to a substitution effect whereby the mutant allele increases or decreases the measured transcript by 1 sdu. We used the Mixed Model Power calculator in JMP-Genomics (Cary, NC, USA) to evaluate the sample size needed to detect an effect of this magnitude given varying levels of clonal variation, batch effects, and mutation differences.

For the baseline scenario where there are neither batch nor clonal effects, 80% power to demonstrate that one SNP has an effect that is at least 1 sdu different from the other nine SNPs is achieved with 8 replicates of each of the ten clones (Figures 2.11a, f). 16 replicates would enable detection of an effect as small as 0.7 sdu, but 4 replicates would

only be powered to detect a substitution effect of 1.5 sdu. However, the experimental data indicates that individual clones generally do vary, as a consequence of genetic background effects if the transfected cell line was not isogenic, or growth differences among aliquots. Modelling these differences as a random effect of just 0.2 sdu among the ten clones demonstrates a dramatic reduction in power to detect the main effect (Figures 2.11b, g). With 8 replicates, only an effect size of 1.7 sdu is reliably detected, though 40% power is still obtained for an effect size of 1 sdu. Doubling the size of the experiment only slightly improves the power, whereas 4 replicates only facilitates detection of effect sizes of 2 sdu. If we further consider the scenario with a batch effect whereby half the clones have an additional random effect of 0.2 sdu (perhaps because they were grown at a different time), then power reduces yet again, as expected (Figures 2.11c, h).

A perhaps more realistic scenario is where different edits of the same polymorphic site also have different impacts on gene expression. This could either be because the precise nature of the deletion matters, or because the independent clones have slightly different growth properties. We modelled this scenario by allowing for two different clones representing the causal variant, also with a 0.2 sdu random effect difference, the same as the effect of the other 9 guide RNAs. In this case (Figures 2.11d, i), 80% power is never achieved, so it would take greater levels of replication at least of the putative causal variant to see a substitution effect in the range of 1 sdu.

A related situation is where more than one of the polymorphisms in the credible interval is responsible for the eQTL effect – for example, three sites in high LD might each account for 0.33 sdu, summing to a combined effect of 1 sdu. To model this, we set 3 of the guide RNAs to be causal, with the other 7 non-functional, but retained 0.2 sdu

differences among clones. Figures 2.11e, j show that power is greater than the same scenario with one causal variant, and approximately the same as with one causal variant and no differences among the remaining clones. Power is actually greater with fewer replicates (red and blue curves), but with 8 replicates 80% power still only detects an effect size of 1 sdu, which is three times larger than the presumed individual effect sizes of the contributing causal variants.

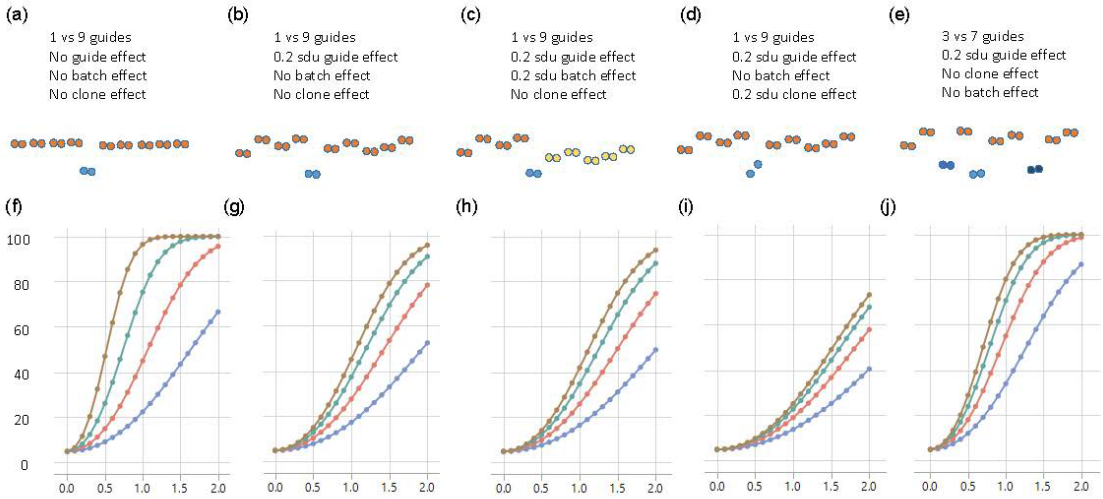

Figure 2.11 Power curves of Fluidigm-based single cell clone regulatory assessment of simulation studies. (a)-(e) diagrams five different scenarios, and the corresponding panels (f)-(j) show the power calculations for exceeding a nominal p-value of 0.05, with blue, red, green and brown curves representing 2, 4, 8 and 16 technical replicates of each clone respectively. The y-axis is the power from 0 to 100 percent, and the x-axis is the effect size of eQTL in standard deviation unit.

2.4 Discussion

Multiple studies have recently reported good success in mapping regulatory intervals using high throughput approaches in human cells. A previous study[37] scanned across over 100 kb of regulatory DNA in the *TP53* and *ESR1* genes using positive selection for

proliferation to enrich cells with aberrantly low expression of the target transcription factors, defining several intervals enriched for signals that overlap with transcription factor binding sites. This approach is however dependent on the ability to select on locus, and similar to methods that sort on the basis of an engineered selectable fluorescence protein[94], only identifies high-impact sites without necessarily discriminating effects of polymorphic sites. Another approach[95] used CRISPRa to map enhancer elements by virtue of activation of regulatory protein-DNA interactions, filtering a handful of short DNA stretches from hundreds of kb of intergenic sequence in the *IL-2RA* gene, but again without the ability to resolve which of the SNPs in a credible interval are responsible for an eQTL. Expression CROP-seq is powered to fine-map eSNPs with 10%-20% effect size within credible intervals by characterizing hundreds of CRISPR/Cas9 genetically mutated single-cell transcriptomes in parallel[96]. Tewhey et al,[31] first demonstrated the utility of massively parallel reporter assays, including the ability to discriminate between alleles at a site. Their results and findings from others[97,98] imply that at least 5% of all polymorphisms in regulatory DNA have the potential to regulate target gene expression. The concern remains though that such effects may be artefacts of short reporter genes assayed outside the context of chromatin and complex regulatory interactions.

Our approach instead borrows from classical quantitative genetic screens in model organisms such as drosophila and yeast. The objective is to create a panel of genetic perturbations in an isogenic background, evaluating the quantitative impact of each variant relative to the frequency distribution of effects of all other perturbations. For example, P-element insertion screens cleanly identified dozens of genes influencing aging, bristle number and aspects of fly behaviour[99,100]. Closer to our experiments, another study[101]

engineered a tiling path across the regulatory region of the *TDH3* gene in Saccharomyces cerevisiae and used flow cytometry to quantify gene expression of hundreds of strains, drawing inferences about the impact of stabilizing selection on transcription. We reasoned that a similar approach should be powerful for moderate-sized laboratories without extensive experience in human cell culture. Even though we, and others, have successfully documented regulatory effects of CRISPR/Cas9 mutagenized candidate mutations of large effect[55,102], the results here applied to typical moderate-effect size eQTL do not support this as a general protocol. The remainder of the discussion deals with multiple constraints on the effectiveness of single cell clone-based screening to dissect credible regulatory intervals in human cell lines.

The first constraint is variability in mutability of targeted regulatory sites. Our approach was mainly limited in three ways: the requirement of nearby PAM sequences and the short distance between the cut site and targeted SNP, the variable efficiency of different gRNAs, and the distinct Indel pattern for each SNP targeted gRNA. We started with a list of 250 candidate polymorphisms, approximately 10 each in two independent eQTL intervals of 13 genes, but discovered that only two-thirds of these were suitable CRISPR targets, either because there was no nearby PAM sequence or the target was in repetitive DNA for which it was not possible to design a guide RNA with a unique target sequence. Up to 20% of the remaining sites were predicted to have high probability off-target sites elsewhere in the genome, which may not matter for a scan of cis-acting effects but is not ideal. Subsequently, we chose 10 sites as a pilot, and screened an average of 24 single cell clones for each site (23.9 ± 6.7) by Sanger sequencing of the targeted region. As shown in online Table S1B

(https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter3_supp_tables), the pilot group had an average of 4 clones each with Indels on both alleles (3.8 ± 1.8). The ratio of clones with Indels on both alleles varied from 0% (RG11) to 25% (RG16) so that the theoretical maximum SNP removal rate were different in each gRNA treated group. RG14, 17, 19, 20 and 34 all had designed cut <5bp to the targeted SNP, but their percentage of SNP removal on both alleles varied from 0% to 16%, which could be due to variations in the size of Indel mutations as previously observed[103]. That is to say, many of the CRISPR-induced mutations removed or inserted one or a few nucleotides either side of the polymorphic site without disrupting the polymorphism itself. We conclude that obtaining at least 4 different clones for a minimum of 20 sites associated with a credible eQTL interval would typically require screening of 500 clones following various iterations of guide RNA design, with less than 100% success and at considerable expense. Allelic replacement by CRISPR-mediated homologous repair would be even more difficult. There are more potential optimizations that may help researchers deal with this constraint. Further optimization can be done in transfection, such as the co-transfection ratio of two plasmids. It is possible that different cell lines would have higher efficiency of mutagenesis. Other CRISPR/Cas9 delivery methods, such as lentivirus transduction, can also be beneficial for a more efficient screening.

The second constraint is clonal variability. We started by addressing a major concern with human cell lines, which is mutational accumulation in culture. Previous studies[91] showed that tumor cell lines diverge genetically in as few as a dozen passages, resulting in divergent drug responses and gene expression profiles. Accordingly, single cell cultures of HL60 and the derivative HL60/S4 cell lines are different at the DNA sequence level, and

have significantly different transcriptomes, both with and without chemical stimulation of differentiation. For a considerable proportion of genes, these differences are of a similar order of magnitude as expected eQTL effects, namely 20% to 50% differences in normalized abundance. While this observation strongly supports the decision to mutagenize a single cell clone, genetic differences may not actually be the major source of clonal variation. Mammalian, including human, cells are much more difficult to culture than yeast or bacteria, as thawed aliquots of frozen lines are well known to differ in growth rates and viability. The technical replicates in Figure 2.1 were all grown in parallel, so do not capture this type of batch effect, which we have not sought to quantify. However, we note that parallel culture of the nine mutant clones analysed was made difficult by variable growth rates, and that some thaws failed to grow at all, requiring expansion of new aliquots. Consequently, batch effects of single cell clones are a hidden but likely considerable source of gene expression variability.

A third constraint is expense. Assuming that the cost of RNA sequencing including cell culture, RNA preparation, library construction, and quality control could be reduced to $100 a sample using for example 3' tagging, an experiment with 8 replicates of 20 clones would still cost $16,000. Instead, we adopted a nanoscale quantitative RT-PCR approach, the 48×48 Fluidigm array. Each of the data points in Figure 2.10 is actually the average of 4 technical replicate qRT-PCR reactions on one plate at a cost of just $1.20 per assay (not including culture and RNA preparation). Technical repeatability is very high with repeated measures typically within 10%, also allowing measurement of dozens of genes simultaneously, so Fluidigm, or similar methods like Nanostring, provides a feasible approach in theory.

However, the fourth constraint, statistical power, emerged as the most serious impediment. A typical eQTL explains between 10% and 20% of the variance in expression of the gene it influences, which corresponds approximately to each allele increasing or decreasing transcript abundance between 0.5 and 1 standard deviation units. We modelled the power to detect such an effect in 80% of experiments given the variance components observed in our experiments and found that in the best-case scenario, 8 biological replicates would be needed to reliably detect a 1 sdu effect. However, addition of modest batch effects, subtle guide RNA differences within a locus, and small differences between different mutations induced by the same clone, power drops considerably. All such effects are apparent in Figure 2.10, suggesting that the single clone analyses, while demonstrably capable of discriminating very large regulatory effects of 2 or more sdu, are not generally likely to be detected with this approach. Cell lines other than HL60 may provide more repeatable results than we observed, which may improve power under some circumstances. It is possible that cell lines other than HL60 may provide more repeatable results than those described here, which may improve power under some circumstances. In this sense, independent valuation of the magnitude of batch effects for different cell lines under different growth conditions may be advisable, though we doubt that it will make single cell mutagenesis an optimal screening approach.

Finally, a fifth constraint is the assumption that each eQTL can be reduced to a single eSNP. This is the parsimonious assumption and fits readily with the conception that regulatory SNPs exert their effects by altering the binding affinity for a specific transcription factor. Even though most eQTL span 100 or more polymorphisms in a credible interval, the general assumption is that prioritizing variants according to functional

criteria and evolutionary conservation, using scores such as CADD or LINSIGHT, reduces the search space to fewer than ten candidates. However, given that these variants are in tight linkage disequilibrium with similar frequencies[61], if they have similar functional scores, then it is possible that the observed univariate eQTL effect is actually due to the summation of two or more smaller contributing effects. Under this scenario power to detect multiple causal variants is also reduced.

These considerations and the overwhelmingly negative results of our experiments lead us to the recommendation not to pursue single clone-based profiling as a general approach to fine mapping of regulatory variants. Despite the conceptual limitation that effects are evaluated outside the context of normal chromatin, massively parallel reporter assays seem to be more powerful and subject to less experimental constraint.

CHAPTER 3. FINE-MAPPING WITHIN EQTL CREDIBLE INTERVALS BY EXPRESSION CROP-SEQ

ABSTRACT: The majority of genome-wide association study (GWAS)-identified SNPs are located in noncoding regions of genes, and are likely to influence disease risk and phenotypes by affecting gene expression. Since credible intervals responsible for genome-wide associations typically consist of 100 or more variants with similar statistical support, experimental methods are needed to fine map causal variants. We report here a moderate-throughput approach to identifying regulatory GWAS variants, expression CROP-seq, which consists of multiplex CRISPR/Cas9 genome editing combined with single cell RNAseq to measure perturbation in transcript abundance. Mutations were induced in the HL60/S4 myeloid cell line nearby 57 SNPs in three genes, two of which, rs2251039 and rs35675666, significantly altered *CISD1* and *PARK7* expression, respectively, with strong replication and validation in single cell clones. The sites overlap with chromatin accessibility peaks, and define causal variants for inflammatory bowel disease at the two loci. This relatively inexpensive approach should be scalable for broad surveys and is also implementable for the fine mapping of individual genes. The above is published in *Biology Methods and Protocols*. This work was performed in collaboration with the group of Dr. Gang Bao, supervised by Dr. Ciaran Lee and with wet lab experiments performed by Dr. Yidan Pan at Rice University. My contribution was mostly the statistical analysis and writing of the paper. Single-cell RNA sequencing was performed in Georgia Tech by Dr. Dalia A. Gulick.

3.1 Introduction

The majority of genome-wide association study (GWAS)-identified SNPs are located in non-coding regions of genes, and are likely to influence disease risk and phenotypes by affecting gene expression[104]. Fine mapping of causal variants responsible for these signals is important for understanding which genes mediate phenotypic variation, dissecting mechanisms of action, assembling regulatory networks, and designing therapeutic interventions. It is recognized increasingly that GWAS peaks have a complex structure, the resolution of which is limited by linkage disequilibrium (LD) and the presence of multiple independent signals at many loci[55,56,105]. Since GWAS peaks often overlap with expression QTL (eQTL) signals, namely associations with gene expression, transcription-based experimental screening approaches can be used to prioritize likely causal variants within credible intervals that contain 100 or more polymorphisms. Two classes of approach have been reported, CRISPR/Cas9 genome editing[38], and massively parallel reporter assays (MPRA)[33], but have not been developed to systematically scan across the regulatory element(s) of a target gene.

Previous high-throughput CRISPR-based approaches to dissecting the impact of noncoding DNA have focused on defining *cis*-acting regulatory elements, rather than allelic effects of polymorphisms. They have generally utilized selection strategies followed by sequencing of barcodes from bulk cellular populations, assaying for enrichment or depletion of guide RNAs (gRNAs) targeting elements that are required for gene expression. In this way, Sanjana et, al[106] surveyed 700kb around the *NF1*, *NF2*, and *CUL3* loci by selecting for resistance to inhibition of BRAF in a melanoma cell line when transcription

of the genes is reduced, and Rajagopal et, al[35] tiled 40kb around four genes into which they had inserted a green fluorescence protein marker to select for gene expression.

Extending this approach genome-wide, two groups have surveyed function of the majority of binding sites for p53 and Estrogen Receptor-α in the context of oncogene-induced senescence in a breast cancer cell line[37], and for FOXA1 and CTCF mediation of target gene activity in breast and prostate cancer cell growth[36]. These experiments define enhancer elements required for essential gene function, and incidental findings related to the existence of polymorphisms in some elements are reported, but they do not provide a mechanism to systematically scan candidate SNPs in credible intervals.

Here we report an adaptation of the CROP-seq protocol[42], for regulatory fine-mapping. CROP-seq (CRISPR droplet sequencing) involves multiplex CRISPR/Cas9 transfection of a cell line with dozens to hundreds of gRNAs targeting different genes, followed by single cell RNAseq (scRNAseq) transcriptome profiling to monitor the consequences of inferred editing of the target gene. Even though not all cells are edited, the ability to detect which gRNA was present in each sequenced cell allows quantitative comparison of the effect of loss of function of the gene. In expression CROP-seq, we instead transfect dozens of gRNAs targeting different eSNPs in a credible regulatory interval and use the scRNAseq to monitor abnormal expression of linked transcripts (Figure 3.1). Microdeletion or mutation of the SNP in one hundred or more cells provides sufficient power to detect up- or down-regulation of expression consistent with most eQTL effect sizes. In a single experiment, we screened 57 SNPs in eQTL intervals of three genes associated with inflammatory bowel disease (*CISD1*, *PARK7*, and *DAP*), and showed, with

replication and subsequent validation, that in two cases a single SNP located within an open chromatin peak is likely responsible for the genetic association.

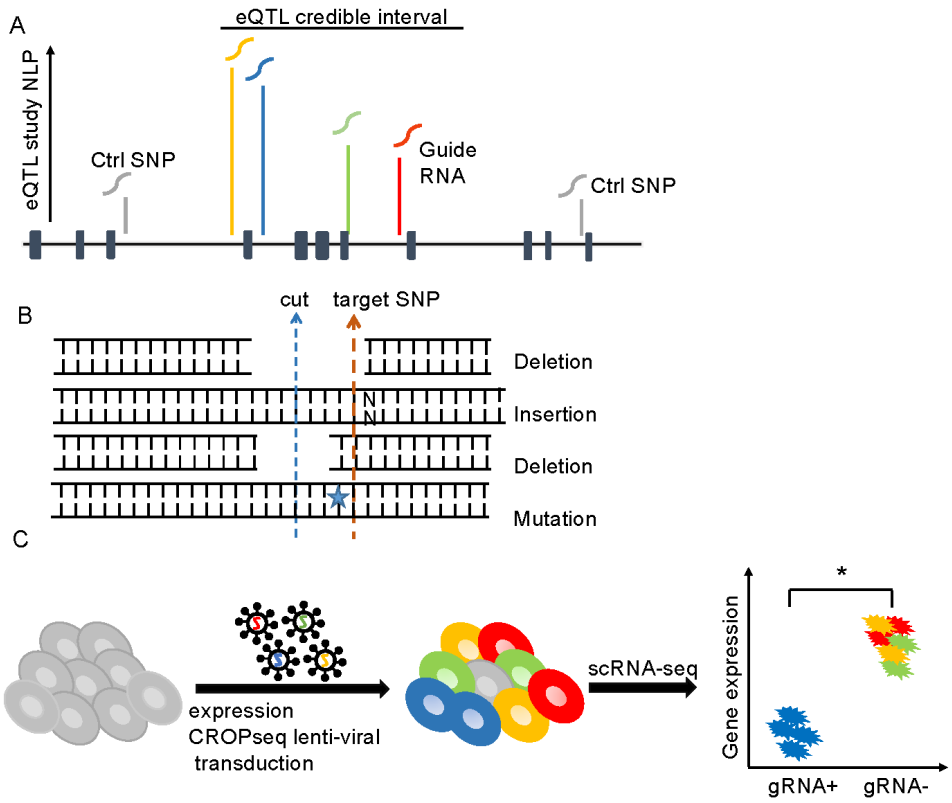

Figure 3.1 Experimental design of expression CROP-seq screening of eSNPs. (A) SNPs were selected with various eQTL p-values from one or two credible intervals for each eGene. Additional SNPs in low LD with the credible interval were selected as control SNPs. Each SNP was targeted by a single gRNA with minimal predicted off-target effect. The horizontal black line represents a hypothetical locus with exons indicated by solid blocks. (B) The Cas9 editing site may be a few bases away from the targeted SNP and can introduce four possible genetic alterations: deletion of both the cutting site and target SNP; insertion; deletion of only the cutting site; and mutation of the target SNP. (C) Pooled CROP-seq lentiviral libraries with 67 gRNA were transduced into the HL60/S4 cell line. Most cells were transduced with a single gRNA. Red, green, yellow and blue represent four different gRNAs. A few cells have zero (grey cell) or multiple gRNAs. After 10X single-cell RNA-seq identified the gRNA of each cell, differential expression of the linked transcript is evaluated between cells with the gRNA relative to cells with all other gRNAs.

3.2 Materials and methods

3.2.1 gRNA design and cloning

Approximately 20 SNPs were chosen for each gene based on prior eQTL mapping in peripheral blood mononuclear cells (PBMC)[57], along with five positive controls targeting the coding regions of the essential genes, *TUBB* and *RUNX1*, three negative controls that have no perfect target in the human genome, and one non-SNP targeting control. Each SNP was targeted for mutation, micro-deletion, or micro-insertion, by one single gRNA predicted in silico with COSMID software[26] to have a minimal likelihood of inducing off-target effects. The chromosomal position of each SNP and the flanking sequences were obtained from dbSNP[71]. All 19-base sequences followed by the correct *S. pyogenes* Cas9 Protospacer Adjacent Motif (PAM) sequence (NGG) inside the window were screened. gRNAs with GC rate over 80% or less than 40% were filtered out to ensure better cutting performance. The gRNAs with the shortest distance from the cut site to targeted SNP (in most cases less than 10 bases) and minimal predicted off-target effects were used in our study. All selected gRNAs (listed in online Table S1, https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter3_supp_tables) have only one perfect match to the whole reference genome, and negative controls had no perfect match in the human genome.

The CROPseq-Guide-Puro plasmid[42] (Addgene, Watertown MA, catalog number #86708, originally from Christoph Bock's lab) was digested by Esp3I (NEB, R0734S). For each designed gRNA sequence, a pair of annealed oligos was cloned into the vector before the gRNA scaffold and after the U6 promoter. Clones were pooled for Maxi-prep (Qiagen,

Hilden Germany, catalog number #12165) following the manufacturer's protocol. The gRNA distribution in the plasmid prep was validated by next-generation sequencing.

In order to estimate the nature and rate of editing, single cell clones were generated from the same cell population that was transduced by the two lentivirus vectors in preparation for scRNAseq. Integrated gRNA sequences of each single cell clone were identified by amplifying a 300bp to 400bp fragment surrounding the relevant target SNP from 73 single cell clones representing one of 19 individual gRNAs. Next-generation sequencing was used to characterize the edited sequences. Online Table S2 (https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter3_supp_tables) reports the exact edit in each clone, and shows that 92% of the cells contained at least one edited allele, with 29% showing a single edited allele. The remainder either had two different edits, or only a single edit (implying biallelic editing, or that the alternate allele was not amplified). Additional columns show whether the target SNP was disrupted by a mutation within 3bp of the target (67% of all clones) or whether the target SNP was directly disrupted (46%).

3.2.2 CROP-seq lentivirus library construction and transfection

Lentivirus production from lentiviral vectors CROPseq-Guide-Puro and lentiCas9-Blast[107] (Addgene, 52962) and was performed following Addgene's standard lentivirus production protocol using the Lenti-X 293T cell line (Takara, Kusatsu Japan, catalog number #632180). LentiCRISPRv2GFP[108] (Addgene, catalog number #82416) was used as the reporter in each transfection. Lentivirus was pelleted by using L-90K ultracentrifuge (with SW32-Ti rotor, 25,000 rpm for 1.5 hours at 4 degree) and dissolved in 100 µl 1xPBS.

Spinfection of HL60/S4 was performed according to the protocol from Feng Zhang lab [16]. Cells were seeded in a 24-well plate at a density of 1×10^6/ml with 5 μg/ml of polybrene (EMD Millipore, Burlington MA, catalog number TR-1003-G). Up to 10 μl of concentrated lentivirus was then added to each well, and cells were centrifuged at 1200×g for 1.5 hours at 33ºC. HL60/S4 cells were first transduced by CROPseq-Guide-Puro lentivirus. 24 hours after spinfection, cells were re-plated at a density of 5×10^5/ml with 2 μg/ml puromycin (SigmaAldrich, P8833) selection for 8 days or until no viable cells were observed. Cell viability was monitored every 24 hours, the media was changed every 48 hours, and cell density was maintained under 1×10^6/ml. The multiplicity of infection (MOI) was calculated, and the group with the least non-zero MOI was marked as HL60/S4-PuroR and used for downstream experiments for achieving optimal single gRNA assignment in the cell population.

After 3 days of recovery in regular culture media, HL60/S4-PuroR was transduced by lentiCas9-Blast lentivirus using spinfection with the same protocol as the CROPseq-Guide-Puro lentivirus transduction. 24 hours after spinfection, cells were plated at a density of 5×10^5/ml with 10 μg/ml blasticidin (Research Products International, Mt Prospect IL, catalog number B12200) selection for 7 days. After 3 days of recovery, cells were further selected by dual drug selection for 3 days (1 μg/ml puromycin and 5 μg/ml blasticidin) to remove residual non-puroR-blastR cells. Cells were then cultured in normal media for 10 days for global gene expression recovery.

3.2.3 Single-cell RNA sequencing and data processing

Single-cell RNA sequencing libraries were prepared using 10X Genomics (Pleasanton, CA) Chromium single cell 3' reagent kit V2 (PN-120267) and V3 chemistry (PN-1000092) with fresh cells for replicate one and replicate two, respectively. The average cDNA library size was 484 bp and 505 bp for replicate 1 and replicate 2, respectively. Sequencing was performed on an Illumina NextSeq 550 system in high-output mode, generating paired-end libraries (28bp for read1 and 98bp for read2). Raw sequence data was first de-multiplexed from BCL files into FASTQ files by using "cellranger mkfastq", with 10X Cell Ranger software. The human reference genome (hg38) was supplemented with 67 gRNA artificial chromosomes, each of which include 241 bp U6 promoter sequences, 8 bp gap sequences between U6 promoter and gRNA, 20 bp gRNA sequences and 261 bp backbone sequences downstream of the gRNA. This 67 gRNA extended hg38 was indexed by "cellranger mkref" with extension ".fa" and ".gtf" files as input. Single cell gene counts were generated by "cellranger count" by aligning reads to the extended hg38 by STAR aligner[76] with default settings. The estimated total number of cells detected was 8,671 and 10,087 for replicate 1 and replicate 2, respectively. The average total sequencing read depth per cell were 58,413 and 45,636 for replicate 1 and replicate 2, respectively.

Each cell was distinguished by a cell barcode and a gRNA sequence, and in the majority of cases a single gRNA was uniquely assigned to each cell (Figure 3.2A). To confirm that the scRNAseq profiles adequately represent rates of lentiviral transformation, we amplified and sequenced the integrated gRNA sequence from 96 single-cell derived CRISPR/Cas9 clones, observing a similar distribution of gRNAs (Figure 3.2B).

The count of UMI (unique molecular identifiers) for each gRNA per cell was quantified by "cellranger count". The gRNA-cell expression matrix was extracted from the cell-gene expression matrix. Only cells with a single gRNA expressed were included in the downstream analysis. If the gRNA UMI count was greater than 0, it was coded as 1, otherwise coded as 0. This updated gRNA-cell identity matrix was appended to the cell gene expression matrix, providing the gRNA assignment for each uniquely assigned cell.

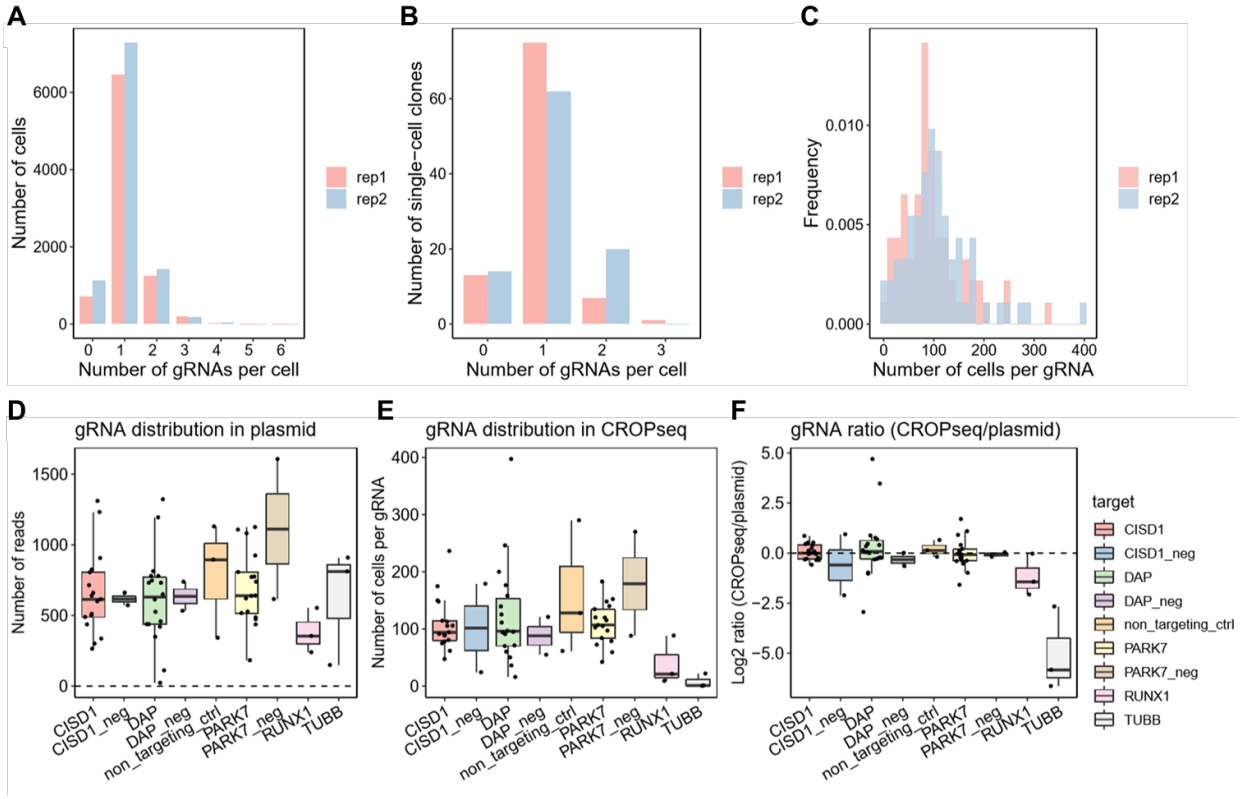

Figure 3.2 Guide RNA distributions. (A) The distribution of number of gRNA per cell detected from scRNAseq. (B) Similar distributions were observed for amplified gDNA insertions in 96 single cell clones. (C) Histogram showing the number of cells containing each gRNA in the two replicates. (D) Raw read counts of each gRNA in the pooled library with respect to the candidate eSNPs or negative controls. (E) Number of cells assigned to each gRNA in scRNAseq shows a similar profile. (F) Log_2 ratio of

normalized RNA to DNA implying no deviation from expected equivalence, except for positive controls in the essential genes *RUNX1* and *TUBB*.

3.2.4 Expression data quality control and normalization

The R package Seurat V3.0[109] downloaded from https://github.com/satijalab/seurat was used for single-cell RNAseq expression data processing and analysis. Low quality cells with less than 200 genes expressed as well as lowly expressed genes detected in fewer than six cells were filtered out. Next, cells that had between 2,000 to 7,000 expressed UMIs and cells having less than 25% mitochondrial counts were retained. After quality control, 8,192 cells and 16,372 genes were kept for replicate 1, and 8,921 cells and 16,407 genes were kept for replicate 2. Then gene expression measurements were normalized by dividing by the total UMI counts and multiplying by the scaling factor 10,000, and transformed to logarithm base 2.

Linear dimensional reduction principal component analysis, PCA, was first performed with the top 2,000 identified highly variable genes and default settings in Seurat[109]. The "JackStraw" function implemented in Seurat was used to determine the significant PCs. Nonlinear dimensional reduction by uniform manifold approximation and projection (UMAP)[110] was then performed based on the top 20 significant PCs with default settings. Each cell was assigned a score summarizing expression of G2/M and S phase gene markers implemented in Seurat package, and thereby classified into either G2M, S or G1 phase according to its cell cycle score. The UMAP projection in Figure 3.3 suggests some clustering of cells by cycle identity (f) which also correlates with read depth (e) and number of detected genes (d). However, individual transcripts do not cluster with respect to these properties (a-c).

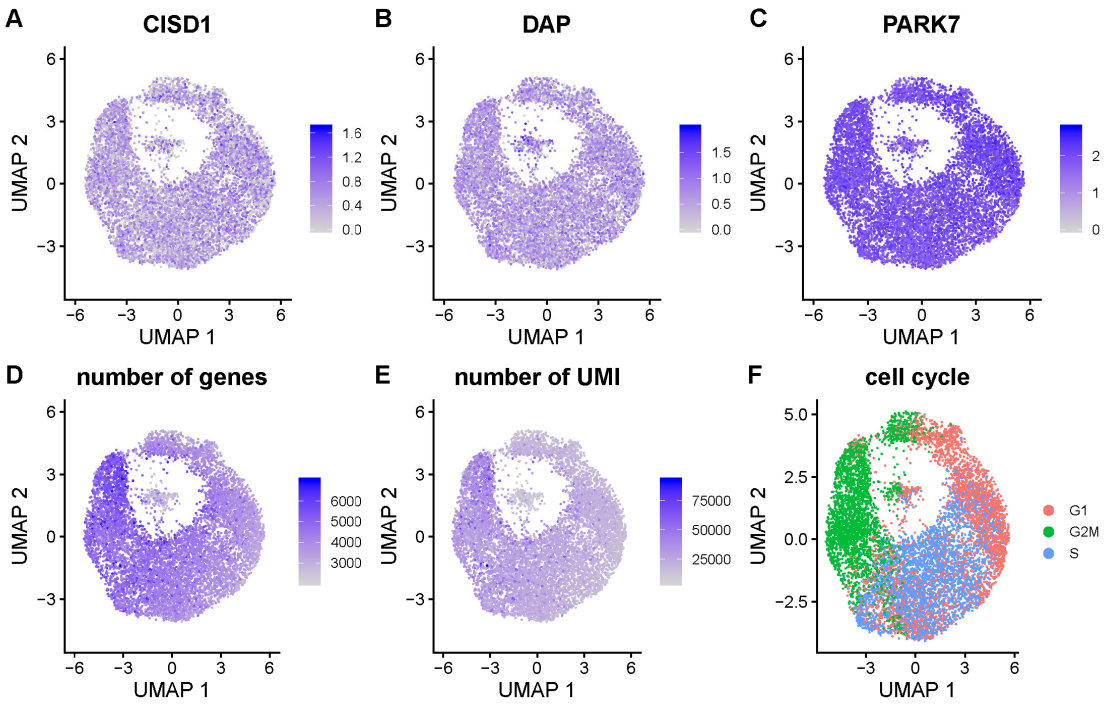

Figure 3.3 UMAP visualization of single-cell transcriptome profiles. Each panel shows the UMAP projection implemented in Seurat [18] of the first 20PC of UMI abundance variation. Cells are color coded according to: log2 normalized UMI counts of (A) CISD1, (B) DAP or (C) PARK7; (D) number of genes; (E) raw UMI counts detected in each cell; and (F) inferred cell cycle state. The UMAP projection is the same as in Figure 3.4A.

3.2.5 Hypothesis Testing

While more complex models, for example cell cycle fitting, or matching cells according to UMI count, were considered, they did not change the conclusions. We therefor report the simplest statistical approach to hypothesis testing. Each of the candidate eSNPs from one credible interval was fit with the normalized expression of its respective target eGene. Univariate linear regression was performed with "lm" function in R. Student's *t*-test was conducted to test the null hypothesis that the coefficient of eSNP equals to zero in the regression. Bonferroni correction was applied for the simultaneously

performed independent t-tests within each locus, namely alpha = 0.05 divided by 20 tests per gene for a gene-wise nominal adjusted critical value of 0.0025.

3.2.6 Validation in CRISPR edited cells with single gRNAs

We further validated the effects of the two identified eSNPs on expression of the target genes *CISD1* and *PARK7* with two single guide RNA approaches. First, HL60/S4 cells were transduced by lentiCas9-Blast and CROPseq-Guide-Puro lentivirus with the same gRNAs targeting rs2251039, rs35675666, or a negative control, using the same protocol as described before but as individual gRNAs in three separate transfection experiments. DNA was extracted from the bulk edited cells, the targeted regions were amplified by PCR, and genotypes assessed by Sanger Sequencing (Eurofins Genomics) using the Synthego ICE strategy[111]. Online Table S4 (https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter3_supp_tables) shows the proportions of most commonly edited alleles at the two loci, accounting for 87% of the rs2251039 and 79% of the rs3567566 edits. In both cases the bulk of the edits are indels adjacent to the SNP. Transcript abundance in duplicate bulk RNA extracts was estimated by real time quantitative reverse transcription polymerase chain reaction (qRT-PCR).

Second, these bulk edited cell suspensions were single cell sorted and seeded into 96-well plates with standard culture media and cultured for 14 days. Half of the cells for each single cell clone were taken at day 7 for genotyping using next-generation sequencing of the targeted region. From these, a set of single cell clones that expanded successfully and had the SNP removed/ affected were chosen for qRT-PCR, including 7 rs22510396 gRNA-edited clones, 6 rs35675666 gRNA-edited clones, and 4 clones edited with a non-

targeting negative control gRNA. The sequences of the targeted alleles are shown in online Table S5 (https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter3_supp_tables). RNA from each selected single cell clone was extracted using a RNeasy Mini Kit (Qiagen, cat. no. 74104) and reverse transcribed with iScript™ cDNA Synthesis Kit (Biorad, cat. no. 1708891) following standard protocols. Cycle thresholds for *CISD1*, *PARK7*, *GAPDH* and *ACTB* were quantified by qRT-PCR with three technical replicates. The $2^{-\Delta\Delta Ct}$ method was used to analyse the qPCR results, in which gene expression in cells with gRNA targeting *CISD1* or *PARK7* was normalized by the average of corresponding expressions in negative controls as well as the average of two housekeeping genes. Similar results were obtained with each single control gene.

3.3 Results

Multiplex CRISPR/Cas9 editing of myeloid HL60/S4 cells, followed by single-cell RNA sequencing (scRNAseq), was used to monitor the impact of candidate regulatory SNP disruption on gene expression of three genes in a single experiment. We screened 57 candidate SNPs along with 10 control SNPs, using lentiviral transfection of a single-cell clone of the HL60/S4 myeloid human cell line. Approximately 20 SNPs were chosen for each gene based on prior eQTL mapping in peripheral blood mononuclear cells (PBMC)[57], along with five positive controls targeting the coding regions of the essential genes *TUBB* and *RUNX1*[12], three negative controls that have no perfect target in the human genome, and one non-SNP targeting control. Each SNP was targeted for mutation, micro-deletion, or micro-insertion, by one gRNA predicted to have a minimal likelihood of inducing off-target effects[26]. Two lentiviral vectors were used to successively infect HL60/S4 cells, the first one encoding both the puromycin resistance gene, and a single gRNA (positioned such

that transcripts containing the guide would be captured by RNAseq), the second encoding both a blasticidin resistance gene and the Cas9 enzyme. This design facilitates identification of which guide(s) from the pool of 67 guides in the transformation mix, each single cell has taken up.

Each cell was distinguished by a cell barcode and a gRNA sequence, and in the majority of cases a single gRNA was uniquely assigned to each cell (Figure 3.2A). To confirm that the scRNAseq profiles adequately represent rates of lentiviral transformation, we amplified and sequenced the integrated gRNA sequence from 96 single-cell derived CRISPR/Cas9 clones, observing a similar distribution of gRNAs (Figure 3.2B). The distribution of cells per unique guide ranged from 10 to 550, with an average of 117.3 ± 66.5 cells, ensuring sufficient statistical power to detect eQTL with moderate to high effect sizes (Figure 3.2B). Furthermore, each of the 67 gRNAs were evenly distributed in the transfection mix of cloned DNA plasmids. There were no significant fold changes in guide abundances in scRNAseq relative to DNA plasmid levels, with the exception of the essential genes RUNX1 and TUBB (Figure 3.2D, E). These results confirm the efficiency of Cas9-mediated editing and imply that disruption of the regulatory regions of the three target genes did not compromise cell viability.

We characterized the transcriptional profiles of 6,358 and 6,974 single gRNA assigned cells in two biological replicates, with on average 58,413 and 45,636 sequencing reads per cell. Cells with two or more gRNAs were excluded from the eQTL analysis. UMAP projection shows that the expression of *CISD1*, *DAP* and *PARK7* was uniformly allocated among the clusters (Figure 3.3A-C). There was some clustering of the cells with respect to total number of reads and of UMI, which to some extent correlates with cell

cycle stage (G1, S or G2/M) (Figure 3.3D-F), while a small number of low-transcript abundance cells were also excluded from further analysis. Figure 3.4A shows that cells with a single gRNA (for example targeting to rs2251039) (orange) and cells with gRNAs other than the one targeting to rs2251039 (blue) were also evenly distributed with respect to the clustering.

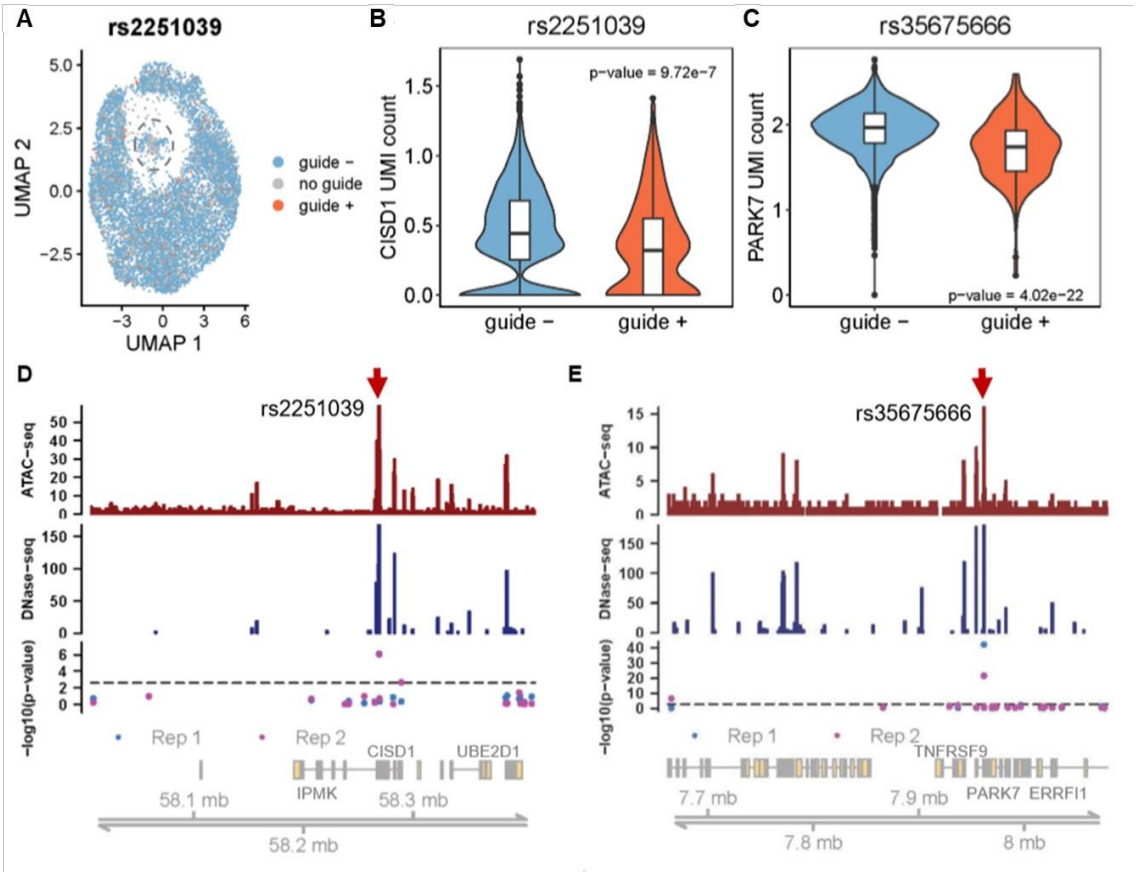

Figure 3.4 Identification of causal variants by expression CROP-seq. (A) Nonlinear dimensional reduction of 20PCs of single cell transcriptome profiles by UMAP in Seurat (18). Cells are color coded as orange rs2251039 gRNA; blue gRNAs other than rs2251039; grey without any gRNA. Excluded cells with abnormally low number of UMI indicated by the dashed circle. (B, C) Violin plots show kernel density distributions of the expression of normalized log2 *CISD1* and *PARK7* UMI counts of cells with (+) or without (-) gRNAs targeting rs2251039 or rs35675666. Boxplots show the median, first and third quantiles of the data. (D, E) Chromatin accessibility of identified expression CROP-seq peaks for *CISD1* and *PARK7*. Top and middle histograms show HL60 ATAC-seq (GSM2083754) peaks and HL60 DNase-seq (ENCSR000ENU) peaks, respectively. The third panel shows the negative log10 p-value of student's *t*-test statistic

corresponding to the genomic location of the tested SNPs in two biological replicates, with the gene-wise Bonferroni adjusted p-value = 0.0025 threshold indicated by the dashed line. The bottom panel is a schematic of all gene transcripts annotated from the UCSC gene table (hg38). Red arrows point to the two inferred causal eSNPs in both replicates.

Univariate linear modelling was sufficient to resolve individual eSNP effects observed in two replicates of the experiment conducted several months apart. In the first replicate, two SNPs, one in *CISD1* (rs2251039), and one in *PARK7* (rs35675666) were identified as putatively causal ($p<10^{-6}$ and $p< 10^{-20}$; both with Bonferroni corrected p-value < 0.0025). The same two SNPs replicated in the second experiment, at similar significance levels (Figures 3.4B, C). Only two other nominally-significant associations were observed, in a single replicate at *DAP* and a single replicate at *CISD1*. Moreover, we also examined if the knockout or mutation of targeted SNP would also influence the expression of adjacent genes. We tested the association between *CISD1* eSNPs with *IPMK* or *UBE2D1* expression, as well as between *PARK7* eSNPs and *ERRFI1* (*TNFRSF9* abundance was too low to assess), and between *DAP* eSNPs and *ANKRD33B* expression. None of the candidate eSNPs showed significant association with the nearby transcripts in either replicate (online Table S3, https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter3_supp_tables).

Both of the significant SNPs were also among the most significant hits in the CAGE study that motivated sampling of the three genes[57]. Further evidence that they are likely causal is provided by the observation that they both lie under chromatin accessibility assay peaks. Figures 3.4D, E show the location of each assayed SNP at *CISD1* and *PARK7* relative to ATACseq and DNase-seq (ENCSR000ENU) profiles of HL60/S4 cells[113]. The majority of the nonsignificant expression CROP-seq SNPs lie between ATAC or DHS

sites. Furthermore, rs2251039 is located just 28bp upstream of the transcription start site of *CISD1* and is within a binding motif for the Bhlhe40 transcription factor, a known regulator of cytokine production in T-cells[114].

The function of both SNPs was validated using qRT-PCR on both bulk edited cells and in single cell clones, with results illustrated in Figure 3.5. Bulk transfection of HL60 cells with single guides resulted in down-regulation of the associated *CISD1* and *PARK7* transcripts relative to cells transfected with non-targeting control gRNAs to a similar degree as inferred in the expression CROP-seq assays. More precise evidence for down-regulation of gene expression after disruption of the targeted SNP was obtained by qRT-PCR of 6 single cell clones containing indels in or adjacent to rs35675666 in *PARK7*, or 7 single cell clones containing indels in or adjacent to rs2251039 in *CISD1*. Relative to expression in 4 non-targeting control clones and normalized to the unaffected housekeeping genes *GAPDH* and *ACTB* with the $2^{-\Delta\Delta Ct}$ method[115], all targeted transcripts showed between 20% and 87% reduced abundance, with $p<0.005$ for both genes.

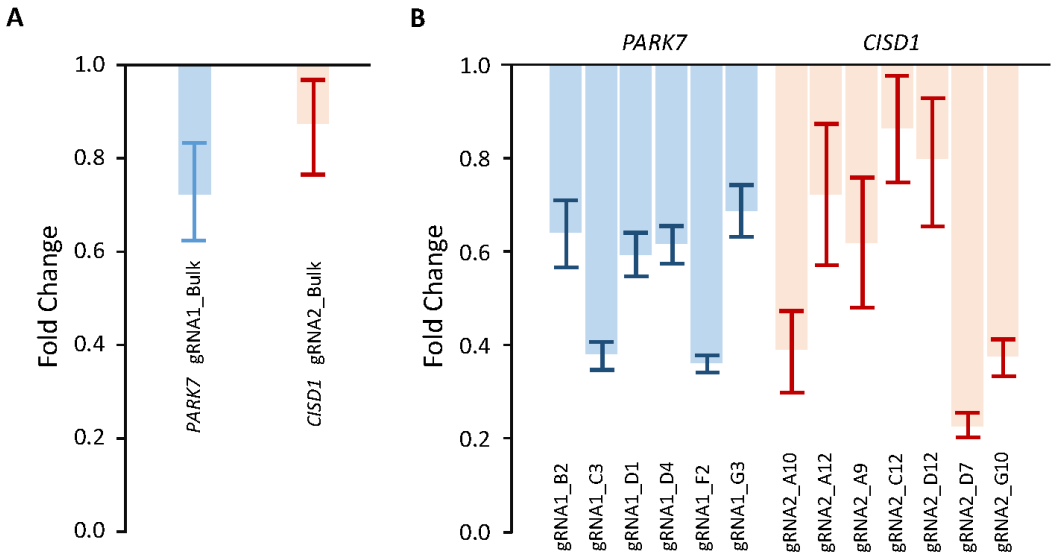

Figure 3.5 Targeted gene expression change of CRISPR-edited cells with single gRNA. qRT-PCR was performed on both bulk CRISPR-edited cells (A) and single cell clones (B) with gRNAs targeting rs35675666 (gRNA1, *PARK7*) and rs2251039 (gRNA2, *CISD1*) respectively. Cycle threshold (Ct) values were normalized relative to 4 biological replicates with cells expressing Cas9 and a non-targeting gRNA using the $2^{-\Delta\Delta Ct}$ method [24]. The expression of the transcript associated with each targeted SNP was reduced in both qRT-PCR designs, with significance (one tailed t-test of deviation of estimated fold reduction relative to unity) indicated by the indicated p-values: bulk-gRNA1, p =0.02; bulk-gRNA2, p=0.05; single cell clone gRNA1 p=0.0004; single cell clone gRNA2, p=0.0015.

3.4 Discussion

Published genome-editing strategies for interrogating regulatory elements either utilize CRISPRi or CRISPRa to inhibit or activate transcription[43,116], or rely on assays that select for essential gene function[36,37,106] or reporter gene expression[94]. Neither approach is suitable for systematically screening the function of each of the candidate SNPs in a credible interval of a typical gene. Massively parallel reporter assays have been used to this end more successfully. For example, Van et, al[32] evaluated 32,373 variants at 3,642 eQTL by inserting 180bp oligonucleotides encompassing each SNP in front of a minimal

promoter, finding 842 polymorphisms that drive reporter expression in a lymphoblast cell line at different levels. An even larger scale experiment by van Arensbergen *et al*[32] surveyed 5.9 million variants, namely 57% of all known common variants in the human genome, by associating short DNA fragments with a barcode and assaying tag abundance in hepatic and erythroid cell lines. They identified over 30,000 candidate eSNPs, most cell-type specific, and described enrichment with various chromatin features. Impressive as these studies are, there is always the caveat that enhancer activity outside normal chromatin context may not be accurate, and perusal of the SuRE database[32] suggests that many sites have large but nonsignificant effects since the majority of the cloned fragments do not drive expression. Hence, false negative rates are not known, and complementary assays that systematically interrogate credible intervals in the same promoter context should also be informative.

Our approach is to directly measure expression of a gene after genome-editing of a set of regulatory polymorphisms.[117] Targeted reporter assays analyzing RNA from bulk preparations of clonal cell lines, have low power to resolve typical eQTL effects that explain in the range of 10% to 20% of the variance of the target gene. The CROP-seq single-cell eQTL screening strategy gains power from the sequencing of thousands of cells in parallel. The effect size of rs2251039 in *CISD1* corresponds to a reduction of around 0.5 standard deviation units due to gene editing, equivalent to an eQTL explaining between 5% and 10% of the variance (depending on the allele frequency and assumptions about whether one or both alleles are disrupted in the CROP-seq). The much larger rs35675666 effect at *PARK7* could correspond to a three-times larger-effect eQTL, or may reflect more efficient gene editing by the particular gRNA, which appears to create

large deletions encompassing the SNP (online Table S2, https://github.com/RuoyuTian/PhD_thesis/tree/master/chapter3_supp_tables).

Interestingly, rs35675666 is located in the first intron of the *PARK7* transcript and is a GWAS SNP for ulcerative colitis (p-value = 5×10^{-9}) and inflammatory bowel disease (p-value =1×10^{-15})[117]. Although McCole et, al[118] argued that *ERRF11* is a strong candidate gene in the interval due to the impact of ErbB receptor feedback inhibition on epithelial apoptosis and possibly barrier function, the absence of effect on *ERRF11* transcript abundance calls into question that inference and instead promotes *PARK7* as the likely causal gene. *PARK7* encodes a C56 peptidase family member that has been shown to function as a regulator of mitochondrial respiration and lysosomal function[119]. Autosomal recessive loss of function leads to early-onset Parkinson's disease, and reduced expression may conceivably disrupt autophagy or oxidative stress sensing, both of which are implicated in ulcerative colitis[120].

In theory, expression CROP-seq should be powered to fine map eSNPs within credible intervals that explain just a few percent of the expression of a target gene. We were able to confirm the identity of autoimmune disease-associated GWAS variants in two loci, but did not detect the third eQTL or resolve the secondary associations that are nevertheless present at each of the loci we tested. Comprehensive fine-mapping will often require 100 or more gRNAs per gene, but is well within the scope of the experimental pipeline described here. Limitations include the inability to target all SNPs due to absence of appropriate PAM sequences, reduced power for genes expressed at levels close to the limit of detection in scRNAseq, and appropriateness of the cell line(s) chosen for the assay. Replication is likely to be important for the confident identification of relatively small

effect eSNPs, particularly given that cell passaging, mutation, and random variability during cell culture can affect the transcriptional background[91]. Future experiments may also use prime editing[121] to specifically replace one allele with the alternate allele, rather than inducing mutations at or near the site. Finally, we also show that integration with functional annotation data may help to validate inferred eSNPs and identify the likely transcription factors they bind. In all, our method facilitates the genetic screening of non-coding variants and the transcriptional interpretation of risk variants in the post-GWAS era.

CHAPTER 4. CLASSIFYING THE TOLERANCE OF PROMOTER REGIONS BY THE BURDEN OF RARE VARIANTS ACROSS TISSUES

ABSTRACT: A large majority of genetic variants are rare in the population, some unknown fraction of which may have large regulatory effects through which they contribute to disease risk. However, the study of rare variants lags behind common variants, in part due to the insufficient sample size of risk alleles. Importantly, some of these may also be responsible for observed common variant effects, due to synthetic association with disease risk in GWAS. Moreover, investigation of the functionality of rare variants across individual tissue types has just started. Here, I used GTEx V8 data from 740 European individuals across 49 tissues to show that there is enrichment of rare alleles in the promoter regions of individuals who have extreme expression across tissues. The rare allele burden gradually weakens with distance from the TSS. Moreover, significant depletion of rare allele associated with extreme expression is observed in loci with low relative promoter polymorphism across tissues. From these observations, I develop a novel gene categorization system that annotates genes according to the degree of intolerance to regulatory rare variants and relative regulatory region polymorphism. Genes that are intolerant to rare regulatory mutation and has constraint in regulatory region are informed with tissue cluster, implying its potential application in updating the existing regulatory rare variant prediction and genetic scores with tissue information.

4.1 Introduction

Rare mutations make a major contribution to disease risk for individual people. However, rare variant associations cannot be detected by GWAS due to the insufficient sample size of the risk allele. On the other hand, rare risk alleles may be responsible for common variant GWAS associations if there is "synthetic association" between multiple rare risk alleles and a GWAS-tagged common SNP. In these cases they will usually go undetected. A handful of recent studies have shown that rare variants are nevertheless associated with extreme gene expression in peripheral blood and other tissues[102,122]. Notably, the Bayesian framework algorithm RIVER (RNA-informed variant effect on regulation) integrates functional genomics data to compute the probability that a given rare variant has a regulatory impact[122], prioritizing the regulatory rare variants as likely causal.

Current approaches utilize exome sequencing data and evolutionary constraint estimates to develop scoring systems that quantify genic intolerance to mutation[123-125]. By comparing the expected with observed PTV (protein-truncating variants) of each gene from whole exome sequencing data, genes are classified into null, recessive and haploinsufficient. One corresponding score, pLI, quantifies the probability of being loss-of-function intolerant[123]. Genes with pLI $\geq$ 0.9 are intolerant to loss-of-function mutation, while genes with pLI $\leq$ 0.1 are tolerant to loss-of-function mutations. A similar genetic score RVIS (Residual Variation Intolerance Score) takes the studentized residual of a regression of the number of missense and truncating variants on the sum of all variants, but is restricted to protein coding regions[125]. Genes with negative RVIS are intolerant to functional variations, and are enriched in OMIM. Furthermore, a similarly calculated score ncRVIS (noncoding RVIS), compares the predicted and observed prevalence of regulatory

variants, and performs well in predicting dosage sensitive genes[124]. However, none of those scores use the non-coding region rare variants and evolutionary constraint measures to categorize genes, and have not emphasized tissue specific disease-associated genes.

Here, I performed a regulatory rare variant burden test on the large GTEx dataset with 740 European individuals including both RNA sequencing, and whole genome sequencing data across 47 tissues and 2 cultured cell lines. I used this data to develop a novel gene categorization system that systematically places genes into tissue clusters according to the degree of intolerance to regulatory rare variants and the relative regulatory region polymorphism. I compare the new score to pLI and RVIS and show that it provides an alternative method for prioritizing rare promoter variants as potentially causal.

4.2 Materials and methods

4.2.1 Dataset

I used sequencing data from the current release version 8 of the Genotype-Tissue Expression (GTEx) project. Samples were collected from 54 non-diseased tissue sites sampled at autopsy from 980 deceased individuals. RNA-seq, WES and WGS was available. RNA-seq data were downloaded from the GTEx portal fully processed, filtered and normalized, as used for eQTL analysis by the GTEx consortium[126]. Out of 980 individuals, 866 individuals had WGS data, and after filtering out non-European ancestry individuals to avoid complications of population stratification, 740 individuals were retained. 47 tissues and 2 cultured cell lines were included in the analysis. The number of samples with both RNA-seq and WGS data for each tissue is summarized in Table 4.1.

Table 4.1 GTEx V8 tissues and samples with both RNA-seq and WGS data

Tissue	Sample size	Tissue	Sample size
Adipose Subcutaneous	492	Esophagus gastroesophageal junction	281
Adipose visceral omentum	402	Esophagus mucosa	424
Adrenal gland	200	Esophagus muscularis	396
Artery aorta	338	Heart atrial appendage	322
Artery coronary	180	Heart left ventricle	334
Artery tibial	489	Kidney cortex	65
Brain amygdala	119	Liver	183
Brain anterior cingulate cortex BA24	136	Lung	444
Brain caudate basal ganglia	173	Minor salivary gland	118
Brain cerebellar hemisphere	158	Muscle skeletal	602
Brain cerebellum	189	Nerve tibial	449
Brain cortex	184	Ovary	140
Brain frontal cortex BA9	158	Pancreas	252
Brain hippocampus	151	Pituitary	220
Brain hypothalamus	157	Prostate	186
Brain nucleus accumbens basal ganglia	182	Skin not sun exposed suprapubic	440
Brain putamen basal ganglia	154	Skin sun exposed lower leg	518
Brain spinal cord cervical c-1	115	Small intestine terminal ileum	144
Brain substantia nigra	101	Spleen	185
Breast mammary tissue	337	Stomach	269
Cells cultured fibroblasts	416	Testis	277
Cells EBV-transformed lymphocytes	116	Thyroid	494
Colon sigmoid	273	Uterus	108
Colon transverse	305	Vagina	122
		Whole blood	574

4.2.2 Rare variants burden test

Rare variants are defined as variants with minor allele frequency smaller than 0.05 and were extracted from the WGS genotype data. If a gene has multiple transcripts, the transcription start site (TSS) was set as the most upstream TSS. For all analysis, I heuristically defined the promoter region as $\mathrm{TSS} \pm 1kb$. Rare variants located in other

regulatory regions were also extracted for some analyses, including 20-21 kb, 10-11 kb, 5-6 kb, 3-4 kb regions upstream of TSS, or 3-4 kb, 5-6 kb regions downstream of TSS. Regulatory region rare variants genotypes were available for 17, 158 genes, excluding the *HLA* gene complex, from 740 European individuals.

The burden test conducted here was introduced explicitly in Zhao et, al[102]. In brief, for each gene, the transcript abundance of each gene in each individual was sorted into bins according to rank of relative expression. Each bin has 6 individuals. Number of expression bins vary in tissues depending on sample size. For each bin of each gene, the number of rare alleles in a specified regulatory region of individuals in that bin was counted, and the sum of rare allele counts for each bin over all 17,158 genes was computed. Typically, this results in a "smile plot" where the counts are highest in the bottom and top percentile bins, implying an excess burden in individuals with extreme expression. To evaluate the relationship between rare variants in regulatory region with expression statistically, a linear quadratic regression analysis

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \quad (1)$$

was performed by fitting the numeric expression bin order (x) with the summed rare allele counts (y) in corresponding bins, where $\beta_0, \beta_1, \beta_2$ are the coefficients, ε is the random error which is assumed to be normally distributed with a mean of zero. The burden tests of rare variants were performed for the 49 tissues. Subsequently, within each tissue, tests were performed with rare variants in the previously specified seven regulatory regions at increasing distances from the TSS.

4.2.3 Computation of nucleotide diversity

Nucleotide diversity is defined as the average number of nucleotide differences per site between two randomly chosen DNA sequences, denoted by π,

$$\pi = \sum_{ij} x_i x_j \pi_{ij},$$

where x_i and x_j is the frequency of the ith and jth sequence in the population, respectively, π_{ij} is the number of nucleotide differences per nucleotide site between the ith and jth sequence[127]. I computed the nucleotide diversity by VCFtools[128] with 740 European individuals WES and WGS genotypes. To compute the coding region nucleotide diversity, π_{cod}, I used WES data of 5,430 genes. To compute the promoter region nucleotide diversity, π_{prom}, I used the WGS data of 17,153 genes, and the promoter region is defined as TSS$\pm 1kb$.

4.3 Results

4.3.1 The burden of rare variants in extreme expression across tissues

Significant quadratic linear regressions of fitting expression bin (x) to number of rare variants in individuals (y) were initially observed in peripheral blood samples of 472 genes from 410 individuals[102]. Similarly, using a different methodology, burdens were also observed in GTEx v6 data for 449 individuals across 44 tissues for rare regulatory variants in 10kb TSS region for outlier expression both in single and multiple tissues[122].

I systematically performed the burden test with data for 740 individuals and 49 tissues from GTEx v8 RNA-seq and whole-genome sequencing data, performing each test across tissues over regulatory regions in 1kb-sliding windows. The “smile” plots in Figure

4.1 show enrichment of rare variants in individuals in the extremes of either high or low expression. Each incremental step towards low or high expression is the first order of expression bin (x) plus a constant, the coefficient of x in the regression (1). A significant quadratic relationship is observed for most of the tissues in each of five tissue clusters (see explanation in section 4.3.4). GTEx v8 data provides sufficient sample size for all tissues. There is no correlation between sample size (Figure 4.2a) or significance (Figure 4.2b, c, d), and the enrichment of rare variants within 2kb region of TSS shows in most tissues (Figure 4.2b, e).

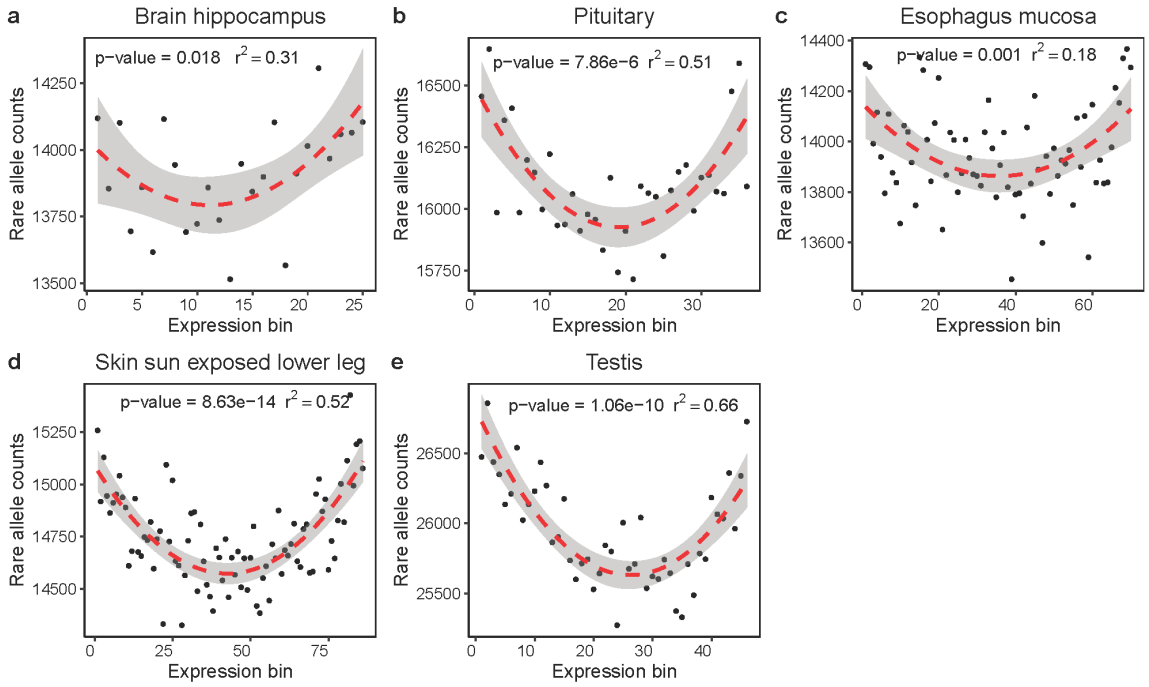

Figure 4.1 Linear quadratic regression of rare allele counts and expression bin. Scatter plots show promoter region (TSS $\pm$ 1kb) rare allele (MAF < 0.05) counts (y) for the corresponding expression bin (x). A fitted linear quadratic regression line is drawn with grey shade showing standard error of prediction. A tissue with significant fitting in each of five tissue cluster is shown: (a) brain hippocampus for B cluster; (b) pituitary for E cluster; (c) esophagus muscularis for C cluster; (d) skin sun exposed lower leg for O

cluster; (e) testis for T cluster. The p-value of each model is reported on the plot accordingly.

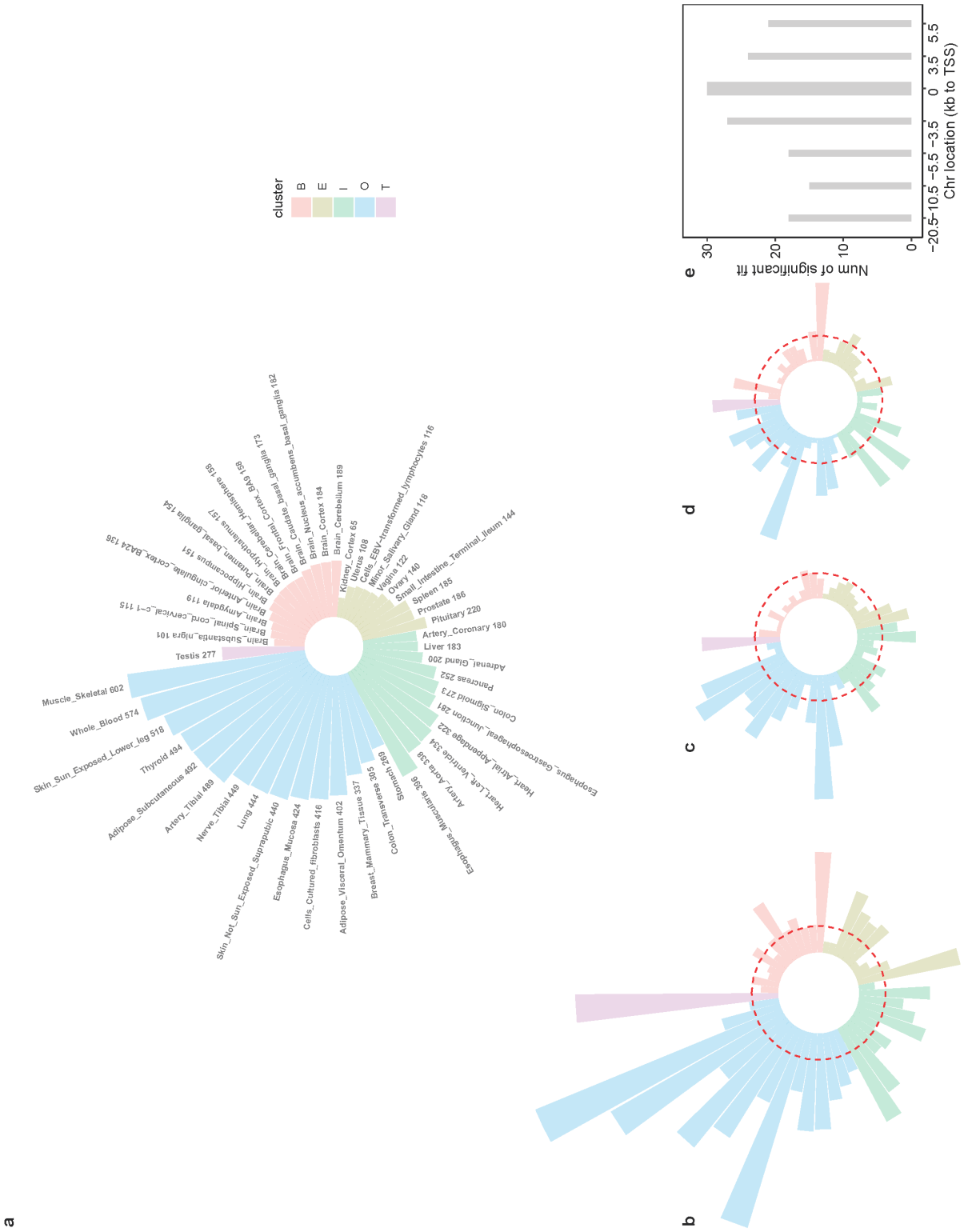

Figure 4.2 Summary of burden tests across tissue and regulatory regions. (a) Circular barplot shows tissues and sample size (Table 4.1) used for burden tests, colored by tissue clusters, B, E, I, O and T. (b) Significance (negative $\log_{10}$ p-value) of each burden test in each tissue within 2kb TSS, or in the regions (c) 5-6 kb upstream and (d) downstream of the TSS. The red dashed line is the significant level (α) at 0.05, $-\log_{10}\alpha = 1.3$. (e) Barplot shows the number of significant regressions at each region around TSS across tissues. The width of each bar is equal to the width of sliding window.

The magnitude of enrichment decreases as the 1kb sliding window moves away from TSS (Figure 4.2c, d), and there are fewer tissues showing the enrichment. However, the burden of distant rare variants, located within 20-21 kb upstream of TSS, is still observed in 18 tissues (Figure 2e). These results confirm that the rare variants near the promoter are likely to lead to extreme gene expression values, and that in some cases distal enhancer sequence also show a more modest impact. It is interesting that the O cluster has the largest number of tissues showing in rare regulatory enrichments, whereas the B (Brain) cluster has the minimum. In summary, the burden of rare variants in the TSS region is detected in most of the 49 tissues, and the regulatory rare variants are enriched most heavily in regions close to the TSS.

4.3.2 The linear correlation of coding and promoter region nucleotide polymorphism

To quantify the genetic variation of regions in genes, I computed the coding region nucleotide polymorphism (π_{cod}) for 5,430 genes from WES data, and the promoter region nucleotide polymorphism (π_{prom}) for 17,153 genes by WGS data of 740 European individuals, where promoter region is defined as within 2kb of TSS. By merging the two sets of nucleotide polymorphism, 5,416 genes with both π_{cod} and π_{prom} were retained for analysis. The linear regression (Figure 4.3) fitting π_{cod} to π_{prom} is highly significant with p-value = 9.4e-6, and can be described as

$$\pi_{prom} = 0.04 + 0.04\pi_{cod} + \varepsilon \qquad (2)$$

The coefficients were estimated using the "lm" function in R. The positive intercept indicates that there is an excess of genetic polymorphism in promoter regions relative to adjacent coding regions. Moreover, only 3 genes have zero π_{prom}, whereas multiple genes have zero coding region nucleotide polymorphism.

The residual of the fit of π_{prom} on $\pi_{cod}$was extracted for each gene, representing the relative genetic diversity. Genes with the top five positive residuals indicating greater promoter polymorphism are *MUC12*, *TRBV5-6*, *ALG1L2*, *LINC02014* and *HLA-DPA1*, labelled in grey in Figure 4.3. *MUC12* encodes a component of the epithelial protective barrier and as associated with colorectal cancer. The *HLA* region is involved in immune defense and is known to be highly polymorphic. Genes with the bottom five negative residuals are *RP11-809H16.5*, *RP13-131K19.2*, *KC6*, *AC008079.10* and *RP11-64K12.9*, which coding regions are unconstrained. Functional enrichment analysis of additional genes can be explored further.

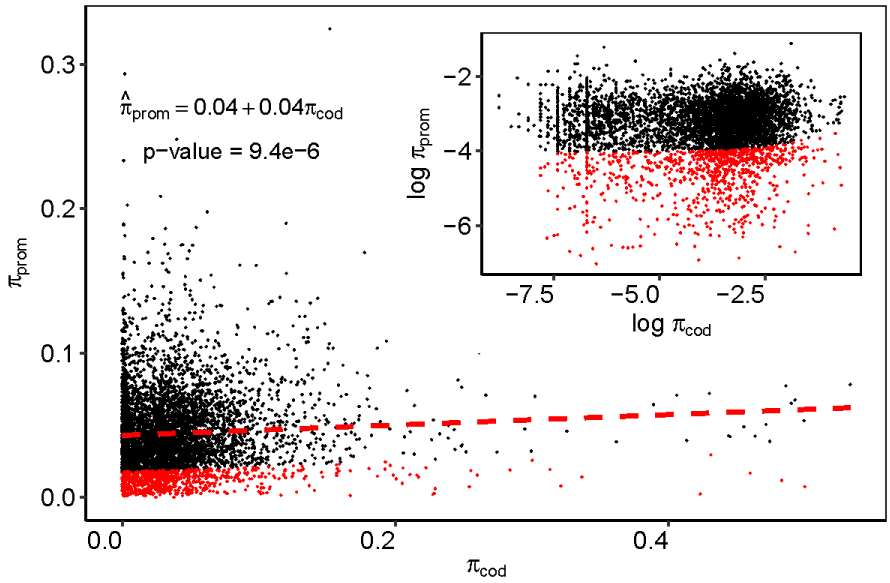

Figure 4.3 Linear relation between coding region and promoter region (TSS±1*kb*) nucleotide diversity. Each dot represents one of 5,430 genes. The red dashed line is the fitted regression line according to the model in (2). The red dots are genes with the bottom 10% regression residuals indicating the highest level of relative promoter constraint. Logarithm transformed scatter plot is embedded in the right upper corner, and genes with either $\boldsymbol{\pi_{prom}}$or $\boldsymbol{\pi_{cod}}$ or both equal zero are removed.

4.3.3 Genes with low relative polymorphism of promoter to coding are intolerant for rare regulatory variants

Next, I asked whether rare regulatory variants are depleted in genes with relative low promoter polymorphism. I performed one-tail unpaired student's *t*-tests across all tissues. Genes in the bottom 10% of the residual of regression (1) are significantly depleted for rare promoter region (within 4kb of TSS) alleles (MAF<0.05) that are also in the top and bottom 5% of expression bins. This significant regulatory variant depletion was observed for all 49 tissues (Table 4.2).

Table 4.2 p-value of unpaired one tail student's *t*-test, contrasting the mean number of rare alleles (MAF<0.05) in top and bottom 5% extreme expressed individuals against all other genes also in the bottom 10% residual of fitting equation 2.

Tissue	p-value	Tissue	p-value
Adipose subcutaneous	2.40e-12	Esophagus gastroesophageal junction	7.06e-14
Adipose visceral omentum	1.12e-11	Esophagus mucosa	6.28e-16
Adrenal gland	6.59e-17	Esophagus muscularis	4.73e-15
Artery aorta	4.00e-16	Heart atrial appendage	1.34e-13
Artery coronary	1.10e-14	Heart left ventricle	1.19e-15
Artery tibial	6.36e-17	Kidney cortex	1.49e-12
Brain amygdala	5.51e-18	Liver	2.58e-16
Brain anterior cingulate cortex BA24	8.12e-19	Lung	6.03e-18
Brain caudate basal ganglia	3.74e-18	Minor salivary gland	1.62e-08
Brain cerebellar hemisphere	2.83e-19	Muscle skeletal	3.64e-13
Brain cerebellum	3.37e-18	Nerve tibial	8.57e-19
Brain cortex	4.77e-15	Ovary	2.30e-13
Brain frontal fortex BA9	4.77e-20	Pancreas	4.49e-10
Brain hippocampus	6.57e-18	Pituitary	2.08e-17
Brain hypothalamus	1.19e-16	Prostate	2.75e-14
Brain nucleus accumbens basal ganglia	4.86e-18	Skin not sun exposed suprapubic	1.31e-18
Brain putamen basal ganglia	1.37e-07	Skin sun exposed lower leg	1.85e-20
Brain spinal cord cervical c-1	2.38e-11	Small intestine terminal ileum	2.13e-14
Brain substantianigra	2.97e-10	Spleen	5.90e-16
Breast mammary tissue	6.09e-21	Stomach	3.74e-15
Cells cultured fibroblasts	2.11e-17	Testis	2.39e-28
Cells EBV-transformed lymphocytes	4.73e-08	Thyroid	4.72e-18
Colon sigmoid	4.20e-14	Uterus	7.16e-11
Colon transverse	1.61e-14	Vagina	5.39e-16
		Whole blood	2.09e-19

4.3.4 A novel gene classification based on tissue-specific regulatory tolerance

Zhao et, al observed that genes not represented on the MetaboChip are significantly more likely to be enriched for rare regulatory variants[102]. Here I propose that disease-associated genes or highly conserved genes are depleted for rare regulatory variants in promoter regions. Since variants with MAF = 0.01-0.05 are not strongly subject to

purifying selection, I chose a more constrained see of rare variants with MAF < 0.01 for the following gene classification. By plotting the residual of the fitting of π_{cod} to π_{prom} against the ratio of rare variants in either the bottom or top 5% of extreme expression individuals to the total number of rare variants in the promoter region, I divided genes into four quadrants: set I, low residual and low ratio; set II, low residual and high ratio; set III, high residual and high ratio; set IV, high residual and low ratio. The decision boundary reported here is $residual = -0.025$ and $ratio = \frac{\overline{ratio}}{2}$ but similar results are seen with residual = 0 and the mean ratio. I hypothesized that set I genes would be depleted of regulatory rare variants that are selected against and functionally conserved in one or more tissue.

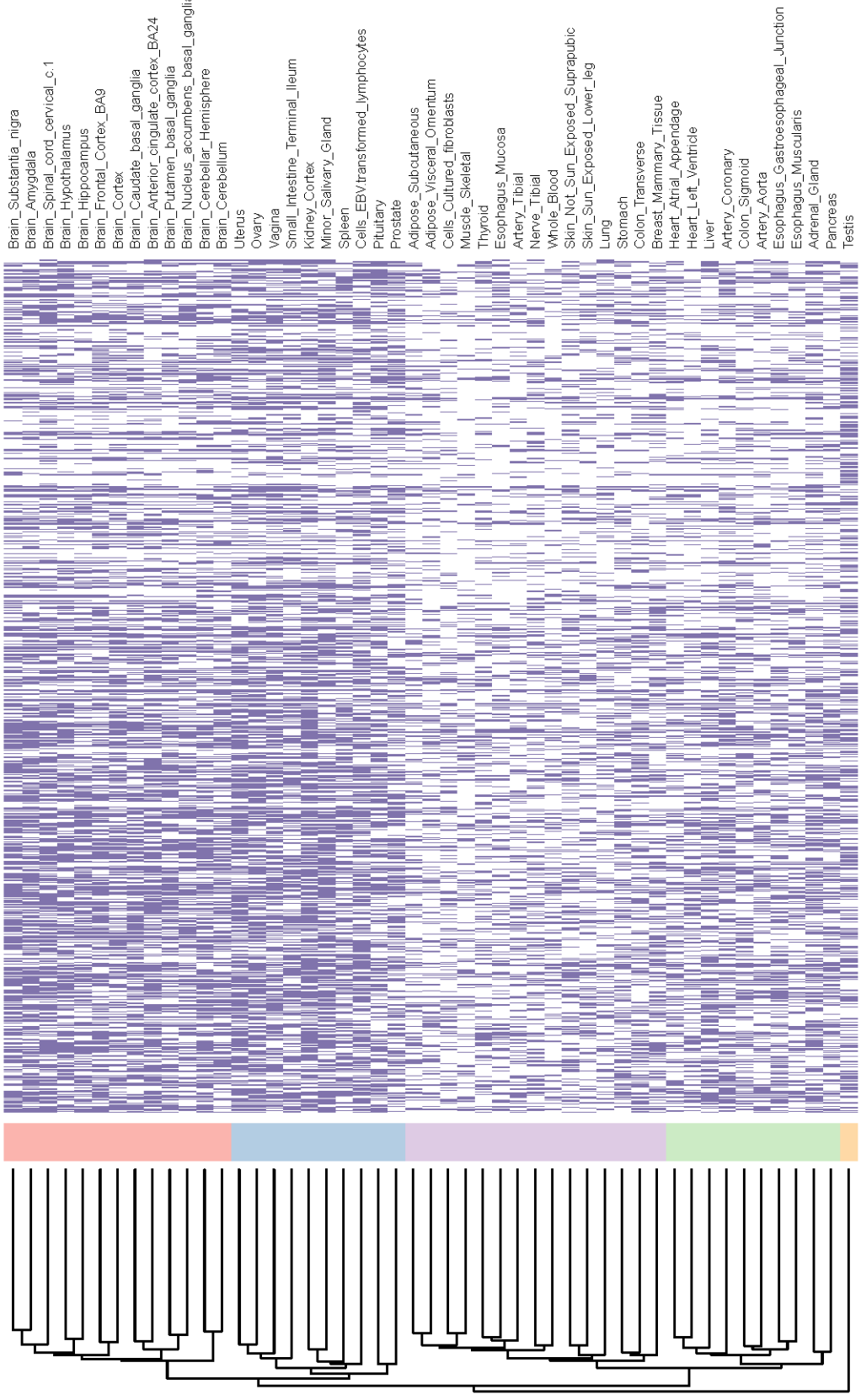
Brain_Substantia_nigra
Brain_Amygdala
Brain_Spinal_cord_cervical_c.1
Brain_Hypothalamus
Brain_Hippocampus
Brain_Frontal_Cortex_BA9
Brain_Cortex
Brain_Caudate_basal_ganglia
Brain_Anterior_cingulate_cortex_BA24
Brain_Putamen_basal_ganglia
Brain_Nucleus_accumbens_basal_ganglia
Brain_Cerebellar_Hemisphere
Brain_Cerebellum
Uterus
Ovary
Vagina
Small_Intestine_Terminal_Ileum
Kidney_Cortex
Minor_Salivary_Gland
Spleen
Cells_EBV.transformed_lymphocytes
Pituitary
Prostate
Adipose_Subcutaneous
Adipose_Visceral_Omentum
Cells_Cultured_fibroblasts
Muscle_Skeletal
Thyroid
Esophagus_Mucosa
Artery_Tibial
Nerve_Tibial
Whole_Blood
Skin_Not_Sun_Exposed_Suprapubic
Skin_Sun_Exposed_Lower_leg
Lung
Stomach
Colon_Transverse
Breast_Mammary_Tissue
Heart_Atrial_Appendage
Heart_Left_Ventricle
Liver
Artery_Coronary
Colon_Sigmoid
Artery_Aorta
Esophagus_Gastroesophageal_Junction
Esophagus_Muscularis
Adrenal_Gland
Pancreas
Testis

Figure 4.4 Hierarchical clustering of tissues by the presence of set I genes. The presence and absence of genes in tissue is color coded by purple and white. Side color bar annotates tissue cluster B, E, O, I and T.

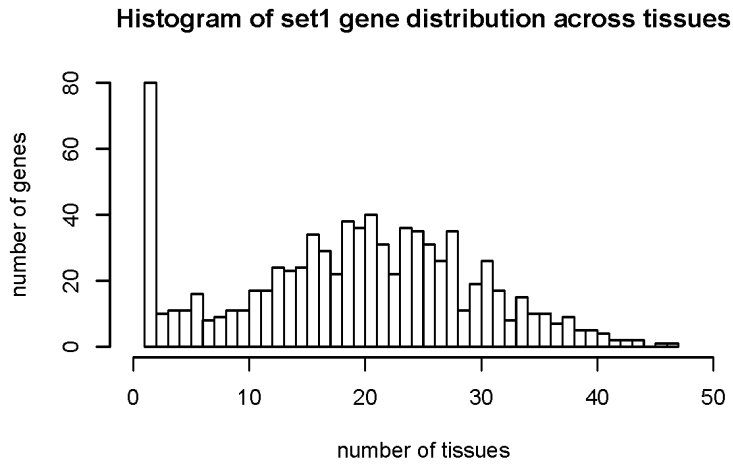

Figure 4.5 Histogram of set I gene distribution across tissues.

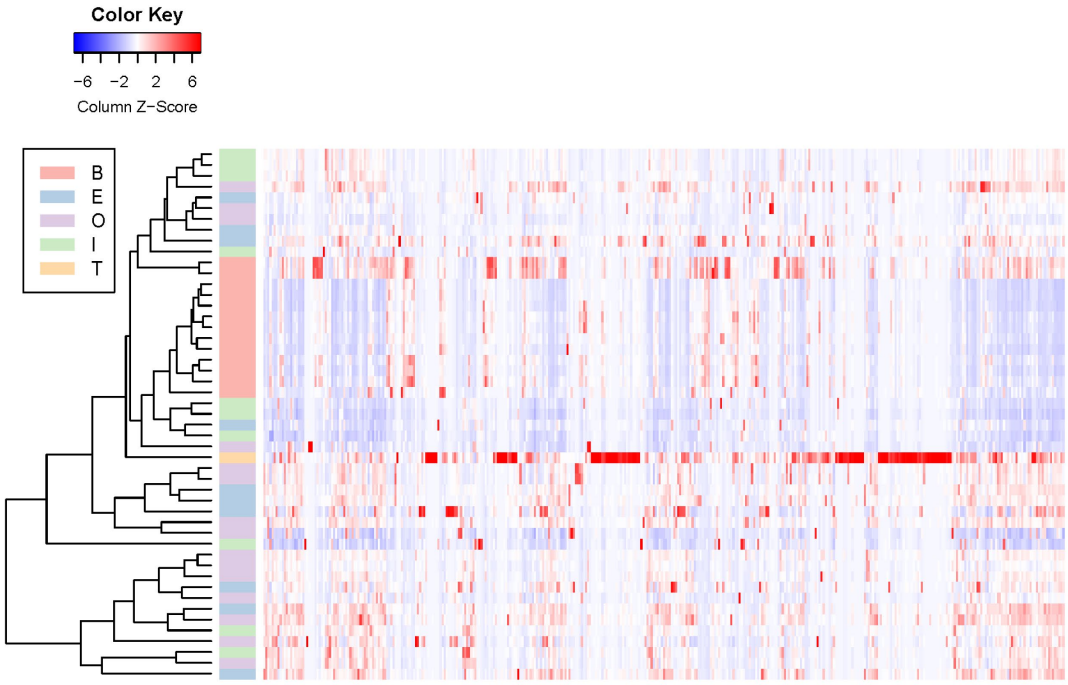

Figure 4.6 Heatmap of the medium expression of set I genes across tissues. Gene expression is normalized by row. Side color bar show five tissue clusters.

There were 841 genes across 49 tissues in set I. Each gene was coded as 0, not present, or 1, present in set I of a particular tissue. Tissues with 841 features are clustered by hierarchical clustering in Figure 4.4, which defines five tissue clusters: Brain (B), Endocrine (E), Internal (I), Other (O) and Testis (T). It is interesting that all of the brain tissues are clustered together in B which has the most genes, consistent with selection against regulatory variation in genes that play key roles in normal brain function. "Testis" is an outlier, indicating that regulatory constraint is independent of all other cluster. Set I genes are observed on average in approximately 20 tissues, while 55 genes are tissue-specific (Figure 4.5). The set I genes have relatively low expression in cluster B, are highly expressed in testis, and vary in abundance in the other clusters (Figure 4.6).

4.3.5 Comparison with other tolerance scores

Two other widely used genetic scores, pLI[123] and RVIS[125], quantify the probability of intolerance to mutation. I also compared the pLI and RVIS score of genes that are enriched in 80% of tissues in each tissue cluster, with the exception of testis. Kruskal-Wallis non-parametric ANOVA tests were applied, none of which showed significant co-occurrence of different groups of enriched genes, implying that my classification is orthogonal to the existing scores. A possible reason is that both of them utilize genetic variation in the coding region for computation of the scores. A fourth score was recently described by the same group that developed RVIS, designed to identify tolerance to variation in distal enhancer elements, which reported a similar result. The observation that my classification by the relative polymorphism of promoter region and coding region, and depletion of regulatory rare variants, leads to meaningful clustering by tissue type implies there is novel potential for annotating tissue-specific likelihood of regulatory effects.

4.4 Discussion

In this chapter, I demonstrate that rare variants in the promoter region are regulatory and have large effects leading to enrichment in individuals with extreme values of gene expression. Integrating these observations with measures of nucleotide diversity, I next demonstrate that rare regulatory alleles with large effect are depleted in genes that are evolutionary constrained specifically in the promoter region. Filtering genes with relatively low promoter nucleotide polymorphism and rare variant depletion by tissue leads to categorization of tolerance to mutation which appears to be organized into functional tissue groups.

Other sets of genes may also be interesting to investigate. For example, set II genes are evolutionary constrained in promoter regions but have relative high abundance of rare regulatory variants of large effect. Those rare variants maybe not selected against, and are less likely to show association with disease risk. Another interesting analysis would be family-based studies of rare diseases, evaluating segregation of rare risk alleles with large effect size.

In terms of relative expression level of set I genes in tissues, it is interesting that brain tissues have relatively low expression levels, whereas testis has the most abundant transcript abundance. I did not explore the possible reasons here, but it would be worthwhile to analyze case/control expression data of neuronal diseases and ask if those enriched genes are differentially expressed genes in specific disease contexts.

In summary, the study develops a unique tissue-gene categorization system based on quantifying the genic intolerance to rare variants that have large regulatory effect in

promoter regions. As more datasets associating gene expression and rare genetic variation with disease appear, it should be possible to develop a validated tool that prioritizes regulatory variants for experimental validation of their role in pathogenesis using the methods developed in Chapters 2 and 3.

CHAPTER 5. CONCLUSION AND DISCUSSION

This thesis focuses on experimental and computational functional validation of disease-associated risk variants, which is a major challenge of human genetics in the post-GWAS era. My contributions include design of a CRISPR/Cas9 based regulatory variant fine-mapping approach, and development of a gene-tissue categorization system to prioritize pathogenic rare variants that are candidates for genetic screening.

Firstly, I performed a regulatory variant identification assay in CRISPR/Cas9 genome edited single-cell clones. I observed constraints from the variability in mutability of targeted regulatory sites; clonal variability; high expense and low statistical power. This led me to conclude that this single-cell clone based method is not recommended for scanning within credible intervals of 100 or more SNPs. Secondly, I designed an expression CROP-seq method, a CRISPR/Cas9 pooled genomic screening with massively parallel single-cell RNA-seq as readout. With expression CROP-seq, mutagenesis was introduced simultaneously targeting 60 SNPs of interest to cells with the same genetic background, and transcript abundance was quantified at the same time by robust single-cell RNA-seq. It largely reduced the impact of some of the sources of variability found in the previous method, and greatly enhanced statistical power. Expression CROP-seq comprehensively fine-maps eSNPs within credible intervals that explain just a few percent of the expression of a target gene. A direct outcome of utilizing this approach was the successful identification of two causal regulatory SNPs that altered IBD associated risk gene expression, which also reside in open chromatin regions. Subsequently, I aimed to extend the application of this approach to rare variants screening. I categorized genes by their

tolerance of promoter region and quantity of rare regulatory variants with large effect across 49 tissues with GTEx v8 dataset. Genes constrained by promoter polymorphism and having deficit of rare regulatory variants are clustered by tissue, which supports the use of tissue-gene genomic annotation for prioritization of GWAS tagged risk loci in functional validation.

Some limitations of CRISPR/Cas9 technology bring up constraints of the demonstrated screening approaches. The first constraint is accessibility. The Protospacer Adjacent Motif (PAM) sequence (NGG) should be always 3-4 bp downstream of the cut site. In my gRNA design, the cut site was restricted to less than 10 base away from target SNP, yet even then not all of the SNPs within a Bayesian credible set were editable. For example, the top SNP of the primary signal of *CISD1*, rs146577551, is not targetable. Thus, the important finding that multiple independent eQTL regulates gene expression[55,56,105,129] cannot be tested in for all eGenes. The second issue is off-target effects. I chose gRNAs with minimal predicted off-target sites throughout entire genome. Off-target effect can be minimized by improving off-target prediction algorithms in silico, using optimized reagents and controlling Cas9 delivery time[130,131]. Even though off-target effects are a major concern in gene therapy, they should also be considered in genetic screening applications. The third constraint is DNA repair pathway. To enhance editing efficiency, I chose non-homologous end joining (NHEJ) based pathway to introduce microindels to the cut site, instead of homology-directed repair (HDR). In theory, NHEJ reaches high editing efficiency up to 20%-60%[24], while HDR is only 0.5%-20%. Even though it would be ideal to replace the reference allele with an alternate allele by HDR with an introduction of template DNA, I utilized NHEJ out of consideration of editing efficiency and power. In

my case, NHEJ achieved incredibly high editing efficiency, 90% on average. Causal eSNPs identified by expression CROP-seq are thought to alter gene expression by changing transcription factors and regulatory element binding affinity. eSNPs altering gene expression by other mechanisms would not be amenable to be identified by expression CROP-seq. An emerging technique in CRISPR/Cas9 field, prime editing[121], may help to improve precision. Prime editing obtains 20%-50% editing efficiency, comparable with NHEJ. An advantage of prime editing is that it is able to mediate targeted insertions, deletions and all 12 base-base conversions, including four transition mutations and eight transversion, without introducing double strand breaks and template DNA. Thus, prime editing both achieves high editing efficiency and replaces alleles on targeted sites, which could be a modification for expression CROP-seq.

Even though targeted genes I selected were all with medium to high expression in HL60/S4, and the identified two causal SNPs downregulated target gene expression after editing, I expect that expression CROP-seq is also able to identify those SNPs with upregulatory effects. This would be an advantage over MPRA[33], which only detects regulatory elements that are required to increase reporter expression. Another advantage of CROP-seq is that the gRNA sequence itself serves as barcode. Other pooled CRISPR single-cell RNA-seq screening methods, like Perturb-seq[40], all rely on sequencing the barcode to identify individual cells. A potential problem is misidentified cell edits caused by frequent recombination of barcode and gRNA. The barcode is always designed to be several kilobases from gRNA. During viral infection, it is feasible that barcode mismatches with the unpaired gRNA, with a recombination rate that may be as high as 50%. Researchers need to pay extra attention when applying such methods[132].

Cancer genomes can be highly instable. For example, some cancer cell lines have big loads of mutations and copy number variation. If SNPs of interest have more than two copies, the effect size can be overestimated in the cell line relative to those estimated in an eQTL study. One the other hand, if the SNP of interest has only one copy or zero copies in a cell, the estimated effect size is underestimated or completely undetectable in the cell. Some screening results may be biased.

As previously discussed, a meaningful proportion of eQTL are cell-type and tissue-specific. The importance of validating eSNPs in relevant cell types is now clear, which also helps to explain the nature of GWAS signals. Applying expression CROP-seq in primary T cells should generate fruitful results for immune associated eQTLs, for example. One of the major concerns is the potential variability of genetic background in different batches of primary T cells. Since they are isolated from individuals, there might be variation introduced by the transcriptional background. It will be necessary to estimate such variation before genome editing. On the other hand, we can also treat it as an advantage for the application of disease specific loci screening. For example, targeted SNPs in primary T cells from IBD patients and healthy controls may have different effects. This application will be able to identify context dependent regulatory SNPs, which might be associated with the pathogenesis of autoimmune diseases. A successful study using CROP-seq in primary T cell is available for reference[133].

Gene-tissue classification can also be extended for population specific categorizing systems. In chapter 4, only European individuals were included in my analysis, and about 100 non-Europeans were excluded due to low sample size. Rare alleles exhibit variable

effect contributions to rare diseases among populations[134]. The analysis could be applied to other populations with sufficient divergent genomic data. [135].

RIVER improves rare regulatory variants prediction by using both outlier gene expression information and genomic annotation as priors[122]. And both priors are evaluated as an average across tissues, instead of considering individual tissue. My categorizing system emphasizes that the intolerance of promoter and depletion of rare regulatory variants of genes can be shared among tissues, and also be observed in single tissues. This system can be served as a resource to update existing gene constraint scores with tissue information, like pLI[123], RVIS[124,125]. The next step will be focusing on the prediction of tissue-specific individual pathogenic rare variants by adding regulatory effects that result in extreme expression. The combination of both tissue and regulatory information will inevitably increase the accuracy and provide tissue functional information in genetic screening of rare variants.

REFERENCES

1 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001).

2 Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-389 (2005).

3 DeWan, A. *et al.* HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* **314**, 989-992 (2006).

4 Yang, Z. *et al.* A variant of the HTRA1 gene increases susceptibility to age-related macular degeneration. *Science* **314**, 992-993 (2006).

5 McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356-369 (2008).

6 Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

7 Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43**, 1066-1073 (2011).

8 Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Phil. Trans. R. Soc. B* **368**, 20120362 (2013).

9 Lloyd-Jones, L. R. *et al.* The genetic architecture of gene expression in peripheral blood. *Am. J. Hum. Genet.* **100**, 228-237 (2017).

10 Göring, H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* **39**, 1208-1216 (2007).

11 Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science* **325**, 1246-1250 (2009).

12 Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519-523 (2014).

13 Kasela, S. *et al.* Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. *PLoS Genet.* **13**, e1006643 (2017).

14 Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580-585 (2013).

15 Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*, 787903 (2019).

16 Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167-170 (2010).

17 Jinek, M. *et al.* A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).

18 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823 (2013).

19 Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233 (2013).

20 Jacobs, J. Z., Ciccaglione, K. M., Tournier, V. & Zaratiegui, M. Implementation of the CRISPR-Cas9 system in fission yeast. *Nat. Commun.* **5**, 5344 (2014).

21 Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 227 (2013).

22 Shen, B. *et al.* Generation of gene-modified mice via Cas9/RNA-mediated gene targeting. *Cell Res.* **23**, 720 (2013).

23 Belhaj, K., Chaparro-Garcia, A., Kamoun, S. & Nekrasov, V. Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR/Cas system. *Plant Methods* **9**, 39 (2013).

24 Maruyama, T. *et al.* Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nat. Biotechnol.* **33**, 538-542 (2015).

25 Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822 (2013).

26 Cradick, T. J., Qiu, P., Lee, C. M., Fine, E. J. & Bao, G. COSMID: a web-based tool for identifying and validating CRISPR/Cas off-target sites. *Mol. Ther. Nucleic Acids* **3** (2014).

27 Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187 (2015).

28 Wang, X. *et al.* Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nat. Biotechnol.* **33**, 175 (2015).

29 Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521-530, doi:10.1038/nbt.2205 (2012).

30 Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome research* **23**, 800-811 (2013).

31 Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519-1529 (2016).

32 van Arensbergen, J. *et al.* High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* **51**, 1160 (2019).

33 Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271-277, doi:10.1038/nbt.2137 (2012).

34 Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014).

35 Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167-174 (2016).

36 Fei, T. *et al.* Deciphering essential cistromes using genome-wide CRISPR screens. *Proc. Natl. Acad. Sci.* **116**, 25186-25195 (2019).

37 Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* **34**, 192 (2016).

38 Gasperini, M. *et al.* A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377-390. e319 (2019).

39 Adamson, B. *et al.* A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867-1882. e1821 (2016).

40 Dixit, A. *et al.* Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853-1866. e1817 (2016).

41 Jaitin, D. A. *et al.* Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883-1896. e1815 (2016).

42 Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297 (2017).

43 Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol. Cell* **66**, 285-299. e285 (2017).

44 Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40**, 638 (2008).

45 Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5-22 (2017).

46 Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare Variants Create Synthetic Genome-Wide Associations. *PLoS Biol.* **8**, e1000294, doi:10.1371/journal.pbio.1000294 (2010).

47 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).

48 Wright, C. F. *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet. Med.* **20**, 1216-1223 (2018).

49 Farnaes, L. *et al.* Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *NPJ GENOM MED* **3**, 1-8 (2018).

50 Kingsmore, S. F. *et al.* A Randomized, Controlled Trial of the Analytic and Diagnostic Performance of Singleton and Trio, Rapid Genome and Exome Sequencing in Ill Infants. *Am. J. Hum. Genet.* **105**, 719-733 (2019).

51 Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).

52 Frésard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat. Med.* **25**, 911-919, doi:10.1038/s41591-019-0457-8 (2019).

53 Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881-888 (2008).

54 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci.* **106**, 9362-9367 (2009).

55 Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173 (2017).

56 Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505 (2018).

57 Zeng, B. *et al.* Comprehensive Multiple eQTL Detection and Its Application to GWAS Interpretation. *Genetics* **212**, 302091 (2019).

58 Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535-552 (2014).

59 Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236 (2015).

60 Li, M. J. *et al.* Predicting regulatory variants with composite statistic. *Bioinformatics* **32**, 2729-2736 (2016).

61 Liu, L. *et al.* Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat. Commun.* **10**, 330 (2019).

62 Huo, Y., Li, S., Liu, J., Li, X. & Luo, X.-J. Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. *Nat. Commun.* **10**, 670 (2019).

63 Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246 (2011).

64 Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119 (2012).

65 Wang, Y. *et al.* Autocrine motility factor receptor promotes the proliferation of human acute monocytic leukemia THP-1 cells. *Int. J. Mol. Med.* **36**, 627-632 (2015).

66 Czimmerer, Z. *et al.* Identification of novel markers of alternative activation and potential endogenous PPARγ ligand production mechanisms in human IL-4 stimulated differentiating macrophages. *Immunobiology* **217**, 1301-1314 (2012).

67 Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv*, 447367, doi:10.1101/447367 (2018).

68 Fairfax, B. P. *et al.* Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. *Science* **343**, 1246949, doi:10.1126/science.1246949 (2014).

69 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310 (2014).

70 Liu, L., Tamura, K., Sanderford, M., Gray, V. E. & Kumar, S. A molecular evolutionary reference for the human variome. *Mol. Biol. Evol.* **33**, 245-254 (2015).

71 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308-311 (2001).

72 Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61-D65 (2006).

73 Cradick, T. J., Qiu, P., Lee, C. M., Fine, E. J. & Bao, G. COSMID: a web-based tool for identifying and validating CRISPR/Cas off-target sites. *Mol. Ther. Nucleic Acids* **3**, e214 (2014).

74 Breitman, T., Selonick, S. E. & Collins, S. J. Induction of differentiation of the human promyelocytic leukemia cell line (HL-60) by retinoic acid. *Proc. Natl. Acad. Sci.* **77**, 2936-2940 (1980).

75 Okazaki, T., Bell, R. M. & Hannun, Y. A. Sphingomyelin turnover induced by vitamin D3 in HL-60 cells. Role in cell differentiation. *J. Biol. Chem.* **264**, 19076-19080 (1989).

76 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

77 Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).

78 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

79 Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305-W311 (2009).

80 Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

81 Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-W97 (2016).

82 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303 (2010).

83 Van der Auwera, G. A. *et al.* From FastQ data to high‐confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43**, 11.10. 11-11.10. 33 (2013).

84 Preininger, M. *et al.* Blood-informative transcripts define nine common axes of peripheral blood gene expression. *PLoS Genet.* **9**, e1003362 (2013).

85 Sanchez-Freire, V., Ebert, A. D., Kalisky, T., Quake, S. R. & Wu, J. C. Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns. *Nat. Protoc.* **7**, 829-838, doi:10.1038/nprot.2012.021 (2012).

86 Coudray-Meunier, C. *et al.* A Novel High-Throughput Method for Molecular Detection of Human Pathogenic Viruses Using a Nanofluidic Real-Time PCR System. *PloS One* **11**, e0147832-e0147832, doi:10.1371/journal.pone.0147832 (2016).

87 Jang, J. S. *et al.* Quantitative miRNA Expression Analysis Using Fluidigm Microfluidics Dynamic Arrays. *BMC Genom.* **12**, 144, doi:10.1186/1471-2164-12-144 (2011).

88 Collins, S. J., Gallo, R. C. & Gallagher, R. E. Continuous growth and differentiation of human myeloid leukaemic cells in suspension culture. *Nature* **270**, 347 (1977).

89 Gallagher, R. *et al.* Characterization of the continuous, differentiating myeloid cell line (HL-60) from a patient with acute promyelocytic leukemia. *Blood* **54**, 713-733 (1979).

90 Leung, M.-F., Sokoloski, J. A. & Sartorelli, A. C. Changes in microtubules, microtubule-associated proteins, and intermediate filaments during the differentiation of HL-60 leukemia cells. *Cancer Res.* **52**, 949-954 (1992).

91 Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325 (2018).

92 Ramirez, R. N. *et al.* Dynamic gene regulatory networks of human myeloid differentiation. *Cell Syst.* **4**, 416-429. e413 (2017).

93 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886-D894 (2018).

94 Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167 (2016).

95 Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111 (2017).

96 Pan, Y., Tian, R., Lee, C. M., Bao, G. & Gibson, G. Fine-mapping within eQTL Credible Intervals by Expression CROP-seq *Biol. Meth. Protoc.* (2019).

97 Kreimer, A. *et al.* Predicting gene expression in massively parallel reporter assays: a comparative study. *Hum. Mutat.* **38**, 1240-1250 (2017).

98 Kalita, C. A. *et al.* High-throughput characterization of genetic effects on DNA–protein binding and gene transcription. *Genome Res.* **28**, 1701-1708 (2018).

99 Norga, K. K. *et al.* Quantitative analysis of bristle number in Drosophila mutants identifies genes involved in neural development. *Curr. Biol.* **13**, 1388-1396 (2003).

100 Magwire, M. M. *et al.* Quantitative and molecular genetic analyses of mutations increasing Drosophila life span. *PLoS Genet.* **6**, e1001037 (2010).

101 Metzger, B. P., Yuan, D. C., Gruber, J. D., Duveau, F. & Wittkopp, P. J. Selection on noise constrains variation in a eukaryotic promoter. *Nature* **521**, 344 (2015).

102 Zhao, J. *et al.* A burden of rare variants associated with extremes of gene expression in human peripheral blood. *Am. J. Hum. Genet.* **98**, 299-309 (2016).

103 van Overbeek, M. *et al.* DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. *Mol. Cell* **63**, 633-646 (2016).

104 Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396-1400 (2019).

105 Zeng, B. *et al.* Comprehensive Multiple eQTL Detection and Its Application to GWAS Interpretation. *Genetics*, genetics. 302091.302019 (2019).

106 Sanjana, N. E. *et al.* High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545-1549 (2016).

107 Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783 (2014).

108 Walter, D. M. *et al.* Systematic in vivo inactivation of chromatin-regulating enzymes identifies Setd2 as a potent tumor suppressor in lung adenocarcinoma. *Cancer Res.* **77**, 1719-1729 (2017).

109 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* (2019).

110 Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38 (2019).

111 Hsiau, T. *et al.* Inference of CRISPR Edits from Sanger Trace Data. *bioRxiv*, 251082, doi:10.1101/251082 (2019).

112 Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-1101 (2015).

113 Ramirez, R. N. *et al.* Dynamic Gene Regulatory Networks of Human Myeloid Differentiation. *Cell Syst.* **4**, 416-429.e413, doi:https://doi.org/10.1016/j.cels.2017.03.005 (2017).

114 Lin, C.-C. *et al.* Bhlhe40 controls cytokine production by T cells and is essential for pathogenicity in autoimmune neuroinflammation. *Nat. Commun.* **5**, 3551, doi:10.1038/ncomms4551 (2014).

115 Livak, K. J. & Schmittgen, T. D. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* **25**, 402-408, doi:https://doi.org/10.1006/meth.2001.1262 (2001).

116 Wang, H., La Russa, M. & Qi, L. S. CRISPR/Cas9 in genome editing and beyond. *Annu. Rev. Biochem.* **85**, 227-264 (2016).

117 de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256-261, doi:10.1038/ng.3760 (2017).

118 McCole, D. F. IBD Candidate Genes and Intestinal Barrier Regulation. *Inflamm. Bowel Dis.* **20**, 1829-1849, doi:10.1097/mib.0000000000000090 (2014).

119 Krebiehl, G. *et al.* Reduced basal autophagy and impaired mitochondrial dynamics due to loss of Parkinson's disease-associated protein DJ-1. *PloS One* **5** (2010).

120 Graham, D. B. & Xavier, R. J. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature* **578**, 527-539, doi:10.1038/s41586-020-2025-2 (2020).

121 Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149-157, doi:10.1038/s41586-019-1711-4 (2019).

122 Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239 (2017).

123 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285 (2016).

124 Petrovski, S. *et al.* The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet.* **11**, e1005492 (2015).

125 Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).

126 Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*, 787903, doi:10.1101/787903 (2019).

127 Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* **76**, 5269-5273, doi:10.1073/pnas.76.10.5269 (1979).

128 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

129 Zeng, B. *et al.* Constraints on eQTL Fine Mapping in the Presence of Multisite Local Regulation of Gene Expression. *G3-Genes Genomes Genetics* **7**, 2533-2544 (2017).

130 Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184-191, doi:10.1038/nbt.3437 (2016).

131 Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S. & Yang, S.-H. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Mol. Ther. Nucleic Acids* **4**, e264, doi:https://doi.org/10.1038/mtna.2015.37 (2015).

132 Xie, S., Cooley, A., Armendariz, D., Zhou, P. & Hon, G. C. Frequent sgRNA-barcode recombination in single-cell perturbation assays. *PLoS One* **13**, e0198635, doi:10.1371/journal.pone.0198635 (2018).

133 Shifrut, E. *et al.* Genome-wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function. *Cell* **175**, 1958-1971.e1915, doi:https://doi.org/10.1016/j.cell.2018.10.024 (2018).

134 Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci.* **108**, 11983, doi:10.1073/pnas.1019276108 (2011).

135 Bustamante, C. D., De La Vega, F. M. & Burchard, E. G. Genomics for the world. *Nature* **475**, 163-165, doi:10.1038/475163a (2011).