THE UNIVERSITY OF CHICAGO

DISSECTING THE GENETIC BASIS OF HUMAN IMMUNE-RELATED DISEASES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY
ZEPENG MU

CHICAGO, ILLINOIS
MARCH 2024

To my mother and father.

Where is the Life we have lost in living?

Where is the wisdom we have lost in knowledge?

Where is the knowledge we have lost in information?

–T. S. Eliot, the Rock (1934)

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

Rarely in our lifetime do we have an opportunity to reflect on all the people that have helped and supported us, in small and big ways, to bring us to where we are today. As I reach the end of this five-year journey, I would like to take this opportunity to express my gratitude toward everyone who has had or is still having a substantial and meaningful impact on my education and life.

First, I must thank my advisor, Dr. Yang Li for the unconditional support and excellent training that I received from him. I have learned so much from you, not only on doing daily research, but also on ways of thinking and analyzing scientific questions in a critical way, and not to follow the path walked by others, but to always follow the questions we care. Thank you for giving me the space and trusting me to freely explore my sometimes not-so-clear passions and interests, allowing me to do things more independently. I would also like to thank all current and past members of the Li Lab for the supportive and thought-stimulating environment they created.

Thank you to my Thesis Committee members, Drs. Bana Jabri, Matthew Stephens, and Luis Barreiro for offering guidance and feedback on my work. I am especially grateful to Dr. Bana Jabri, for supporting my immature ideas on the interactions between regulatory T cells and $CD8^+$ effector T cells, and for allowing me to learn immunological experiments from the very beginning in her lab. Thank you to Megan Borregard, for helping me to refine and crystallize my hypothesis and come up with actionable experimental plans and, most importantly, teaching me with all the patience how to do the immunology experiments and spending long hours with me in front of the bench. These knowledge and skills will be useful to me forever.

Thank you to Sue Levinson for always helping me in every way possible and making my graduate life much easier. To the Section of Genetic Medicine community: first, thank you to Dr. Yoav Gilad for allowing me to attend your lab meeting, to learn and exchange

ideas with the members of your lab. Thank you for creating an open environment that allowed me to learn from people that I would otherwise have no chance to know. Thank you to the people in this community that I can learn from and also have fun together, especially Wenhe Lin, Lili Wang, HyunKyung Kim and Dr. Jinghui Li. To administrators in the Section of Genetic Medicine, Tamiko Charley and Sandra Dantzler, thank you for always helping me in countless ways.

My journey as a student in bioinformatics and human genetics started before my graduate school thanks to the International Student at Large (ISAL) program, here at the University of Chicago with Dr. Eileen Dolan. Thank you for accepting me in your lab when I was still an undergraduate knowing little about computational biology. Thank you to Omar El-Charif, who directly taught me computational skills during my time at the Dolan Lab. Through the several months in the lab, I gained hands-on experience in doing computational research and became fully confident in my ability in pursuing a Ph.D. degree in this field.

To my parents and grandparents, thank you for unconditionally support my choice and decision in every step during my growth and education. Your love and support transcend physical distance and I can constant feel it thousands of miles away on this side of the globe. Thank you to my twin brother, Zekun Mu, for always being together with me along the journey, especially during the pandemic when many international students cannot meet their family. Thank you for having hours-long video calls with me, during which you often remind me of how little I know about immunology, so that I always know I need to do better.

ABSTRACT

In the past 15 years, genome-wide association studies (GWAS) have identified thousands of genetic variants associated with a plethora of human complex traits, and greatly expanded our knowledge on their genetic architectures. Around 90% of GWAS variants are located in noncoding regions, suggesting that they may have regulatory effects. However, statistical integration of GWAS and expression quantitative trait loci (eQTL) can only explain ~25% of GWAS variants, highlighting the challenges in uncovering the molecular mechanisms underlying GWAS signals. Here, we show that genetic variants affecting pre-mRNA splicing (sQTL) can pinpoint GWAS loci missed in eQTL. Together, eQTL and sQTL explain 40% of GWAS variants. We demonstrate that primary, steady-state eQTL and sQTL are largely shared among immune cell types in peripheral blood, suggesting that it is challenging to pinpoint the exact cell type in which these shared QTL mediate disease risk. I also show that disease samples from the correct tissue can capture regulatory effects distinct from healthy or *in vitro* stimulated immune cells. We then investigate the regulatory effects on chromatin accessibility and relevance to disease. We find that chromatin accessibility (caQTL) greatly complement eQTL/sQTL in explaining disease GWAS. Through extensive comparisons with caQTL, eQTL and GWAS, I conclude that partially due to enhancer pleiotropy, even though caQTL can correctly reveal the effect of disease variants on chromatin accessibility, it is much harder to pinpoint the true causal gene and causal context when a caQTL has no effect on gene expression. This highlights the crucial importance to study more disease-relevant contexts and integrate caQTL with multi-modal annotation data for deeper understanding of the causal mechanisms of disease GWAS. Taken together, our results offer a more comprehensive understanding of the regulatory landscape of molecular phenotypes, especially for chromatin accessibility and gene expression, and their roles in the genetic predisposition of immune-related diseases. We highlight the complexity in layers of gene regulation cascade and provide a

more cautionary tale in the interpretation of molQTL-GWAS data. We believe our work form a basis for future studies to further extend the boundaries of our knowledge of how genetic variants can shape human phenotypes through regulation of a network of molecular features.

CHAPTER 1

INTRODUCTION

1.1 Mapping the genetic architecture of human complex diseases

The majority of human diseases—like many other complex traits—are jointly determined by genetic factors and environmental factors. The study of the genetic architecture of human diseases aims at determining the number and location of genetic variants impacting disease risk, and the joint distribution of their effect sizes and allele frequencies. It is well-established that the genetic architecture of human diseases falls into a wide range of distribution. Mendelian diseases are typically caused by variants in a single gene, hence also known as monogenic diseases. Earlier genetic studies based on linkage analysis in pedigrees have successfully identified causal genes for many monogenic diseases. Monogenic diseases are usually highly heritable, and have high penetrance rate, but they also tend to be less prevalent in the population due to negative selection on strongly deleterious mutations. At the other end of the spectrum are polygenic diseases, which can be jointly affected by hundreds to thousands of genetic variants together. Polygenic diseases are more common in population—rheumatoid arthritis (RA), for example, could affect ~0.5%-1% of the global population[1]. Although many polygenic diseases are not as deleterious or life-threatening as monogenic diseases, the systemic and chronic nature of these disease can nonetheless lead to significant years of life lost and reduction in patients' quality of life[2]

Mapping the genetic architecture of polygenic diseases proved to be challenging due to several reasons. First, it requires technology capable of identifying genetic variants across the whole genome at scale. Second, given that most genetic variants underlying complex diseases have too small effect sizes to be detected by the traditional linkage analysis[3], a large cohort of cases and controls is needed for sufficient power to map disease-

associated variants. In the past two decades, the construction of the reference genome pivoted by the Human Genome Project (HGP) and the steady decrease in the cost of genotyping by DNA microarray, and next-generation sequencing (NGS) have made it feasible to collect genotype data from huge cohorts. Statistical imputation algorithms can further improve the number of genetic variants that can be reliably genotyped and included in downstream analysis, especially for common single nucleotide polymorphisms (SNPs). As a result, the association between each genetic variant and disease status in a population can be tested genome-wide with a linear regression, while adjusting for confounding factors such as population admixture, gender and other demographic variables. Nowadays, a typical study can test anywhere between 5 to 10 million SNPs, depending on the sample size and genotyping platform. This framework, known as genome-wide association study (GWAS), has become the "paradigm" of mapping the heritable component of complex diseases. To date, there have been more than 6,000 publications, 550,000 lead associations documented in the GWAS Catalog[4].

Not only did the success of GWAS significantly expanded the catalogue of disease-associated variants, but it has also shed light on the genetic architecture of human diseases. One major discovery from GWAS in the past 15 years is that most complex diseases have a genetic component and are often polygenic, including metabolic diseases like type 2 diabetes, psychiatric diseases like schizophrenia, and immune-related diseases like rheumatoid arthritis, inflammatory bowel diseases and asthma, for which at least a hundred genome-wide significant SNPs have been found. Given the "missing heritability" problem—which posits that current GWAS signals only explain a fraction of total genetic variation—it is postulated that new disease-associated variants will continue to be identified as GWAS sample size grows[5]. Notably, as the prevalence of complex, chronic diseases steadily increases in the U.S. and around the globe[6], it is imperative to establish more complete understanding of the diseases mechanisms. Human genetics and GWAS offer

a unique opportunity to understand disease pathogenesis, which could in turn translate into better disease prediction, management, and treatment.

1.2 Elucidate causal mechanisms of noncoding GWAS variants with molecular QTL

Contrary to the ever-increasing number of significant variants identified in GWAS, very few of them have been successfully translated into biological discoveries, which entails knowing the causal variants, downstream causal genes and functional contexts. This gap in knowledge highlights the challenges in obtaining causal mechanisms from statistical associations, and can be partially attributed to the difficulty in linking GWAS variants to genes, due to at least three reasons. First, about 90% of GWAS variants lie in noncoding regions[7]. Even though this represents an enrichment in protein-coding sequences, which only account for ~2% of the entire human genome, the finding means that the vast majority of GWAS hits cannot be directly linked to a gene or changes in amino acid residues in protein products. Second, the success of GWAS relies on linkage disequilibrium (LD) between true causal SNPs and co-segregating SNPs in the population. Consequently, in most cases, the lead SNP identified by GWAS is not causal, but merely "tagging" the true causal SNP. Third, by testing the association between germline variants and a phenotype, GWAS is able to find trait-associated variants that function in virtually any context (i.e. cell types, cellular states, or developmental stages), which means that it is impossible to pinpoint the exact context in which their effects impact disease risk directly from GWAS results.

To address these issues, recent years have witnessed concerted efforts in developing strategies to study noncoding GWAS variants. One of the most important and fruitful approaches is the statistical integration of summary statistics from GWAS and molecular

quantitative trait loci (molQTL). Molecular QTL refers to the genetic variants that are associated with the levels of molecular features, such as gene expression, pre-mRNA splicing, chromatin accessibility, histone modification, DNA methylation at CpG sites, and protein levels[8–16]. Molecular QTL can be categorized into *cis*- or *trans*-QTL depending on the distance between a genetic variant and the genomic location of the tested feature, such as the transcription start site (TSS) of a gene. The majority of molQTL mapping studies to date have only focused on *cis*-QTL[1]. Mapping *trans*-QTL is extremely challenging because: (1) the vast number of SNP-feature pairs genome-wide leads to huge multiple testing burden; (2) the small effect sizes of *trans*-QTL require large sample size to detect. However, recent innovations in statistical methods and efforts to meta-analyze multiple QTL datasets can lead to biological discoveries in *trans*-QTL that greatly complement our current understanding of gene regulation landscape in the near future[17,18].

The rationale for integrating GWAS with molQTL comes from the observation that noncoding elements in the genome have regulatory functions, such as activating gene expression. Furthermore, studies have shown that 80.6% GWAS lead SNPs either lie directly within or are in close LD with SNPs in open chromatin regions[19], suggesting that GWAS SNPs might perturb the activity of regulatory elements. A more compelling line of evidence for the regulatory role of GWAS variants is that they are significantly enriched for expression QTL (eQTL) compared to randomly chosen variants with matched minor allele frequencies (MAF) in the genome[20].

Motivated by these discoveries, statistical methods have been developed to integrate GWAS and QTL, with the goal of nominating genes that can potentially mediate the disease risk conferred by genetic variants. This family of methods includes colocalization to test whether GWAS and QTL data share the same causal SNP[21], PrediXcan/TWAS to determine whether the genetic component of gene expression is associated with disease

1. For the sake of simplicity, all references to in this thesis QTL means *cis*-QTL unless specified otherwise

status[22,23], and Mendelian Randomization (MR) to evaluate the causal relationship between gene expression and disease[24].

1.3 The success and limitation of eQTL

There have been many large-scale studies to collect both genotype and gene expression data from a wide range of human tissues, and to perform eQTL mapping. Earlier studies used easily accessible samples like cell lines or whole blood, allowing for the profiling of gene expression from hundreds to nearly a thousand individuals[25,26]. However, these datasets were unable to study tissue- or cell type- specific eQTL. To address this question, the Gene-Tissue Expression (GTEx) consortium was the most comprehensive catalogue of eQTL/sQTL from more than 50 tissues with, in several cases, over 300 post-mortem donors[9]. It has been immensely successful at revealing the commonalities and specificities of tissue gene regulation. But one drawback of it is that tissue samples are a mixture of multiple cell types. Given the presence of cell type-specific gene expression, variations in cellular compositions across population can lead to spurious eQTL signals. Moreover, genetic effects that are specific to a rare cell population can be masked by more abundant cell types[27]. This is particularly crucial for the study of genetics in IRD because (1) human immune system consists of a huge diversity of immune cells, and (2) many immune cells that are important in IRD have low cell numbers but profound effects in disease pathogenesis. Indeed, the collection of whole blood rather than individual immune cell types in GTEx greatly hampers its utility in studying IRD GWAS. Fortunately, there is a plethora of studies like BLUEPRINT, DICE and ImmuNexUT that solely focus on the cell types and subtypes in immune cell compartment[28–30]. Although some of these studies have relatively small sample sizes, they provide a high-resolution map of gene regulation in diverse peripheral immune cell types, thus offering a unique opportunity to investigate the genetics of immune-related diseases.

However, current eQTL datasets are not without limitations. Chun *et al.*[31] reported that eQTL from the three immune cell types (B cell, $CD4^+$ T cell, monocyte) colocalized with only 21% of autoimmune disease loci. Although the conclusion in this study is largely consistent with previous analysis, it is among the first to formally question the limitation of eQTL datasets. Another study estimated that *cis*-eQTL from GTEx data only mediate ~11% of trait heritability, again suggesting that eQTL have limited power in explaining GWAS signals[32].

One of the most obvious reasons for the limited overlap between GWAS and eQTL is statistical power. While GWAS typically consists of hundreds of thousands or even millions of individuals, most eQTL studies only contain hundreds of individuals, raising the possibility that weaker eQTL are yet to be found. However, whether significantly increasing the sample size in eQTL studies can lead to better understanding of GWAS data remains an open question. On one hand, larger studies have identified more colocalization events between eQTL and GWAS compared to previous, smaller studies. On the other hand, our previous comparison between eQTLGen data (~30,000 samples) and BLUEPRINT data (~300 samples) suggests that increasing the sample size by 100-fold only increases the number of colocalized GWAS loci by 1.5-fold, even though almost every gene included in eQTLGen has at least one primary eQTL[33]. Collectively, these results show that power cannot fully explain why most GWAS loci are not eQTL, and other properties of GWAS and QTL SNPs should be considered.

Below, I will discuss two aspects that are the major concerns of the following chapters in this thesis, namely, disease-specific regulatory effects overlooked by the use of healthy samples in current eQTL data and the effects of disease variants on chromatin accessibility.

1.4 Chromatin accessibility QTL may complement eQTL data

While eQTL data has greatly expanded the understanding of GWAS variants, it is not guaranteed that all genetic effects on phenotype are mediated through gene expression. In other words, other types of molQTL can capture biological mechanisms that are missed by eQTL. For instance, RNA splicing can change the outcome of protein products without having detectable effects on gene expression, which is often measured as the sum of sequencing reads overlapping exons. Recently, Mostafavi *et al.* proposed that GWAS and eQTL are capturing systematically different sets of genetic variants due to the history of natural selection[34]. The authors observed that closest genes to GWAS loci are more selectively constraint than eQTL genes and have more complex regulatory landscape. They further argued that negative selection on functionally important genes purges large-effect eQTL variants around promoters, whereas GWAS study design is still able to capture variants around these genes. This theoretical analysis suggests that the overlap between GWAS and eQTL SNPs is low not because of technical reasons, but due to selection history[34]. Therefore, the intrinsic limitation of eQTL data may not be salvaged by better experimental design. We hypothesis that other molecular phenotypes can complement eQTL data in that they regulate gene expression through distinct biological processes.

Chromatin accessibility is a measure of "open" chromatin regions that are depleted with histone complex occupation in the genome. First assayed by DNase I hypersensitive sites sequencing (DNase-seq) and more recently by transposase-accessible chromatin with sequencing (ATAC-seq), these accessible regions represent active regulatory elements as they need to be accessible to transcription factor (TF) binding to function[35,36]. Studying caQTL offers several advantages over eQTL. First, it is easier to map causal genetic effects on chromatin accessibility than on gene expression, given that chromatin lies upstream of expression in the gene regulation cascade. As genetic information "flows through" the regulatory cascade, noise and other biological mechanisms unaccounted for

could dilute the signal of causal genetic variants. Indeed, chromatin accessibility tend to be more heritable than gene expression[11]. Second, chromatin accessibility can measure the activity of various kinds of regulatory elements, including promoters and enhancers. Some enhancers might be too far away from a gene body to be tested in eQTL studies, others might have too weak an effect on expression to be detected in eQTL. Moreover, while a gene can be regulated by multiple enhancers, expression level only reflects their joint effects, which is not necessarily additive[37]. On the contrary, caQTL has the ability to capture the regulatory effect of separate enhancers.

1.5 Combining single-cell multiomics with molQTL mapping

The study of disease genetics has constantly benefited from technological innovations that generate new types or modalities of data. Among them, single-cell genomics has revolutionized how biological samples are collected and analyzed, and proved to be a powerful tool when combined with human genetics. Until recently, most eQTL studies used bulk RNA-seq method, which measures gene expression in a group of cells. When tissues are used, the measured gene expression represents an average level across the cell types in the tissue, obscuring cell type-specific eQTL signals. This could be overcome by using sorted cell populations for RNA-seq libraries. However, in many cases, there is no *a priori* knowledge for choosing the correct cell population to study. It is also difficult to collect sufficient number of cells in bulk RNA-seq for rare cell types or subtypes and for precious samples.

Single-cell technologies offer an unprecedented opportunity to address these shortcomings. Single-cell RNA-seq (scRNA-seq) counts the number of messenger RNA (mRNA) molecules in each single cell and distinguishes unique molecules from amplified products by a unique molecular identifier (UMI) that's attached to each molecule. Current technology allows for the simultaneous profiling of thousands of genes in around 10,000 cells in

each experiment, which can be scaled to collect population-level scRNA data from almost a thousand individuals[38], providing enough statistical power for QTL mapping. Single-cell RNA-seq data can not only identify rare cell types and states, it can also infer continuous cell trajectory. Combined with genetics analysis, this has led to the discovery of eQTL with "dynamic" effects in processes like immune activation and induced Pluropotent stem cell (iPSC) differentiation[39,40].

Recent developments in technologies also enable the collection of epigenetic or even multi-modal data at single-cell resolution. While there have been several sc-eQTL studies in the past years, sc-caQTL has been scarce so far. Bulk caQTL studies have already shown that caQTL colocalize with more GWAS loci and mediate more disease heritability, suggesting the potential of sc-caQTL to reveal novel biological understandings. What is more, sc-ATAC data and scRNA-data from matched samples or multi-modal single-cell data offers the opportunity to interrogate the genetic effects on both chromatin accessibility and gene expression in the same set of samples and even the same cells, which can be used to better understand the dynamic of various steps of gene regulation and their relevance to diseases.

1.6 The immunology and genetics of rheumatoid arthritis

RA is an autoimmune disease characterized by chronic, often-symmetrical inflammation in synovial joints (synovitis) and subsequent destruction of bone and cartilage. As it is a systemic disorder, comorbidities often involve other tissues and organs, including lung, heart, and peripheral nervous system. RA is one of the most common autoimmune diseases, with a global prevalence of around 1%[1].

The exact triggers of RA remain elusive so far, but previous studies have identified several factors associated with higher RA risk. The initial breakdown of self-tolerance before symptoms in the joints likely happens in the mucosal surfaces. Smoking and inhalation

of certain particles in textile factory or silica mines lead to higher RA risk, possibly by inducing the citrullination of arginine residues in protein peptide by peptidyl arginine deiminases (PADs) such as *PADI4* and *PADI6*. Periodontal diseases—disorders in the oral mucosal—have also been linked to the triggering of RA, and might be related to interactions with oral microbiomes, which can in turn stimulate PAD activity[41,42].

RA is a highly heritable disease, with a heritability of around 60% in twin studies and 10-25% in the general population[43]. The genetic architecture of RA has been relatively well-studied in the past 15 to 20 years. The major histocompatibility complex (MHC) locus shows the strongest association with RA risk, and studied have identified *HLA-DRB1* as the major risk gene in this region. Multiple risk alleles in *HLA-DRB1* encodes the same amino acid sequences, leading to the *shared epitope* hypothesis. Later studies established the link between shared epitope and the development of anti-citrullinated protein antibodies (ACPA).

GWAS has led to the discovery of a large number of non-MHC genetic variants associated with RA, with more than 150 genome-wide significant loci identified[44,45]. RA GWAS is also among the largest—therefore the most well-powered—studies for autoimmune and immune-related diseases, offering a unique opportunity to study the genetic basis of the disease. Many genes implicated in the genetic studies of RA are well-characterized immune genes or have known role in RA etiology, including *PTPN22*, *IL6R* and *PADI4*.

Many immune and non-immune cell types have been associated with RA development. Multiple types of immune cells infiltrate the synovium in RA, among which $CD4^+$ T cells are crucial to disease pathogenesis. Given that the shared epitope resides in *HLA-DRB1*, a MHC class II gene, it is expected to engage with $CD4^+$ T cells. Indeed, activated $CD4^+$ T cells are accumulated in inflamed joints of RA patients, and secretes proinflammatory cytokines. Many T cell subtypes, such as type 1 Thelper (T_H1) cells and type 17 Thelper (T_H17) cells, are enriched in RA synovium, and play crucial roles in disease progression.

For instance, interleukin-17 (IL-17) produced by T_H17 cells can lead to bone and cartilage destruction. Human genetic studies have also found that genes specifically expressed in effector memory $CD4^+$ T cells are highly enriched around RA risk loci[46].

Given that systemic breakdown of self-tolerance is a characterization of autoimmunity, regulatory T (T_{reg}) cells also play an important role in RA development. A large number of T_{reg} cells are present in the synovial fluids of RA patients. From the perspective of epigenetics, H3K4me3 histone modifications, which represent active promoters and enhancers, are significantly enriched with RA SNP in T_{reg} cells[47]. But the exact role of T_{reg} cells in RA can be complicated by their interaction with various types signaling molecules and immune cell types. Many functional studies on T_{reg} cells in autoimmunity were performed in other autoimmune disease samples or in relevant mouse models. Nevertheless, the discoveries made in these systems can be generalized to many human autoimmune conditions, including RA. Earlier studies suggested that T_{reg} cells might have defective immunosuppressive functions, inspiring clinical trials that aimed at developing therapies that supplement healthy, functional T_{reg} cells to patients. Although most studies have shown that it is safe to adoptivly transfer *in vitro* expanded T_{reg} cells into patients, the treatment effect of this practice remains unclear, with many early stage clinical trials still ongoing[48]. Subsequent studies found that synovial fluid T_{reg} cells are still functional *in vitro*, suggesting that it is effector T cells that are rendered resistant to T_{reg} suppression in inflamed joints. This resistant phenotype has been shown to be mediated by certain proinflammatory cytokines including IL-6, IL-15, and IL-21[49,50].

Recent studies have also pointed to $CD8^+$ T cells and their potential role in RA development. For instance, one study found that enrichment of $CD8^+$ T cells in synovial fluid and synovial tissue of sero-positive RA patients and identified a group of $GZMK^+$ $GZMB^+$ $CD8^+$ T cells that secrete IFN-$\gamma$and tumor necrosis factor (TNF), but have diminished cytotoxic activity[51]. Similar, another study identified clonally expanded $GZMB^+$ $CD8^+$ T

cells in the PBMC of ACPA+ RA patients[52]. However, many aspects of the role of $CD8^+$ T cells in RA development remain elusive and further studies are necessary.

RA is often regarded as a "prototypical" autoimmune disease in that many features of RA are shared with other types of autoimmune conditions. Therefore, understanding RA pathogenesis can potentially be generalized to other diseases.

Given the diverse disease-related cell states and complex interactions in disease tissue, it is plausible that the key to understanding the genetics of IRD like RA is to study genetic effects under disease contexts. Many studies have used *in vitro* stimulation on immune cells as a proxy for disease-relevant contexts. This has led to the understanding of how immune cells act differently to various challenges from a genomic perspective, as well as the discovery of eQTL that respond to stimuli (also known as response eQTL, or reQTL)[53,54]. Nevertheless, how faithful these artificial conditions recapitulate biological contexts that mediate the genetic predisposition of IRD remains contested. On the contrary, directly studying genetic regulation in disease conditions offer at least three advantages. First, many immune cell types show altered gene expression under disease conditions, and many of these alterations are disease-specific. For instance, monocytes from PBMC of RA patients showed substantial differences in their transcriptomes from that of healthy PBMC[55]. Second, and more crucially, because many IRDs often most strongly affect certain organs and tissues, cell types from those locations can show specific genetic effects that are absent from easily-accessible tissues like peripheral blood. Third, cellular functions can be profoundly affected by disease-specific micro-environment that does not exist in healthy individuals. This is particularly important for IRD where chronic inflammation at the disease site can modulate the migration, expansion and activities of immune cells for a long period, or even permanently.

CHAPTER 2

THE IMPACT OF REGULATORY VARIANTS ON HUMAN IMMUNE-RELATED DISEASES

Note: This chapter is adapted from published article: ***Mu, Z., Wei, W., Fair, B. et al. The impact of cell type and context-dependent regulatory variants on human immune traits. Genome Biol 22, 122 (2021).*** `https://doi.org/10.1186/s13059-021-02334-x`

2.1 Abstract

The vast majority of trait-associated variants identified using genome-wide association studies (GWAS) are noncoding, and therefore assumed to impact gene regulation. However, the majority of trait-associated loci are unexplained by regulatory quantitative trait loci (QTL). We perform a comprehensive characterization of the putative mechanisms by which GWAS loci impact human immune traits. By harmonizing four major immune QTL studies, we identify 26,271 expression QTL (eQTL) and 23,121 splicing QTL (sQTL) spanning 18 immune cell types. Our colocalization analyses between QTL and trait-associated loci from 72 GWAS reveals that genetic effects on RNA expression and splicing in immune cells colocalize with 40.4% of GWAS loci for immune-related traits, in many cases increasing the fraction of colocalized loci by two fold compared to previous studies. Notably, we find that the largest contributors of this increase are splicing QTL, which colocalize on average with 14% of all GWAS loci that do not colocalize with eQTL. By contrast, we find that cell type-specific eQTL, and eQTL with small effect sizes contribute very few new colocalizations. To investigate the 60% of GWAS loci that remain unexplained, we collect H3K27ac CUT&Tag data from rheumatoid arthritis and healthy controls, and find large-scale differences between immune cells from the different disease contexts, including at regions overlapping unexplained GWAS loci. Altogether, our work supports RNA splic-

ing as an important mediator of genetic effects on immune traits, and suggests that we must expand our study of regulatory processes in disease contexts to improve functional interpretation of as yet unexplained GWAS loci.

2.2 Introduction

Genome-wide association studies (GWAS) have identified well over ten thousand genomic loci associated with human diseases and complex traits. However, while the number of trait-associated variants continues to grow, the causal genes and mechanisms at most GWAS loci remain difficult to determine. This difficulty is in part owing to the fact that ~90% of GWAS variants lie in noncoding regions.

Multiple studies have now shown that noncoding trait-associated variants are enriched for expression QTL (eQTL) and enriched in regulatory elements such as enhancers and promoters. These findings suggest that noncoding variants likely affect traits by impacting gene regulation, an interpretation which has motivated many studies to map regulatory QTL—in particular eQTL—in a diverse set of tissues and cell types. While eQTL indeed overlap with many variants that have been associated with complex traits and diseases, several studies that assessed colocalization between GWAS and eQTL variants concluded that only a minority of GWAS loci can be explained by the eQTL detected in available samples. For example, a 2017 study reported that ~21% of variants associated with autoimmune diseases colocalize with eQTL in at least one of three immune cell types they analyzed. In addition, a paper from the GTEx consortium suggests that ~20% of GWAS loci show colocalized effect with eQTL in the tissue most relevant to the trait. Moreover, another recent study estimated that only an average of ~11% of trait narrow-sense heritability could be explained by *cis*-genetic effects on gene expression levels as measured in GTEx. Altogether, these observations suggest that very little is known about the genes and mechanisms by which genetic variants impact traits at the vast majority of GWAS

loci.

There are several possible explanations for the modest overlap between GWAS loci and eQTL. For example, there may exist genetic effects on gene regulatory processes other than steady state gene expression levels that mediate genetic effects on trait. Indeed, we previously showed that RNA splicing is an important regulatory mechanisms that link trait-associated variants to complex traits. Another explanation is that genetic effects are often restricted to trait-relevant cell types and cell states that have not been the subjects of colocalization or eQTL studies. Indeed, because the effect of GWAS loci on gene regulation can be cell type-specific, QTL maps in precise trait-relevant cell types must be available for successful colocalization. Additionally, the effects of GWAS loci have also been reported to be disease-specific, and can be found only when QTL mapping in samples collected from disease patients is available. Finally, genetic effects on gene regulation may sometimes be too small to be detected at current sample sizes, even in the causal cell types, cell states, and disease context. While all these possibilities likely contribute to the modest overlap that has been observed, identifying major contributors would significantly help our design of future human genomics studies.

The large number of GWAS loci without a colocalized eQTL is particularly striking for immune-related trait GWAS given that immune cell types have been the subject of the most eQTL studies, and with the largest sample sizes for eQTL mapping. This study aims to leverage the large number of eQTL studies available for immune cell types to understand how regulatory variants affect common disease risk, with a particular focus on the ~80% of autoimmune disease GWAS loci without a colocalized eQTL. Our approach was to perform a uniform eQTL and splicing QTL (sQTL) mapping across cell types and datasets to evaluate the prevalence of cell type and cell state-specific effects, as well as to quantify the colocalization rates between sQTL and GWAS loci.

To this end, we mapped eQTL and sQTL in four datasets, including (i) a dataset with a

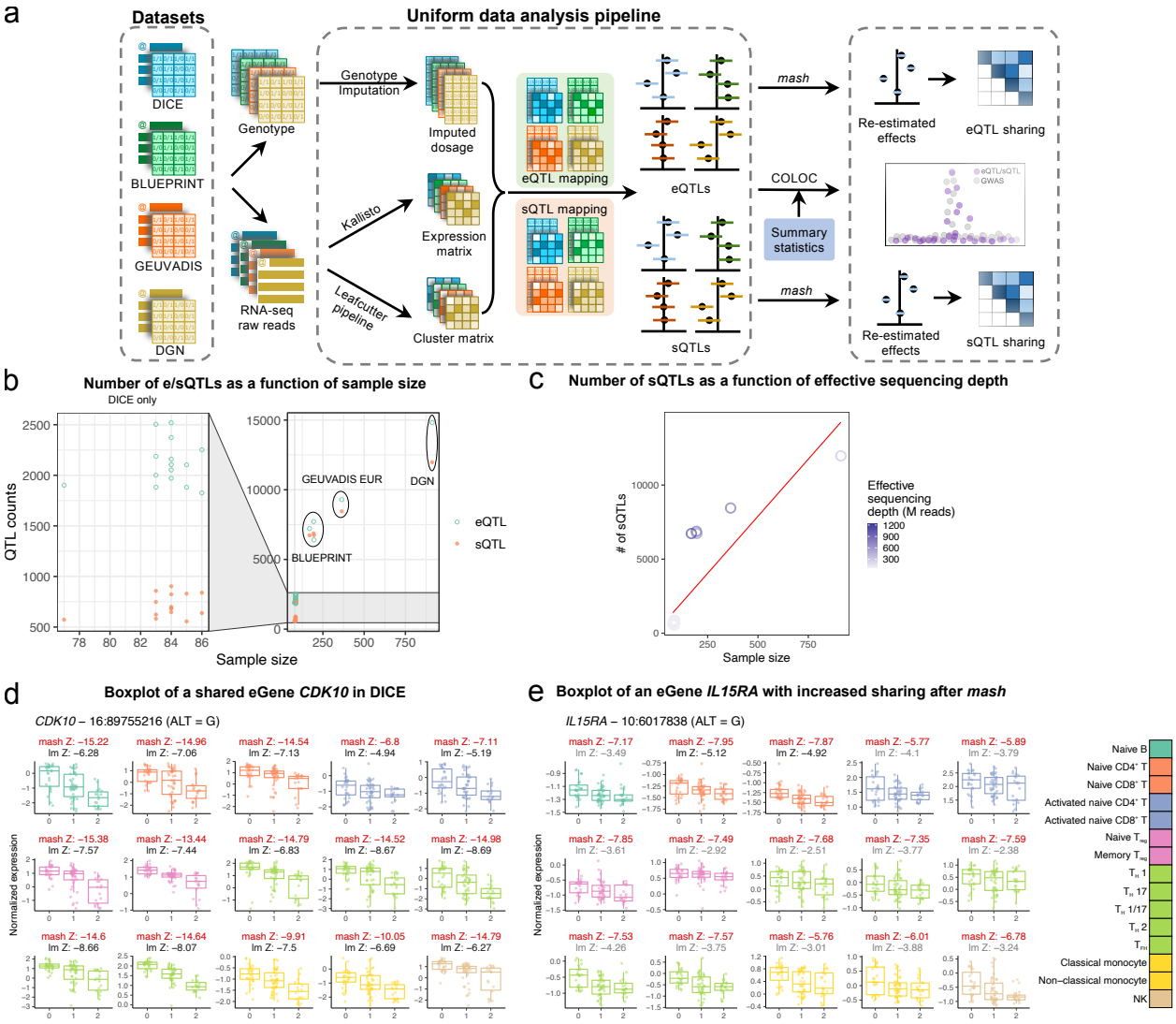

Figure 2.1. Summary of analysis workflow.
a, A uniform computational pipeline to analyze data from four large immune RNA-seq datasets. **b,** Total number of genes and intron clusters with a significant QTL identified in DICE (left) and the three other studies (right) as a function of sample sizes. **c,** Studies with larger effective sequencing depth (BLUEPRINT and GEUVADIS EUR) have more sQTL compared to other studies. Effective sequencing depth = library size $\times$ read-length. Red line represents the fitted simple linear model. **d,** An eQTL for gene *CDK10* is shared by all 15 cell types in DICE despite large differences in baseline expression levels. **e,** An eQTL for gene *IL15RA* is shared across cell types but show cell type-specificity according to linear regression. lm Z: Z-scores of linear model from `FastQTL`, *mash* Z: Z-scores estimated by *mash* (red). The lm Z-scores were colored in grey when the Z-score did not pass statistical significance after `FastQTL` permutation and in black when they were determined to be significant.

large number of different immune cell types but a small sample size (DICE, N = 90); (ii) a dataset with a single tissue-type but with a large sample size (DGN, N = 922); and (iii) two intermediate datasets (BLUEPRINT, N = 197, and GEUVADIS, N = 462). We reasoned that analyzing these datasets in a uniform fashion (**Figure 2.1a**) would allow us to capture both strong QTL with cell type-specificity (using DICE) and weak-effect eQTL that are less likely to be cell type-specific (e.g., using DGN). These data allowed us to evaluate the cell type-specificity of QTL while considering limited statistical power due to small sample size. Moreover, using the uniformly processed QTL, we found that, on average, eQTL and sQTL together colocalize with 40.4% of GWAS loci for autoimmune diseases and blood-related phenotypes, doubling for many GWAS the number of colocalizing loci from previous studies. Interestingly, we found that cell type-specific eQTL account for a very small number of colocalization events, a finding that stands in contrast to several previous studies. Notably, we found that genetic effects on RNA splicing contributed a large number of novel colocalizations, implying that RNA splicing is often impacted by trait-associated variants.

To characterize the remaining 60% of GWAS loci without colocalization, we collected H3K27ac profiles of 5 immune cell types ($CD4^+$ T cells, $CD8^+$ T cells, regulatory T cells, monocytes, and B cells) from rheumatoid arthritis (RA) patients and healthy controls using CUT&Tag. These additional data helped us to better understand the cellular context in which GWAS loci, including those without colocalization, contribute to disease risk. Specifically, our work suggests that to understand the mechanisms underlying many GWAS risk loci, we need to study gene regulation in the context of disease. This stands in contrast to the idea that gene regulation in the disease context merely reflects the consequence of disease, or of response to treatment, which would confound the study of causal genetic mechanisms rather than help it.

Taken together, our work reports a comprehensive analysis of the regulatory effects of

genetic variants on immune cell types, their overlap with GWAS loci and with regulatory regions in a disease context. Our maps of eQTL, sQTL, and gene regulatory regions in diverse immune cell types are available online, which we foresee will aid research on the genetic basis of diseases and on gene regulation in immune cells.

2.3 Results

Harmonized map of eQTL and sQTL in 18 immune cell types

We built a uniform data processing pipeline to harmonize four population-scale RNA-seq datasets (**Figure 2.1a**). The DICE dataset consists of population RNA-seq data for 13 unstimulated immune cell types including various naïve and effector/memory T cell sub-types, classical and non-classical monocytes, B cells, and NK cells. The DICE dataset also includes RNA-seq data from $CD4^+$ and $CD8^+$ T cells that have been activated *in vitro* by engaging T cell receptor (TCR) complex using CD3/CD28 antibodies. Although the sample size in the DICE dataset is the smallest (n = 91) among the four datasets, the large number of sorted cell types makes the DICE dataset ideal to identify cell type-specific genetic effects. The BLUEPRINT dataset consists of RNA-seq data from three cell types (classical monocytes, naïve $CD4^+$ T cells and neutrophils) in 197 individuals. The DGN consortium collected whole blood samples from 922 individuals and, finally, GEUVADIS collected RNA-seq data from 462 lymphoblastoid cell lines (LCL).

Our pipeline, which is described in detail in Methods, includes quantifying RNA expression and splicing levels, imputing genotype data to the same common reference panel, and calling QTL in all datasets using the same strategy. We also designed an approach to harmonize quantification for splicing junction usage across cell types and datasets by first merging textttLeafCutter intron clusters across all samples and then re-calculating intron usage for each sample (Methods). Thus, we produced a harmonized set of introns

that can be readily interrogated. To map eQTL and sQTL, we used `FastQTL`. As covariates for the linear regression, we used three genotype PCs and a number of phenotypic PCs chosen to maximize the number of significant QTL (Storeys q-value < 0.05). In total, we discovered 26,271 genes and 23,121 intron clusters that have a significant QTL in at least one the four datasets at 5% false discovery rate (FDR). As expected, both the numbers of eQTL and sQTL were correlated with sample size (**Figure 2.1b**). In addition to sample size, we found that the number of sQTL identified was also correlated with effective sequencing depths (**Figure 2.1c** and **Figure S2.1**). Notably, while the number of sQTL is roughly linearly related to sample size, datasets with higher effective sequencing depths consistently yielded more sQTL than predicted by a simple linear model. This is most obvious for BLUEPRINT, which used 100 bp single-end or paired-end sequencing when compared to DICE or DGN (both 50 bp single-end). We show a shared eQTL for the *CDK10* gene (**Figure 2.1d**) and an eQTL for the *IL15RA* gene (**Figure 2.1e**) as examples. All gene expression and splicing quantification, as well as all identified eQTL and sQTL are available on Zenodo[56].

Global patterns of eQTL and sQTL sharing across immune cell types

To characterize the cell type-specificity of genetic effects on gene regulation in immune cells, we sought to discern genetic variants that impact gene regulation broadly across many or all immune cell types from those that impact a few or only one cell type. Previous studies have also quantified the sharing and specificity of regulatory QTL[28,29]. However, because the sample sizes of most datasets are small, we speculated that estimates of QTL effect sizes are noisy, which would generally cause studies to underestimate the levels of QTL sharing. We reasoned that our harmonized dataset would allow us to better infer sharing patterns. In particular, we improved our estimates of eQTL and sQTL effect sizes at each locus by statistical shrinkage using *mash*[57]. The *mash* method improves esti-

mates of QTL effect sizes from those that are obtained from applying linear regression in each cell type separately, because *mash* leverages the correlation structure of QTL effect sizes across all cell types to re-estimate QTL effect sizes at each locus. We applied *mash* to calculate posterior mean effect sizes (henceforth referred to as *mash* effect sizes) and corresponding standard errors for the 36,950 unique SNP-gene associations and 116,881 unique SNP-intron associations (q-value below 5%) in the 15 DICE cell types separately. This procedure greatly enhanced estimates of QTL effect sizes in the 15 immune cell types (for two examples see Z-scores in **Figure 2.1d-e**, also see **Figure S2.2** and Methods).

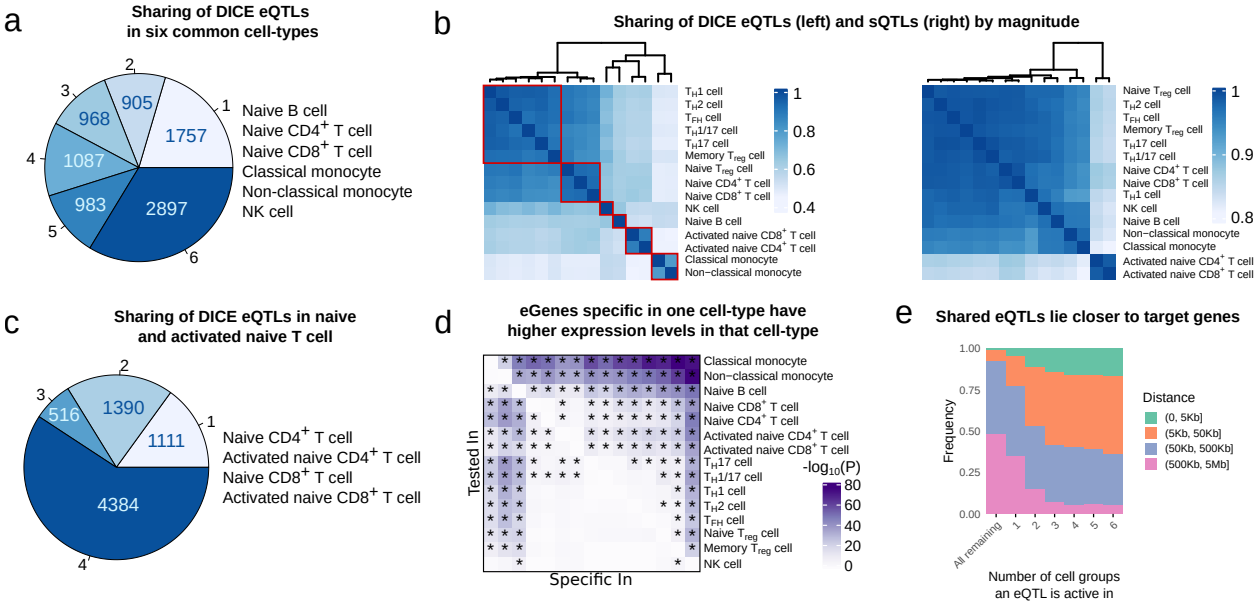

Figure 2.2. Sharing of eQTL and sQTL using *mash*.
a, Estimated number of cell types in which eQTL are inferred to be significant among six common cell types from *mash*. These estimates of sharing are much higher than that from the original study[29]. **b**, Sharing of eQTL (left) and sQTL (right) by magnitude (Fold difference in QTL effect size less than 2). Red square: cells were grouped into six clusters based on eQTL sharing, which resulted in the following groups: (i) Naïve T cells, (ii) Memory and Effector T cells, (iii) Monocytes, (iv) activated T cells, (v) B cells, and (vi) NK cells. **c**, Estimated number of cell types in which eQTL are inferred to be active among naïve and activated naïve T-cells from DICE. **d**, Heatmap showing -$\log_{10}$p values of a differential gene expression analysis that compared the expression level of cell type-specific eGene in the discovery cell type to the expression levels of the eGene in the other 14 cell types. **e**, Distance between eQTL and their target genes stratified by the number of cell groups in which the eQTL is active. To obtain the six cell groups, we grouped the 15 cell types based on similarity in their *mash* effect sizes as described in **b**.

We first asked about the proportion of QTL that are shared across immune cell types based on the estimated *mash* effect sizes. We found that a large fraction (33.7%, n = 2897 of 8597) of genes with an eQTL (eGenes) are shared according to *mash* (Local False Sign Rate [LFSR] < 0.05) across all six distinct major cell types in the DICE dataset (B cell, naïve $CD4^+$ and $CD8^+$ T cell, NK cell, classical monocytes, non-classical monocytes) (**Figure 2.2a**). Our estimates of sharing are therefore much higher than the 5.2% (463 out of 8863) estimated in the original DICE study[29]. In fact, the original DICE study estimated that nearly half of all eGenes are specific to a single immune cell type, while our new estimate suggests that only 20.4% are likely cell type-specific (**Figure 2.2a**).

Using *mash* effect sizes, it is also possible to quantify the amount of QTL sharing in terms of magnitude of effects. We found that over 40% of eQTL have similar *mash* effect sizes (within 2-fold, 34% within 1.5-fold) across pairs of cell types, and this fraction increases to over 90% when considering closely related cell types, such as classical and non-classical monocytes or TH 1 and TH 17 cells (**Figure 2.2b**, left). In addition, we found that the vast majority (> 80% within 2-fold, 67% within 1.5-fold) of sQTL have similar *mash* effect sizes across all immune cell types, with activated $CD4^+$ and $CD8^+$ T cells forming an outlier group in a hierarchical clustering based on estimates of sharing (**Figure 2.2b**, right). These results are consistent with previous work and suggest that the impact of genetic variation on RNA splicing is generally shared across two cell types when the involved mRNA transcripts are expressed in both cell types, which is largely the case for any pair of immune cell types.

In general, the proportion of shared eQTL across cell types captured the lineage relationships among the 15 immune cell types. Specifically, classical and non-classical monocytes clustered together, while B cells and NK cells each formed distinct clusters. Furthermore, despite a high level of QTL sharing (> 80%) among naïive T cells, we found that naïive $CD4^+$, $CD8^+$ and regulatory T cells formed one cluster, while memory and ef-

fector T cells formed another larger cluster. We also observed a higher level of QTL sharing between activated $CD4^+$ and $CD8^+$ T cells compared to that between stimulated and naiïve T cells. This observation suggests that activated $CD4^+$ and $CD8^+$ T cells share similar gene expression programs upon activation, and that differences in genetic effects on gene regulation exist between activated and non-activated cells. Nevertheless, we found that 66.2% of eGenes (n = 4,900) were shared according to *mash* (LFSR < 0.05) between naiïve and activated T cells, suggesting that the overall impact of genetic effects on gene regulation is in most cases the same across activated and naiïve T cells (**Figure 2.2c**). We also calculated share-by-sign and share-by-magnitude excluding the HLA locus (Chr6: 25-35 Mb), and observed no significant difference in eQTL sharing levels. This is not surprising given that the HLA locus contains but a few hundreds genes, which account for 1.26% of all genes included in our *mash* analysis (**Figure S2.3**).

While a large proportion of eQTL and sQTL appeared to be shared across multiple immune cell types, we found that a substantial number of eQTL (2810 eQTL, 27.8%, which include stimulation-specific eQTL) appeared cell type-specific. We asked whether QTL that appeared cell type-specific showed specific features compared to QTL that were shared across immune cell types. We first asked whether genes with eQTL that were specific to a cell type were also more highly expressed in that cell type compared to the other cell types. To test this, we asked whether genes with an eQTL in a cell type A but not in another cell type B, were significantly more highly expressed in cell type A than cell type B. Indeed, we found that this was the case for most cell type-specific eQTL (66.7%, Bonferroni adjusted P value < 0.05, one-sided, paired Wilcoxon rank-sum test), suggesting that variation in gene expression level likely impacts whether a genetic variant has a regulatory effect and/or our ability to detect this effect. This observation was most obvious for classical monocytes, non-classical monocytes and naïve B cells, and is driven by differences in their gene expression levels compared to T cells (**Figure 2.2d**). In ad-

dition to differences in gene expression levels, we found that eQTL that were cell type-specific were located further away from the gene transcription start site in comparison to eQTL that were shared across immune cell types (**Figure 2.2e**). Moreover, cell type-specific eQTL were more highly enriched in enhancers compared to eQTL that were shared (**Figure S2.4**). These observations are consistent with the notion that cell type-specific eQTL tend to impact enhancer activity, while shared eQTL more often impact promoters.

Taken together, our analyses revealed that QTL effects are shared for a large number genes. Nevertheless, we were able to detect a non-negligible number of cell type or cell group-specific QTL. Importantly, these findings and classification show replication across datasets (Supplementary Notes). Thus, we expect our QTL data to be highly replicable in existing or future immune QTL datasets.

Colocalization of immune regulatory QTL with common disease GWAS

Our harmonized eQTL and sQTL data gave us the unprecedented ability to identify genetic variants that impact traits through regulatory effects on immune cell types. We performed colocalization analyses that aimed to determine whether the genetic variants at GWAS loci that are causal for a trait are likely to be the same variants as the causal regulatory QTL. We compiled a set of 72 well-powered GWAS, including 14 for autoimmune diseases (11 unique disease types), 36 blood traits, and 22 other traits, and used `COLOC` to evaluate colocalization (PP4 $\geq$ 0.75)with DICE, BLUEPRINT, and DGN QTL separately (average N = 206,090)[21]. We computed the colocalization rate for each GWAS as the percentage of GWAS loci that show evidence of colocalization out of the total number of associated loci in the GWAS (Methods). We report the main colocalization results of our analyses using BLUEPRINT QTL (3 immune cell types) below, and use the DICE (15 cell types) regulatory QTL to interpret the cell type-specificity of colocalized genes. We reasoned that choosing BLUEPRINT over DICE as the main dataset for this analysis will increase our power for

QTL mapping owing to its larger sample size and will also allow us to identify more sQTL owing to higher RNA-seq coverage and longer read-lengths (**Figure 2.1c**).

When we ascertained colocalization between GWAS loci for the 72 traits and QTL from BLUEPRINT, we observed that colocalization rates between immune regulatory QTL and GWAS hits were higher for autoimmune and blood-related traits compared to other non-immune traits (mean 40.4% versus 27.7%) (**Figure 2.3a**). This observation supports the expectation that a large fraction of colocalized regulatory QTL indeed affect immune traits by impacting gene regulation in immune cell types.

We next focused on autoimmune diseases and blood-related traits. Our regulatory QTL colocalized with a mean of 38.1% (range: 24-47.4%, n = 14) and 41.4% (range: 33.8-50%, n = 36) of autoimmune disease and blood traits GWAS loci (PP4 >0.75), respectively (**Figure 2.3a**). The mean rates of colocalization ranged from 27.4% to 50.2% depending on the choice of posterior probability cutoff for determining colocalization status (PP4, ranging from 0.5 to 0.9, **Figure S2.5**). We chose to use an intermediate cutoff of 0.75 to be consistent with previous studies[31]. Expression QTL colocalized with 6.7-39.3% of GWAS loci with an average of 26.6%, similar to estimates from a previous study[31]. Notably, we found that splicing QTL colocalized with an additional 7.6-21% GWAS loci (average: 13.8%) that did not colocalize with an eQTL, and explain much of the increase in colocalization rates from this study compared to that of previous studies. Interestingly, we observed that the rates of colocalization between GWAS loci and both an eQTL and an sQTL can vary substantially across traits, ranging from 4.5% for systemic lupus erythematosus (SLE) to 28.6% for basophil percentages of granulocyte (BASO%GRAN) (average: 17.2%). Most notably, nearly all colocalized loci associated with SLE (10 out of 16) colocalized only with sQTL (**Figure 2.3a**). Interestingly, the rates of colocalization were not correlated with GWAS sample size nor the number of significant loci, and thus the variation in colocalization rates cannot be attributed to differences in GWAS power

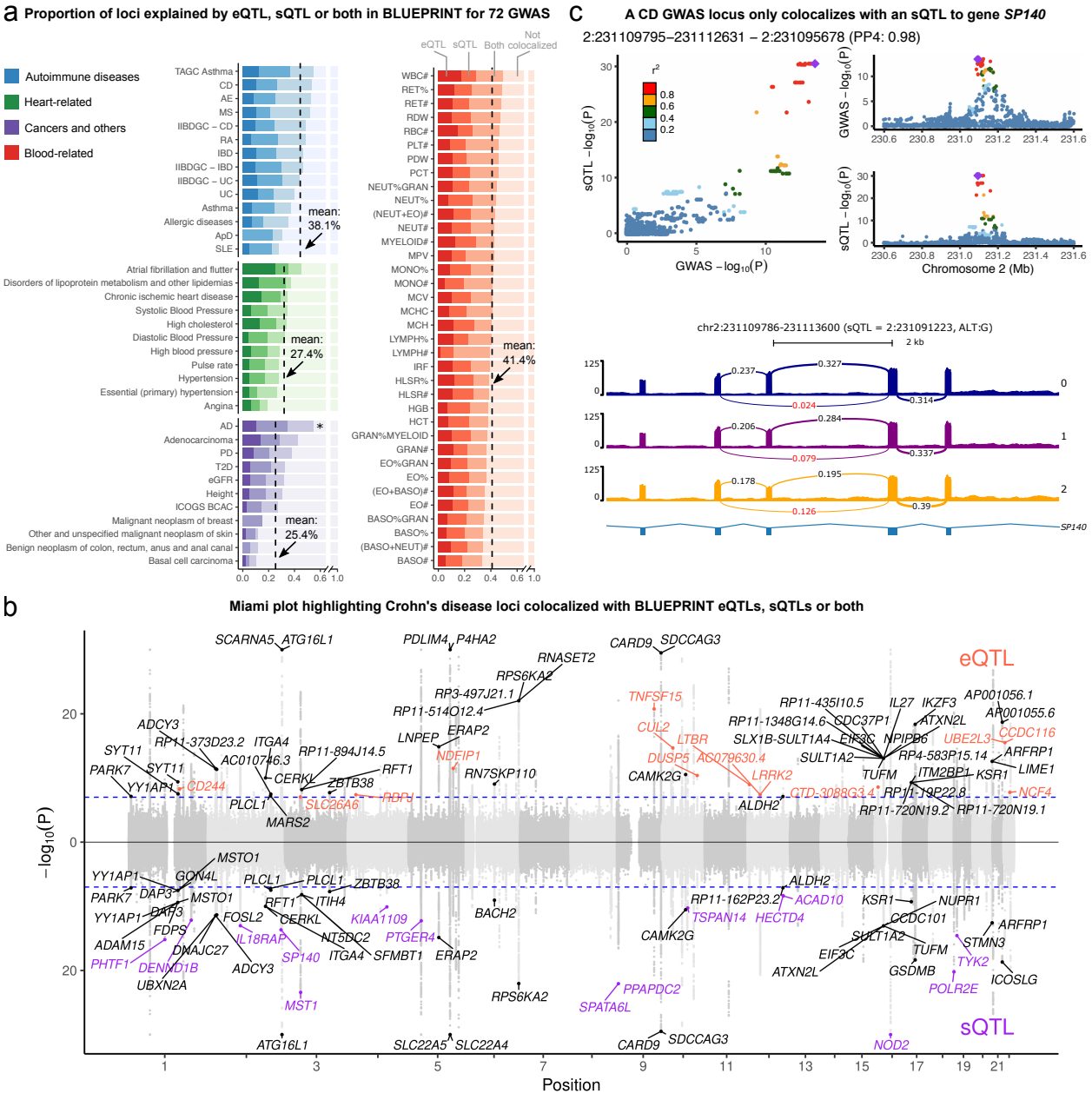

Figure 2.3. Colocalization analysis explained up to 47% of GWAS variants and revealed potential causal SNPs to non-immune traits.
a, Proportions of GWAS loci colocalized with eQTL, sQTL, or both. Dashed line: mean colocalization rate. *: Alzheimers disease (AD) GWAS was not included in the mean calculation owing to the well-documented involvement of microglia in AD. **b,** Colocalization of Crohns Disease (CD) GWAS with eQTL (orange), sQTL (purple), or both (black). GWAS SNPs with -log 10(P) larger than 30 were set to 30 to facilitate visualization. **c,** LocusCompare plot (top) and Sashimi plot (bottom) showing colocalization between a sQTL of an intron in gene *SP140* in T cell and a GWAS locus for CD. Arrows in the Sashimi plot point to the intron affected by the sQTL, labeled with PSI quantification from LeafCutter[58].

(**Figure S2.6**). This result raises the possibility of distinct regulatory architectures for different diseases. We obtained similar rates of colocalization with DICE and DGN, for which 30.7% and 38% immune GWAS loci colocalized with DICE and DGN regulatory QTL, respectively (**Figure S2.7**).

To help the interpretation of these results, we show the colocalizations between immune regulatory QTL and GWAS loci for Crohns disease (CD) as an example (**Figure 2.3b**) (12,194 cases and 28,072 controls)[59]. We included 108 GWAS loci in our colocalization analysis that pass a p-value threshold of 10^{-7} (Methods). Ten and fifteen loci colocalized with only eQTL or sQTL, respectively, while an additional 25 loci colocalized with both eQTL and sQTL. In total, 46% of loci colocalized with an eQTL, an sQTL, or both. Of note, several identified colocalized genes have been extensively studied in terms of CD etiology, including *NOD2* and *ITGA4*, of which the latter is the target for the CD monoclonal antibody drug natalizumab[60,61].

The high rates of colocalization (average: 13.8%) between GWAS loci and sQTL highlight the importance of considering the impact of risk variants on RNA splicing. For example, we identified an sQTL associated with the skipping of the seventh exon in gene *SP140* in T cells that colocalized with a risk locus in both CD GWAS we analyzed (**Figure 2.3c**)[59,62]. *SP140* encodes nuclear body protein SP140[63], which preferentially binds to gene promoters with H3K27me3 modification and regulates multiple immune-related genes[64,65]. Notably, the exclusion of the same exon in *SP140* transcript isoforms has also been associated with risk alleles for other diseases including multiple sclerosis[66].

As expected, immune regulatory QTL colocalized at a lower rates in GWAS of traits that are not autoimmune or blood-related (27.7%). Among the 22 non-immune traits we analyzed, Alzheimers disease (AD) is an outlier, for which 55% of GWAS loci colocalized with a BLUEPRINT QTL. The high rate of colocalization can be explained by the known role of microglia in AD etiology[67]. Nevertheless, it is likely that for most other non-

immune traits GWAS loci, colocalization with immune regulatory QTL reflect a causal effect of the risk variant on disease through non-immune cell types that is also manifested in an immune cell type (Supplementary Notes).

GWAS-eQTL colocalizations across immune cells are highly shared when accounting for statistical power

Several studies have proposed that a large fraction of autoimmune disease risk loci affect gene expression levels in a cell type-specific manner[68,69]. We sought to use our dataset to evaluate this hypothesis by analyzing the cell type-specificity of the eQTL that colocalize with autoimmune GWAS loci. To do this, we focused on the 197 genes with a DICE eQTL (eGenes) that colocalized with at least one of the 14 IRD GWAS in our study. We then evaluated the cell type-specificity using the *mash* QTL effect sizes estimated for the 15 immune cell types from the DICE consortium.

The general pattern of sharing that we observed for colocalized risk loci is that the corresponding eGenes are mostly shared across multiple cell types. The sharing was also apparent across the 6 major groups of immune cells that represent naïve T cells, memory and effector T cells, monocytes, activated T cells, B cells, and natural killer (NK) cells (**Figure 2.4a**). Sixty five of 197 (33.0%) tested genes colocalized in all 6 major immune cell groups. The immune cell groups in which the most colocalized genes were found are memory and naïve T cells, in which 160 and 151 of 197 eGenes colocalized with GWAS loci, respectively. However, only 8, 8, and 4 eGenes showed an effect that appear to be specific to B cells, monocytes, and NK cells, respectively, while 12 eGenes showed an effect only in T cells. These observations suggest that for the vast majority of autoimmune risk loci, the effect of risk variants on gene expression level is not restricted to a single immune cell type or cell group.

We next set to understand the discrepancy between our finding that most GWAS loci

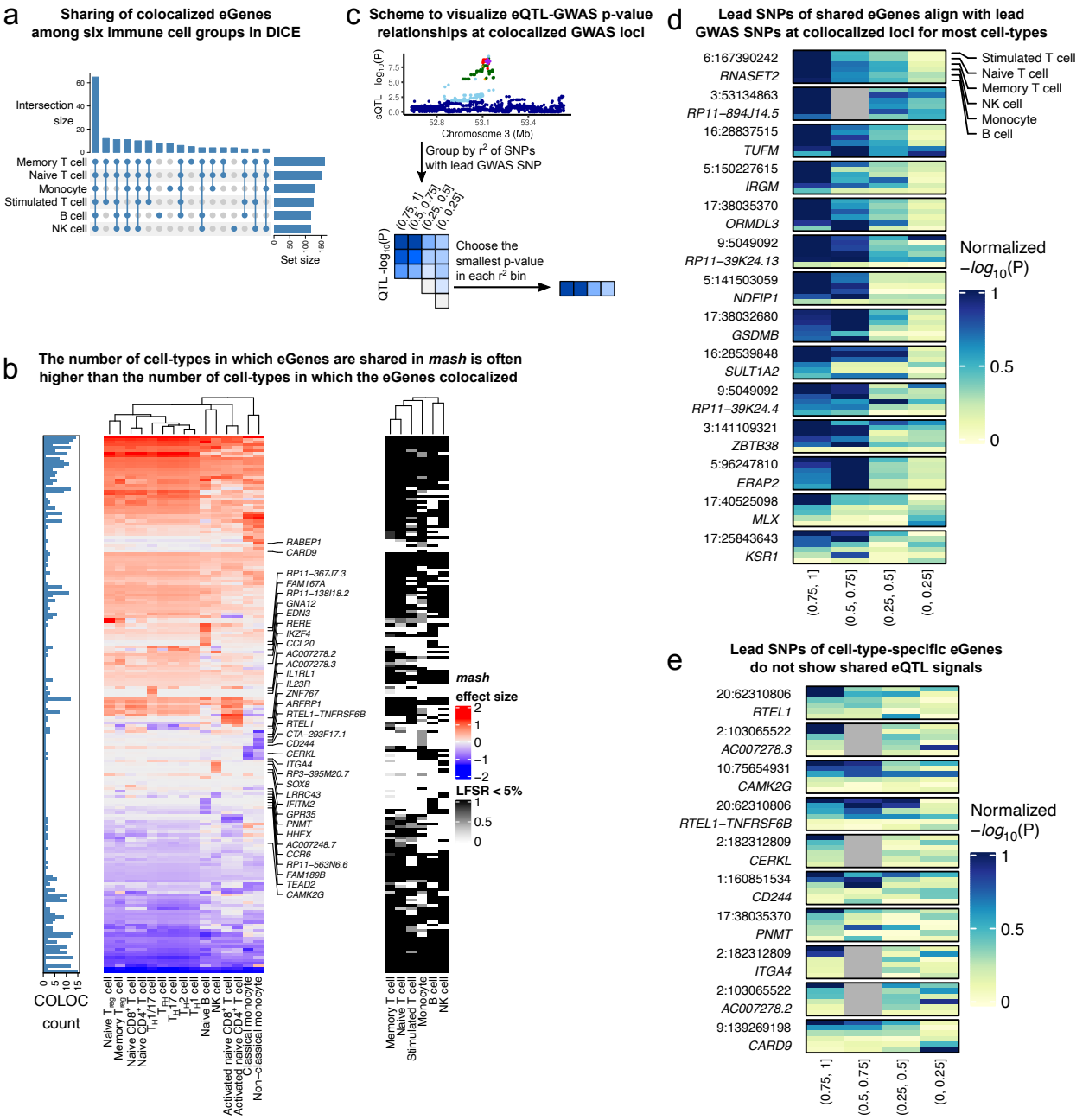

Figure 2.4. *mash* analysis indicates high sharing of QTL among immune cell types.
a, Upset plot showing the sharing of eGenes colocalized with IRD GWAS loci. The majority of colocalized eGenes are shared across the 6 cell groups. **b,** Heatmaps showing *mash* effect sizes of colocalized eGenes (left) and LFSR (<0.05, right). Barplot on the left shows the number of cell types in which eGenes colocalized with a GWAS locus. **c,** Schematic representation of our approach to visualize the QTL association P value of colocalized eGenes across SNPs with different amount of LD with the lead GWAS SNP. **d,** eQTL P values in different LD bins (as described in **c**) at GWAS loci with colocalized eQTL across all 6 cell groups. **e,** Similar to **d**, eQTL P values in different LD bins (as described in **c**) at GWAS loci with colocalized eQTL that were inferred to be cell type-specific.

impact multiple cell types and that of previous work, which suggests more cell type-specificity[46,68]. We first analyzed colocalization status of autoimmune GWAS loci in each cell type separately, which corresponds to the general approach used by previous studies[9,29,31]. We found that, using this approach, the number of cell types with positive colocalization status is generally smallersometimes much smallerthan the number of cell types in which the eQTL effects are shared according to our analysis (**Figure 2.4b**). We speculate that this discrepancy results from the variation in the posterior probabilities of colocalization computed by COLOC, owing to inherent noise in estimating the effect sizes and statistical significance of eQTL (Supplementary Note).

We asked whether this observation was reflective of a general trend across GWAS loci. We reasoned that, under the simplifying assumption that there is only one causal eQTL at each GWAS locus, colocalized loci should show a general pattern where SNPs in high LD with the lead GWAS SNP will show strong associations with expression levels of the colocalized gene, but the eQTL associations will weaken for SNPs in lower LD. Thus, eQTL that colocalize with a GWAS locus in all cell types should show decreasing eQTL association strength for SNPs in decreasing amount levels of LD for most or all cell types. By contrast, eQTL that only colocalize with a GWAS locus in a single cell type, should show these patterns only in a single or a small number of cell types.

To visualize these patterns across many GWAS loci and cell types, we first found the lead GWAS SNPs at every colocalized loci and divided all SNPs within 1Mb into four bins according to their linkage disequilibrium (LD) with the lead SNP (namely, r^2 within ranges of (0, 0.25), (0.25, 0.5), (0.5, 0.75), and (0.75, 1)). Next, for each r^2-bin, we identified the SNP with the smallest eQTL p value for the colocalized eGene in each of the 6 DICE cell groups (**Figure 2.4c**). We then plotted the p values for all the colocalized locus-gene pairs where the *mash* SNPs and the lead GWAS SNPs are in close LD (r^2 >0.8, **Figure 2.4d**, **Figure S2.8a**). We observed that the most significant eQTL are often in high LD with the

lead GWAS SNP for multiple cell groups (rows) when the eQTL were determined to have shared effects. By contrast, for the eQTL we inferred to have a cell type-specific effect, the patterns are strikingly different as the most significant eQTL are more likely to be in lower r^2-bins in most cell types (**Figure 2.4e**, **Figure S2.8b**). These findings support our high estimates of shared regulatory effects of GWAS variants across multiple cell types. These observations also suggests that COLOC is susceptible to noise in QTL mapping, especially when the sample size in QTL mapping is small. More importantly, our data indicates that previous work over-estimated the fraction of GWAS loci with cell type-specific effects on gene expression levels. Indeed, we found cell type-specific colocalization in a single or two major immune cell groups for only 35 of 197 loci (17.8%), while 103 (52.3%) are eQTL in five or more cell groups.

Limited regulatory effects specific to stimulated cells at GWAS loci

Our analysis so far indicates that about 40% of autoimmune GWAS loci have a detectable effect on gene regulation in at least one of the 18 immune cell types analyzed. We next wondered about the mechanism by which the remaining 60% of GWAS loci function. There are several possible explanations for why such a large fraction of GWAS loci do not colocalize with a regulatory QTL identified in our study. One simple explanation is that many of these GWAS loci do not impact disease risk by affecting the expression or splicing of mRNA. Instead, they may affect protein coding sequence or other as yet poorly studied molecular mechanisms, such as alternative polyadenylation[69].

To identify putative mechanisms by which trait-associated variants at uncolocalized GWAS loci function, we asked whether genes in GWAS loci without colocalization were different in terms of expression levels, enhancer density, and sequence constraint compared to those in GWAS loci with colocalization (Methods). Our analysis revealed that genes in loci without colocalization are expressed at a significantly lower levels than com-

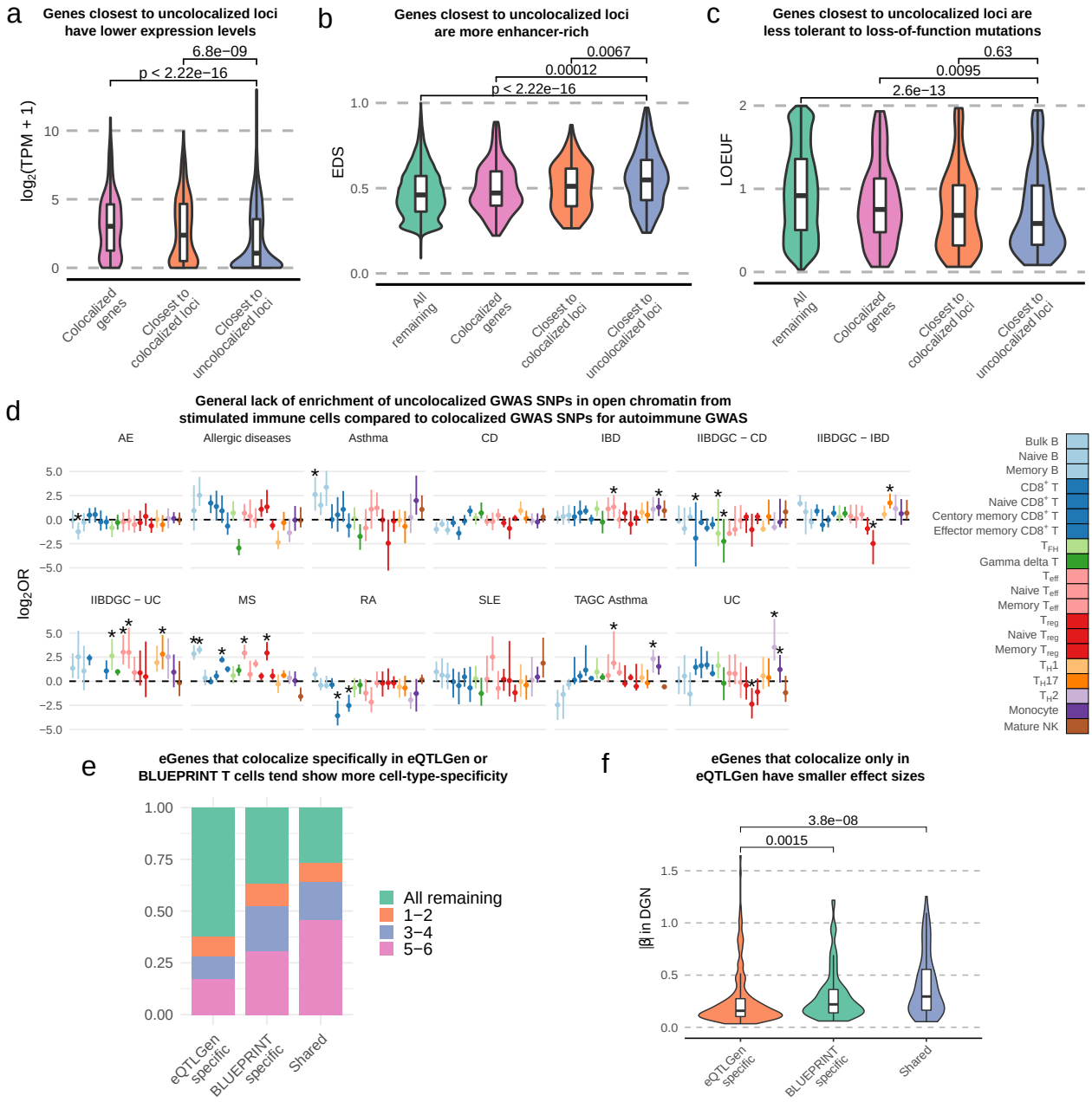

Figure 2.5. Characterizations of uncolocalized GWAS loci.
a, Genes closest to uncolocalized loci are expressed at lower levels compared to colocalized eGenes. **b,** Genes closest to uncolocalized loci have higher EDS. **c,** Genes closest to uncolocalized loci have lower LOEUF[70]. **d,** Forest plot showing the $\log_2$ OR of enrichment of uncolocalized GWAS SNPs in open chromatin of stimulated immune cells compared to colocalized GWAS SNPs (Fishers exact test). Error bars show 95% confidence intervals from bootstrap (Methods); *: FDR <0.05. **e,** eGenes that colocalized only in eQTLGen data tend to be restricted in fewer cell types (in DICE data) compared to eGenes that colocalized only in BLUEPRINT data or eGenes that colocalized in both BLUEPRINT and eQTLGen. **f,** eGenes that colocalize only in eQTLGen data have smaller effect sizes compared to eGenes that colocalize only in BLUEPRINT T cells or eGenes that are shared between eQTLGen and BLUEPRINT.

pared to genes at loci with colocalization (**Figure 2.5a**). In addition, we found a higher enhancer density as measured by EDS[37] (**Figure 2.5b**), and a lower tolerance to loss-of-function mutations as measured by LOEUF[70] (**Figure 2.5c**) for genes in uncolocalized GWAS loci.

Several studies have proposed that many autoimmune disease GWAS loci impact gene regulation in stimulated but not resting immune cells[71,72]. Thus, it is possible that a large fraction of uncolocalized GWAS hits impact gene regulation in stimulated but not unstimulated cells. However, we found in an earlier analysis of DICE RNA-seq data that, although some exceptions exist (**Figure S2.9**), regulatory effects in stimulated $CD4^+$ and $CD8^+$ T cells were largely the same in unstimulated T cells. As a less direct but complementary analysis, we therefore asked whether uncolocalized GWAS loci were more likely to overlap with open chromatin regions in stimulated immune cells compared to colocalized ones using ATAC-seq data from 20 naïve and stimulated immune cells[71]. Again, we found very little support for the hypothesis that a large fraction of uncolocalized GWAS loci impact gene regulation in immune cells that were stimulated. Specifically, we observed very subtle differences in the enrichment of uncolocalized GWAS SNPs in open chromatin regions of stimulated immune cell types compared to that of colocalized GWAS SNPs (**Figure 2.5d**, Methods). When accounting for multiple testing, only 17 out of 254 tests are significant at a FDR of 5%, and the enrichment for these were modest. Thus, these analyses suggest that there are fundamental differences in the mechanisms and genes that underlie colocalized and uncolocalized autoimmune GWAS loci, but the difference cannot be simply explained by regulatory effects that are restricted to stimulated immune cells.

No evidence for GWAS colocalization with small effect eQTL at most unexplained loci

Another explanation for the large number of uncolocalized GWAS loci is that the regulatory effects of many GWAS loci are outside current range of detection owing to small sample sizes. As a simple way to test this, we performed an eQTL analysis for only the lead GWAS SNPs at uncolocalized CD GWAS loci in BLUEPRINT T cells. The smaller number of tests compared to a genome-wide analysis improved our ability to detect eQTL with smaller effect sizes (mean absolute effect size 0.34 versus 0.64 genome-wide, **Figure S2.10**). However, we found that only a small fraction (7.97% on average) of uncolocalized autoimmune GWAS loci showed evidence of a regulatory effect using this approach. This would still leave about half of all autoimmune GWAS loci uncolocalized.

As an additional test, we asked how many uncolocalized GWAS loci could be colocalized using eQTL summary statistics from eQTLGen, which were obtained from a meta-analysis of 31,684 whole blood samples[17], including the 922 DGN samples analyzed in our study. As quality control, we first compared the eQTLGen colocalizations with that of DGN (15,269 common genes) and that of BLUEPRINT (15,373 common genes). Of the 242 autoimmune GWAS loci that colocalized with DGN eQTL, 196 were found to replicate using the eQTLGen dataset (168 of 232 (72.4%) for BLUEPRINT). The higher replication rates for DGN was to be expected given the sample overlaps and that DGN and eQTLGen sampled the same tissue type, whole blood, while BLUEPRINT assayed sorted immune cell types.

Using eQTLGen eQTL, we identified an additional 130 GWAS loci that colocalize in eQTLGen but not in BLUEPRINT, on average accounting for 16.8% (range: 6.6% - 35.8%) of uncolocalized loci from our BLUEPRINT analyses (**Figure S2.11**). These findings suggest that although the gain in colocalization by increasing sample size could be large for some GWAS (e.g., 35.8% for multiple sclerosis), the average increase in colocalization rate

is small. As expected, colocalized eGenes specific to eQTLGen tend to not be eGenes in DICE immune cell types, or were eGenes with cell type-specificity (**Figure 2.5e**). Additionally, the eQTL of colocalized eGenes specific to eQTLGen have smaller effect sizes on average than that of colocalized eGenes specific to DGN, which in turn have smaller effects on average than colocalized eGenes that were identified to be shared in the DICE dataset (**Figure 2.5f**). Thus, despite the substantial improvement in detection power afforded by the large eQTLGen sample size, the average GWAS colocalization rates for eQTL only increased slightly, from 22.9% using BLUEPRINT compared with 29.4% using eQTLGEN. Indeed, even when colocalized loci ascertained in DGN and eQTLGen are combined together, only an average of 35.8% GWAS loci colocalized with an eQTL. While this is a relatively big increase, suggesting that the lack of colocalization at many GWAS loci is due lack of power in our eQTL analysis, these results suggest that increasing the sample size of our eQTL analysis is unlikely to account for the majority of the uncolocalized GWAS loci, at least for eQTL studies on cell types that are well-represented in whole blood.

Condition-specific profiles of H3K27ac in RA patients highlights context-dependent effects in RA pathogenesis

Finally, we hypothesized that the effects of some uncolocalized GWAS loci may be more readily interpretable in the context of the corresponding disease. While stimulation of immune cells in vitro may capture some important regulatory features reflecting disease state, we reasoned that studying immune cells sampled directly from autoimmune disease patients may better help understand the effects of uncolocalized GWAS loci. To this end, we focused specifically on rheumatoid arthritis (RA), an autoimmune disease that primarily affects synovium joints and is often associated with immune cell infiltration that leads to the build up of synovial fluid (SF) that can be collected from a joint aspiration[73].

To obtain regulatory profiles of cells in the context of RA, we first collected peripheral

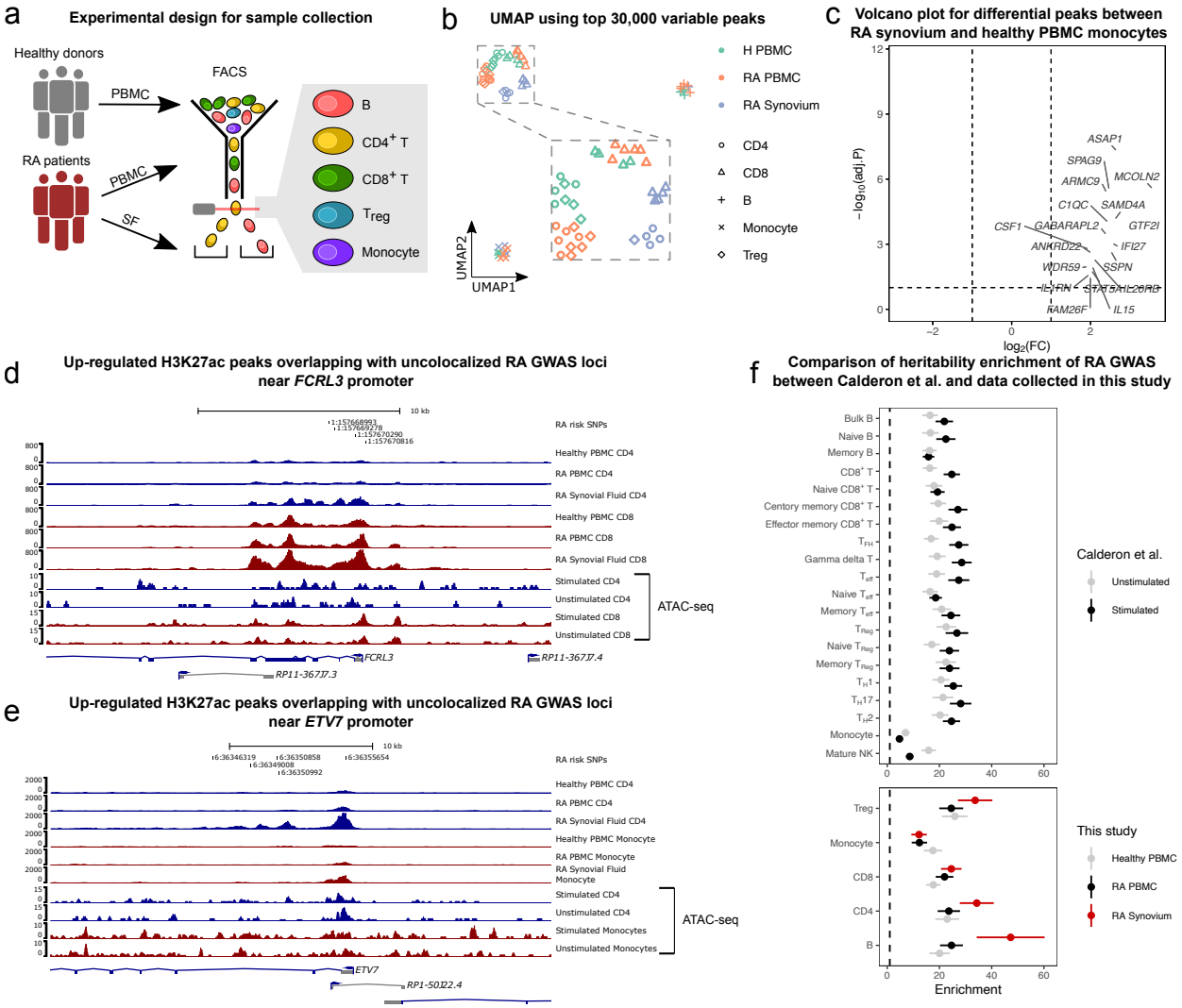

Figure 2.6. H3K27ac profiling in RA samples reveals disease-specific effects.
a, Schematic representation of CUT&Tag experiment design. **b,** UMAP of healthy and RA samples collected from PBMC and synovial fluid. **c,** Volcano plot showing differentially acetylated (H3K27ac) peaks between RA SF and healthy PBMC monocytes. **d, e,** Examples of unexplained GWAS loci that overlap with regions that with higher H3K27ac activity in RA synovial fluid immune cells. **d,** H3K27ac activity at the *FCRL3* promoter is increased in RA SF CD4$^+$ T cells ($\log_2$CPM: 4.02) compared to RA PBMC CD4$^+$ T cells ($\log_2$-fold-change: 1.55, $\log_2$CPM: 2.46, FDR: 0.016) and healthy PBMC CD4$^+$ T cells ($\log_2$-fold-change: 1.72, $\log_2$CPM: 2.30, FDR: 0.0077). For CD8$^+$ T cells, the $\log_2$-fold-change is 1.32 compared to healthy PBMC. **e,** H3K27ac activity at the *ETV7* promoter is increased in RA SF CD4$^+$ T cells ($\log_2$CPM: 6.48) compared to RA PBMC CD4$^+$ T cells ($\log_2$-fold-change: 1.89, $\log_2$CPM: 4.60, FDR: 0.0016) and healthy PBMC CD4 T cells ($\log_2$-fold-change: 2.47, $\log_2$CPM: 4.01, FDR: 5.70×10^{-5}). For monocytes, the $\log_2$-fold-change is 2.02 compared to healthy PBMC. **f,** Heritability enrichment in ATAC-seq peaks (top) and H3K27ac CUT&Tag peaks (bottom) computed using s-LDSC[74]. Error bars represent standard error.

blood mononuclear cells (PBMC) from 6 RA patients and 4 healthy controls, as well as synovial fluid from the same RA patients. We then sorted B cells, $CD4^+$ and $CD8^+$ T cells, regulatory T cells and monocytes using flow cytometry (Methods), and profiled regions marked with H3K27ac using CUT&Tag (**Figure 2.6a**). Using these data, we identified regulatory regions and quantified their activity in 5 immune cell types and 3 different immune contexts corresponding to the peripheral immune context in a healthy state, the peripheral immune context in the disease state, and the immune context at the active site of inflammation. We mapped CUT&Tag 150bp paired-end reads onto the genome using Bowtie 2 and identified peaks using MACS2 for each sample separately[75,76]. We then merged the peaks for all samples, by joining peaks that overlap, to obtain a single consensus peak set that was used for quality control and downstream analyses.

As expected, UMAP visualization of the $\log_2$-transformed read count-per-million ($\log_2$CPM) at the top 30,000 most variable peaks in the consensus set showed separation of the major cell groups (**Figure 2.6b**). In particular, B cells and monocytes formed distinct clusters, while $CD4^+$, $CD8^+$ and regulatory T cells clustered together. Notably, cells from the same biopsy site also formed sub-clusters such that immune cells from healthy and disease PBMC clustered more closely together, while immune cells from synovial fluid clustered separately. Importantly, samples did not cluster according to batch or other technical factors (**Figure S2.12**), indicating that the observed clusters reflect biological differences between cell types and biopsy sites.

We next compared H3K27ac activity between immune cells from the different immune contexts (Methods). The general trend we observed was that H3K27ac profiles in T cells were more different between RA synovial fluid and RA PBMC than between RA PBMC and healthy PBMC. Indeed, we found 2,481 and 2,962 differentially acetylated peaks between RA SF and RA PBMC cells for $CD4^+$ T cells and $CD8^+$ T cells, respectively, compared to the 1,045 and 1,070 differentially acetylated peaks between RA PBMC and healthy PBMC. By

contrast, the H3K27ac profile of monocytes from RA PBMC is more similar to that of RA SF monocytes than that of healthy PBMC monocytes. This finding suggests that monocytes in the peripheral blood of RA patients show similar pathogenesis signatures to synovial fluid monocytes (e.g., at the *IL1B* locus, **Figure S2.13**), and corroborates observations that were made previously using single-cell RNA-seq data[55].

We next studied the 8,117 peaks that showed higher activity in immune cells from RA SF compared to immune cells from healthy PBMC. We found that many of these peaks are located near important genes that are involved in inflammation pathways and disease pathogenesis, such as *CSF1*, which modulates the differentiation of monocytes to macrophages[77], and *IL1RN* (also known as *IL1RA*), which encodes the interleukin-1 receptor antagonist protein that has been associated with autoimmune diseases including RA[78]. Interestingly, IL1RN expression was also found to be upregulated in monocytes treated with synovial fluid from arthritic joints[79]. Overall, we found that genes near peaks with higher activity in RA SF monocytes were enriched in functional annotations such as immune response (P value: 2.96×10^{-14}, hypergeometric test), immune effector process (1.76×10^{-18}), and several pathways including interferon, TNF, NF-$\varkappa$B, and TLR signaling pathways (1.64×10^{-3}, 5.10×10^{-5}, 3.49×10^{-3}, 8.46×10^{-3}, respectively) (Methods). Thus, the H3K27ac profiles of RA SF immune cells revealed elements that appear context-specific and relevant to RA pathogenesis.

We then asked whether differentially active peaks were enriched in unexplained GWAS loci. To answer this question, we overlapped differentially accessible peaks in all immune cells from RA patients with RA GWAS after fine-mapping using SuSiE (Methods)[80]. Strikingly, we found that of the 42 uncolocalized RA GWAS loci, fine-mapped SNPs at 12 loci overlapped with a region with higher activity in RA immune cells (6 loci for healthy PBMC, bootstrap p-value 0.026, Methods). For example, we found that a lead GWAS SNP lies within a differentially active peak at the promoter region of FCRL3 in $CD4^+$ and $CD8^+$ T

cells (FDR: 7.7×10^{-3} for $CD4^+$ and 2.6×10-2 for $CD8^+$ T cells, **Figure 2.6d**). In another example, the RA lead SNP overlaps with an H3K27ac peak located near the promoter of ETV7 which showed higher activity in both RA SF $CD4^+$ T cells compared to the respective cell types from RA PBMC (FDR: 1.6×10^{-3}) and healthy PBMC (FDR: 5.7×10^{-5}). The activity of this regulatory region was also higher in RA SF monocytes compared to healthy PBMC monocytes (FDR: 4.5×10^{-3}, **Figure 2.6e**).

To further assess the relevance of each immune context on the study of disease etiology, we quantified the enrichment of RA heritability in H3K27ac peaks identified in the different immune contexts using stratified LDscore regression[74]. To establish a baseline for comparison, we used accessible chromatin regions identified using ATAC-seq data from unstimulated and stimulated immune cell types (Methods)[71]. Our estimates recapitulated the findings from the original study in which $CD8^+$ T cells and delta gamma T cells showed the largest increase in heritability enrichment subsequent to stimulation (~30-fold vs ~20-fold enrichment for stimulated versus unstimulated)[71]. We then applied the same analysis using our H3K27ac peaks. We found that while the estimated RA heritability was similarly enriched in ATAC-seq peaks from unstimulated immune cells and in H3K27ac peaks from RA PBMC and Healthy PMBC (~20-fold), the heritability enrichment was greater in H3K27ac peaks from RA SF B cells, $CD4^+$ T cells, and Tregs than in ATAC-seq peaks from in vitro stimulation of the same cell types (**Figure 2.6f**).

Our analyses therefore show that there are significant differences in the regulatory landscapes of immune cell types across disease states and immune contexts. In particular, we found that the regulatory landscape of cells extracted from the active site of RA inflammation showed striking differences when compared to that of circulating immune cells in the periphery of both RA patients and healthy individuals. Importantly, we find that the regulatory regions identified in immune cells from RA synovial fluid overlap with many uncolocalized GWAS loci and are the most highly enriched in RA SNP heritability.

Altogether, these observations indicate the importance of studying cell types in the correct disease context in order to elucidate the genetic etiology of a disease.

2.4 Discussion

The goal of this study was to establish a detailed accounting of the effects of genetic variants on gene regulation in immune cells and their overlap with genetic effects on human traits and disease. Recent studies suggested that fewer than a third of GWAS loci colocalize with an eQTL[9,31]. This finding implies that much is left to be understood about the mechanisms by which genetic variants impact human traits.

There are several possible explanations for the small fraction of GWAS loci that colocalize with an expression QTL. Our work evaluated the possibilities that (i) there exist genetic effects on gene regulation other than steady state gene expression levels, (ii) genetic effects are often restricted to cell types and cell states that are causal for the trait, and (iii) genetic effects are often too small to be detected, even in the causal cell types or cell states. These possibilities are not mutually exclusive, but the implications are different for how we should design future human genomics research. For example, if trait-associated variants often impact mRNA splicing but not steady state mRNA expression, then a more widespread focus on mapping the effects of genetic variants on mRNA splicing is needed. If most disease-associated genetic effects are very specific to cell types and cell states that are relevant to the trait, then studying eQTL identified in bulk, unsorted, tissues will have limited success in elucidating the mechanisms underlying most GWAS loci.

Using our harmonized regulatory QTL data, we found that eQTL and sQTL together colocalized with up to 45% of trait-associated loci for the 72 GWAS we analyzed. On average, 40.4% of significant loci from the 50 immune-related GWAS colocalized with a regulatory QTL, a larger proportion compared to an average of 26.4% for the 21 non-immune

GWAS we analyzed (excluding Alzheimers disease, 55.2% colocalized). One of the caveats in our colocalization analysis is the use of the method COLOC. Although COLOC is a very popular method for colocalization analyses, it uses priors that, when altered, can impact substantially the computed posterior probabilities that the causal eQTL and GWAS variants are the same variants. Reassuringly, when we used another colocalization method, HyPrColoc[81], that does not rely on user-defined priors, we were able to replicate nearly all colocalized genes identified using COLOC, indicating that our colocalization analyses are robust and replicable (Supplementary Notes).

Our data also allowed us to ask whether regulatory QTL are likely to be active in many immune cell types, or only in few or a single cell type. We found that at least one third of eQTL (81% of sQTL) are shared across all 15 immune cell types we analyzed from the DICE dataset. For closely-related cell types, we found that the fraction of shared eQTL was as high as 96% (99% for sQTL). Intriguingly, activated and naïve T cells share nearly 70% of detected eGenes. Thus, QTL effects appear similar across many cell types and cell states. One important implication of this finding is that eQTL that colocalize with GWAS SNPs in one cell type are also likely to be active in other cell types. Thus, eQTL that colocalize at a GWAS locus in one cell type should, in general, colocalize in the other cell types. Indeed, after accounting for variability in the posterior probabilities of colocalization reported by COLOC owing to the inherent noise in QTL mapping, we found that the majority of GWAS loci colocalizes with the same QTL in multiple cell types. Altogether, these data questions the notion that the vast majority GWAS SNPs affect gene regulation in a very cell type-specific manner as highlighted in several studies[29,53]. Thus, the use of regulatory QTL from proxy cell types or tissues, e.g., from the GTEx consortium, to identify causal genes may be well justified for a large fraction of GWAS loci.

A noteworthy finding from our colocalization analysis is that genetic variants that impact mRNA splicing often colocalize with a GWAS signal. Indeed, if we considered eQTL

only, our rates of colocalization would be very similar to that of previous studies (26.2% vs 21%)[31]. Instead, when sQTL were tested for colocalization, we found that more GWAS loci colocalized with sQTL than with eQTL. It is worth noting however, that a substantial number of GWAS loci colocalized with both an eQTL and an sQTL. This may be due to horizontal pleiotropy, whereby a genetic locus can influence the expression level of a gene, as well as the splicing of an intron in the same or a different gene. Another possible explanation for this observation is that eQTL effects are often mediated by sQTL or vice-versa. A colocalization analysis for sQTL conditioned on the eQTL would be necessary to tease apart these possibilities but is outside the scope of our work.

Despite a substantial increase in colocalization rates in our study, we find that for most traits, over half of all GWAS loci do not colocalize with a regulatory QTL. Interestingly, we found several differences between genes at colocalized GWAS loci and those at uncolocalized loci. Genes at GWAS loci without colocalized regulatory QTL tend to be more lowly expressed, have higher enhancer density, and are less tolerant to loss-of-function mutations. These findings suggest that genes at uncolocalized GWAS loci may be subject to stronger constraints both at the levels of gene regulation and sequence conservation. Thus, a plausible explanation is that genetic effects at these loci are on average smaller and more cell type or context-specific compared to genes at GWAS loci with colocalization. This hypothesis is consistent with the idea that much larger sample sizes may be required to find the causal QTL effects that explain the associations at GWAS loci without colocalization. That said, our colocalization analyses on QTL datasets with very large sample sizes (DGN: N = 900, eQTLGen: N = 31,684) revealed that the rates of colocalization only increased slightly despite the large increase in our power to detect low-effect QTL. We speculate that an important reason for the modest increase in colocalization is because both DGN and eQTLGen QTL data are from whole blood samples, which are less likely to capture genetic effects that are cell type or context-specific.

One intriguing finding from our analyses is that eQTL identified specifically in in vitro stimulated immune cells from DICE colocalized with only a small number of GWAS loci that did not colocalize with QTL from unstimulated cells. This observation might seem surprising because a recent paper showed that autoimmune disease SNP heritability is more highly enriched in accessible chromatin from in vitro stimulated immune cells compared to naïve immune cells[71]. However, we should note that our data does indeed suggest that SNPs at colocalized and uncolocalized GWAS loci are more highly enriched in open chromatin from stimulated cells compared to unstimulated cells. The differences in the enrichment, however, is negligible, suggesting that stimulation-specific effects can not explain why a large fraction of GWAS loci do not colocalize with the regulatory QTL identified in our study.

One possible explanation for the modest increase in the colocalization rates, when using eQTL identified in stimulated immune cells, is that the immune cells stimulated in vitro only partly recapitulate gene regulation in the in vivo disease context. Thus, although many regulatory elements are primed to be activated subsequent to in vitro stimuli—thereby capturing some of the important regulatory regions relevant to disease—they may require additional factors to fully capture the effects of genetic variants on gene expression levels in the disease context. In support of this, Tsuchiya et al.found that stimulating immune cells in vitro was able to recapitulate gene expression signatures of immune cells from rheumatoid arthritis (RA) patients when 6 different cytokines were used together, but not when the cytokines were used on their own[82].

To better understand the role of context on our ability to interpret GWAS signals, we collected H3K27ac measurements in healthy and RA patients using CUT&Tag to use as proxy for enhancer and promoter activity. Although the sample sizes are too small for a QTL analysis, we were able to use these data to ask whether gene regulatory data in the disease context could aid us to identify putative mechanisms that underlie RA GWAS

hits, in particular for loci with no QTL colocalization. We found that SNPs at 12 out of 42 uncolocalized GWAS loci overlap with regions with increased H3K27ac levels in immune cells from RA synovial fluid. Remarkably, we also found that regions marked by H3K27ac in immune cells from RA synovial fluid were more highly enriched in RA heritability than compared to healthy or RA immune cells collected from peripheral blood. Additionally, our initial analyses suggest that the RA GWAS heritability enrichments in regulatory regions identified in RA synovial fluid immune cells are even higher than in that of in vitro stimulated immune cells. We should note here that caution must be used when interpreting these results as the data type collected in these two studies differ (ATAC-seq versus CUT&Tag). Nevertheless, these preliminary analyses indicate that studying the regulatory effects of genetic variants in the disease context may be critical for discovering the mechanisms behind a large number of GWAS loci without colocalization.

2.5 Methods

Data processing

To harmonize the set of genetic variants across all four datasets, we imputed the genotypes of all individuals in the four studies using the 1000G Phase 3 v5 as a common reference panel (Michigan Imputation Server[83]). Following imputation, only non-duplicated genetic variants with INFO score larger than 0.9 were retained. We filtered variants with Hardy-Weinberg Equilibrium (HWE) p values below 10-5, with missing genotype rate higher than 5%, and with minor allele frequency below 5% using PLINK v1.9[84]. We used the remaining set of variants in all subsequent analyses unless otherwise noted. To exclude outlier individuals, we calculated genotype principal components (PCs) using smartpca[85]. Five outliers in the DICE dataset were identified and removed from downstream analyses.

To quantify gene expression levels, we used Kallisto[86] and summed the transcript per

million (TPM) estimates of all GENCODE 19[87] isoforms to obtain a gene-level TPM. The gene-level TPM were then scaled and quantile-quantile normalized as described before[10]. Gene expression principal components were calculated using the prcomp function in R. To quantify RNA splicing, RNA-seq reads were aligned to the hg19 reference gnome using STAR 2.6.0[88] with the GENCODE 19 annotation. To avoid reads mapping with allelic bias, we used WASP[89] as implemented in STAR 2.6.0 by providing the corresponding genotype data. This is an important step as we found a substantial increase in the number of false positive splicing QTL due to allelic bias in read mapping. Indeed, when reads representing different alleles map to different regions of the genome, QTL mapping will be susceptible to identifying spurious associations between the alleles and read coverage at those genomic regions[58]. Exon-exon junctions were extracted using RegTools[90], and clustered and quantified using LeafCutter[58]. As expected, we observed that the number of exon-exon junctions identified in each sample is positively correlated with the sequencing depth in the DICE consortium (**Figure S2.1**). To harmonize quantification for splicing junction usage across cell types and datasets in all 18 immune cell types, clusters were merged and the merged union was used to re-calculate intron usage in all samples.

MashR analysis in the DICE dataset

To quantify the sharing of eQTL and sQTL in the DICE dataset, we followed the workflow provided by the authors of MashR (`https://github.com/stephenslab/gtexresults`) that was previously described in[57]. Briefly, standard errors of QTL effect sizes were calculated from FastQTL nominal output, which were used together with effect sizes as the input for *mash*. To quantify the correlation structure of the null tests, 30% of all tests were randomly sampled (referred to as the "random" set). To obtain a confident set of QTL for each feature (gene or intron), the SNP with the smallest P-value across all tested SNPs and all cell types were extracted for each feature. This resulted in a feature-by-sample

matrix of effect sizes and their standard errors without missing values referred to as the "strong" set. For eQTL, we included all protein coding genes. For sQTL, we included all introns. Data-driven covariance matrices were calculated from the "strong" set. We then built a *mash* model using the "random" set with the exchange effects (EE) mode to estimate the priors. This model was then applied to the "strong" set to calculate the posterior mean effect sizes (*mash* effect sizes). Significant QTL after *mash* analysis were feature-SNP pairs with local false sign rate (LFSR) below 0.05, as suggested by[57]. The level of QTL sharing was quantified as both overall sharing and pairwise sharing. Overall, sharing was determined to be the number of cell types in which a given feature has a regulatory QTL (LFSR <0.05). Pairwise sharing was quantified both by magnitude and by sign. Share-by-magnitude between two cell types correspond to the proportion of QTL that is significant in one of the cell types and posterior mean effect sizes differ by no more than twofold. Share-by-sign between two cell types correspond to the proportion of QTL that was significant in one of the cell types and had the same sign. The 15 cell types in DICE were grouped into 6 cell groups based on the eQTL sharing-by-magnitude (see **Figure 2.2b**).

Characterization of regulatory QTL

To calculate the distance between eQTL and their target genes, we defined the promoter of each gene as the region 2000 bp upstream and 500 bp downstream of TSS. We tested the enrichment of eQTL in regulatory elements from Ensembl Regulatory Build and consensus ATAC-seq peak set from Calderon et al.[71]. We categorized all ATAC-seq peaks to be either an enhancer or a promoter based on whether they overlap with any promoter region (2000 bp upstream and 500 bp downstream of TSS). The observed and expected number of QTL overlapping with each feature was estimated using the fenrich command from QTLtools[91], and the odds ratios of enrichment were calculated by supplying those

number to Fishers exact test in R. We validated eQTL from DICE in other datasets using $\pi 1$ statistics [92], stratifying eQTL by their levels of sharing across six cell groups estimated by *mash* (specific: in one cell group; intermediate: 2-5 cell groups; shared: 6 cell groups). The 95% confidence intervals of $\pi 1$ was estimated using 1000 bootstraps (i.e., re-sampling DICE eQTL with replacement).

Colocalization

COLOC Colocalization analyses were performed between eQTL/sQTL and 72 publicly available GWAS summary statistics for 11 autoimmune diseases (14 studies), namely, rheumatoid arthritis (RA) [44], Crohns disease (CD) [59,62], ulcerative colitis (UC) [59,62], inflammatory bowel disease (IBD) [59,62], allergy and eczema (AE) [93], asthma, hay fever and eczema (allergy for short) [94], apoptotic dermatitis (ApD) [95], asthma [96,97], systemic lupus erythematosus (SLE) [98] and multiple sclerosis [99]. We also collected 36 GWAS for blood-related traits [100], 11 GWAS related to heart functions and circulation system [101], and several other traits including type 2 diabetes (T2D) [102], Alzheimers disease (AD) [103], Parkinsons disease (PD) [104], estimated glomerular filtration rate (eGFR) [105], height [106], and breast cancer survival [107] and other cancers/neoplasms [101]. We considered the 14 autoimmune and the 36 blood-related GWAS as immune-related, and the rest 22 GWAS as non-immune GWAS.

To assess colocalization between GWAS loci and QTL, we first identified the lead GWAS variants and their flanking region in which colocalization was to be tested. Specifically, all variants available in the GWAS summary statistics were sorted by p-values in increasing order. Starting from the variant with the smallest p-value (lead variant), variants within the 500 Kb window on either side of the leading variant were removed. This resulted in a 1Mbp GWAS locus for colocalization analysis. The same procedure was then applied to the next most significant variant among the remaining variants, until no variant with p value below 10-7 was left. The HLA region (Chr6: 25-35 Mb) was excluded from colocal-

ization. Only GWAS with more than 10 identified loci were included in our analysis. For each GWAS locus identified above, colocalization was tested only if it harbored a regulatory QTL with beta-distribution permuted p value below 0.01 (bpval <0.01) as reported by FastQTL in the 1 Mb window flanking that leading GWAS SNP. Default priors were used for COLOC. We set PP4 >0.75 as the threshold for colocalization. The colocalization proportion was calculated as the proportion of colocalized loci among all identified loci in a GWAS.

Colocalization results were visualized using a function adapted from LocusCompare[108]. For a given locus, SNP with the largest posterior probability from COLOC was defined as the colocalized SNP. r^2 relative to the colocalized SNP were calculated from the genotypes in the QTL study. To visualize the sQTL in the form of a Sashimi plot[109], we first grouped individuals by their genotypes, and then extracted RNA-seq reads that mapped to the cluster that contains the intron to be visualized. To make the coverage comparable between different genotypes, we scaled the read coverage by the number of indivuduals that carry each genotype using the scaleFactor argument in bamCoverage from Deeptools when generating bigWig files[110]. The coverage was then visualized using pyGenomeTracks[111].

Cis-eQTL data of eQTLGen was directly obtained from the website (`https://eqtlgen.org/cis-eQTL.html`)[17]. We also downloaded allele frequencies from 26,609 eQTLGen samples (excluding Framingham Heart Study), which were used in our colocalization analysis. Of note, the DGN dataset is also included in eQTLGen meta-analysis, but does not alter the interpretation of any of our analyses.

HyPrColoc

The GWAS-gene pairs tested in HyPrColoc were selected in the same way as COLOC. We set PP >0.25 as the threshold for colocalization as recommended by the authors[81].

Validation of immune-cell-specific colocalization for non-immune traits

We validated colocalization of 14 non-immune traits (11 heart-related, AD, PD and breast cancer survival) in DICE immune cells using the GTEx V7 eQTL. We first chose several tissues in GTEx that are most relevant to each GWAS trait. For heart-related traits, we chose tissues in heart and circulation system (Artery - Aorta, Artery - Coronary, Artery - Tibial, Heart - Atrial Appendage, Heart - Left Ventricle). For AD and PD, we included the 13 brain tissues (Brain - Amygdala, Brain - Anterior cingulate cortex (BA24), Brain - Caudate (basal ganglia), Brain - Cerebellar Hemisphere, Brain - Cerebellum, Brain - Cortex, Brain - Frontal Cortex (BA9), Brain - Hippocampus, Brain - Hypothalamus, Brain - Nucleus accumbens (basal ganglia), Brain - Putamen (basal ganglia), Brain - Spinal cord (cervical c-1), Brain - Substantia nigra). For breast cancer survival, we used adipose tissues and breast tissue (Adipose - Subcutaneous, Adipose - Visceral (Omentum), Breast - Mammary Tissue). We then identified all the colocalized gene-SNP pairs for these 14 GWAS in DICE, and extracted their P values from GTEx eQTL in the relevant tissues, as well as from DICE eQTL in all immune cell types. Given that a large proportion of eQTL are shared in DICE, we grouped the 15 immune cell types into 6 groups, assigning the smallest P value from all cell types within a given group to that group for each gene. We used Bonferroni correction to adjust P values for multiple testing. Finally, we calculated the proportion gene-SNP pair that has adjusted P value below 0.05 in DICE but not GTEx tissues.

Characterizations of uncolocalized GWAS loci

We restricted this analysis to the loci from the 14 autoimmune GWAS that did not colocalize with any BLUEPRINT QTL. All genes were classified into four categories: genes with an eQTL that colocalized at a GWAS locus, genes that are the closest to a GWAS locus, genes that are closest to a uncolocalized GWAS locus, and all remaining genes. We compared

gene expression level in the three BLUEPRINT cell types separately. The gene expression level values for the three cell types were combined and plotted in **Figure 2.5a**. We also obtained Enhancer-domain score (EDS)[37] and "loss-of-function observed/expected upper bound fraction" (LOEUF)[70] for all available genes and compared the distribution of EDS and LOEUF across the four categories above.

To test the enrichment of uncolocalized loci in ATAC-seq peaks in stimulated immune cells, we constructed a contingency table by counting the number of colocalized and uncolocalized loci overlapping stimulated and unstimulated ATAC-seq peaks, respectively. We then tested the hypothesis that uncolocalized loci were more highly enriched in stimulated open chromatin regions compared to colocalized loci using Fishers exact test. We estimated 95% confidential interval of estimates by bootstrapping uncolocalized GWAS loci 1,000 times with replacement.

We reasoned that regulatory effects of many uncolocalized GWAS loci might be too small to be detected due to small sample sizes. To test this possibility, we ascertained eQTL only at uncolocalized GWAS loci. Briefly, we extracted QTL tests at lead SNP of uncolocalized loci. GWAS locus-gene pairs that have already been tested in COLOC but did not colocalize were filtered. Since it is common for one lead SNP to be associated with many genes, we adjusted the P values by number of tested genes at each loci using Bonferroni correction and picked the gene with the smallest P value. We then calculated the proportion of genes with P value below 0.05. This analysis was applied to each autoimmune GWAS in each cell type in BLUEPRINT dataset.

RA samples collection and analysis

Sample collection and CUT&Tag experiment All of the clinical samples were obtained from Xijing Hospital. Peripheral blood and synovial fluid samples were collected from 6 RA patients at the Department of Clinical Immunology, Xijing Hospital. All of the RA pa-

tients fulfilled the 1987 revised American College of Rheumatology criteria and the ACR 2010 Rheumatoid Arthritis classification criteria[112]. In addition, peripheral blood samples were gathered from 4 healthy individuals. All blood and synovial fluid samples were subjected to gradient centrifugation using lymphocyte separation medium (MP Biomedicals, 0850494) to isolate mononuclear cells, which were cryopreserved for later experiments.

The cryopreserved mononuclear cells were thawed into RPMI/10%FBS, washed once in sterile phosphate-buffered saline (PBS; Beyotime, ST476), and stained with the following antibodies in PBS for 30 min: anti-CD3-APC/Cy7 (Biolegend, 300426), anti-CD4-PE/Cy7 (Biolegend, 357410), anti-CD8-Percp/Cy5.5 (Biolegend, 301032), anti-CD25-PE/CF594(BD Horizon,562525), anti-CD19-FITC (Biolegend,302206), and anti-CD14-APC (Biolegend, 301808). $CD4^+$ T cells (CD3+, $CD4^+$, CD8-), $CD8^+$ T cells(CD3+, CD4-, $CD8^+$), T reg cells (CD3+, $CD4^+$, CD8-, CD25+), B cells (CD3-, CD19+), and monocytes (CD3-, CD14+) were sorted by FACSAria III (BD Pharmingen, San Diego, USA) directly into wash buffer for CUT&Tag, with a maximum of 1×105 cells for each cell type. We profiled H3K27ac (abcam ab4729) for each cell type following the standard CUT&Tag protocol[113]. Samples were processed in different batches, and we ensured to include at least one healthy individual and one RA patient in each batch to minimize batch effects that align with biological differences that we are interested in.

CUT&Tag data analysis

The DNA libraries were subjected to 150bp paired-end (PE) sequencing. Sequencing reads were aligned to human reference genome hg19 using Bowtie 2 with parameters `-local -very-sensitive-local -no-unal -no-mixed -no-discordant -phred33 -minins 10 -maxins 700`[75]. Aligned reads were filtered using Samtools with `-F 1804 -f 2 -q 30`[114]. Samples with fewer than 2M reads were excluded from subsequent analyses. Fil-

tered BAM files for samples that have the same disease status (healthy/RA), tissue-type (PBMC/SF) and cell type were merged. Read coverage was calculated using bamCoverage in 10bp window normalized by RPKM[110]. H3K27ac peaks were called from the merged BAM files using MACS2 with parameters `–format BAMPE –broad –broad-cutoff 0.1 –qvalue 0.1 –extsize 146`[76]. We reasoned that calling peaks from merged BAM files increases the signal-to-noise ratio. To generate a consensus peak set, we merged all the peaks using bedtools merge[115], resulting in 90,412 peaks. We then counted the number of fragments overlapping with the consensus peak set in each sample using featureCounts[116].

Differential peak analysis was performed using limma[117]. We calculated average $\log_2$CPM across samples with the same disease status, tissue-type, and cell type. This average $\log_2$CPM was only used to filter our peaks with low fragments counts. Peaks with average $\log_2$CPM below 2 in all groups were excluded from differential analysis. Then, normalization factors were calculated from the remaining peaks using the TMM method, and counts in each sample converted to $\log_2$CPM. Since samples were processed in different batches, we used ComBat to adjust for batches while including disease status, tissue-type, and cell type as our variable-of-interest. We constructed a contrast matrix comparing RA SF vs. RA PBMC, RA SF vs. Healthy PBMC, and RA PBMC vs. Healthy PBMC in each, and applied the trend method. Differential peaks were defined as $\log_2$-fold-change ($\log_2$(FC)) larger than 1 or smaller than -1, and FDR below 0.1.

We overlapped H3K27ac peaks up-regulated in RA samples with uncolocalized RA GWAS loci. We first fine-mapped RA GWAS summary statistics using SuSiE[80]. Fine-mapping was performed at each locus we used in our colocalization analysis. We supplied GWAS Z-scores, genotype correlation matrix from CEU and GBR from the 1000 Genome Project as the reference panel and the sample size of reference panel to the susie_rss function.

We estimated the enrichment of RA SNP heritability in our H3K27ac peaks using Strat-

ified LD Score Regression (S-LDSC)[74]. We used MACS2 peaks from merged BAM files, which were extended by 500 bp on both sides. To reproduce the heritability analysis from Calderon et al.[71], we used the MACS2 peaks shared by the authors.

2.6 Supplementary Notes for Chapter 2

Patterns of sharing across datasets

To evaluate replication of our findings across datasets, we first verified that QTL that were shared across DICE immune cell types were more likely to be captured in QTL data from other datasets, particularly datasets that consist of RNA-seq from whole blood. To test this, we used Storey's π_1 statistics to estimate the proportion of eQTL identified in DICE that are also eQTL in the other three datasets. In particular, we partitioned the eQTL into six groups representing the level of sharing across the major immune cell lineages sampled by DICE. We estimated that 88.2% ($\sim$2,118 out of 2,401) of eQTL that were shared in all six DICE immune cell groups were eQTL in whole blood from DGN, and approximately 83% in monocyte or T cells from the BLUEPRINT consortium. By contrast, only 8.13-40.4% of eQTL that were detected in only one DICE immune cell-type could be detected in the the DGN or BLUEPRINT data (**Figure 2.7a**). In addition, when we calculated the π_1 statistics starting with only eQTL that were specific to T-cells from DICE, we found that 17.7% were captured in T cells from the BLUEPRINT consortium. Although this rate of replication may seem low, it is 2.2-fold higher than compared to the overall proportion of cell-type-specific DICE eQTL that were captured by BLUEPRINT T cells (**Figure 2.7b**). These patterns of sharing observed are consistent with the sharing pattern and cell-type-specificity patterns of the eQTL as inferred from the DICE dataset. We note also here that, interestingly, up to 40.4% of the cell-type-specific eQTL identified in DICE could be replicated in the DGN dataset, suggesting that a substantial fraction of cell-type-specific

eQTL effects can be detected in RNA-seq from whole blood.

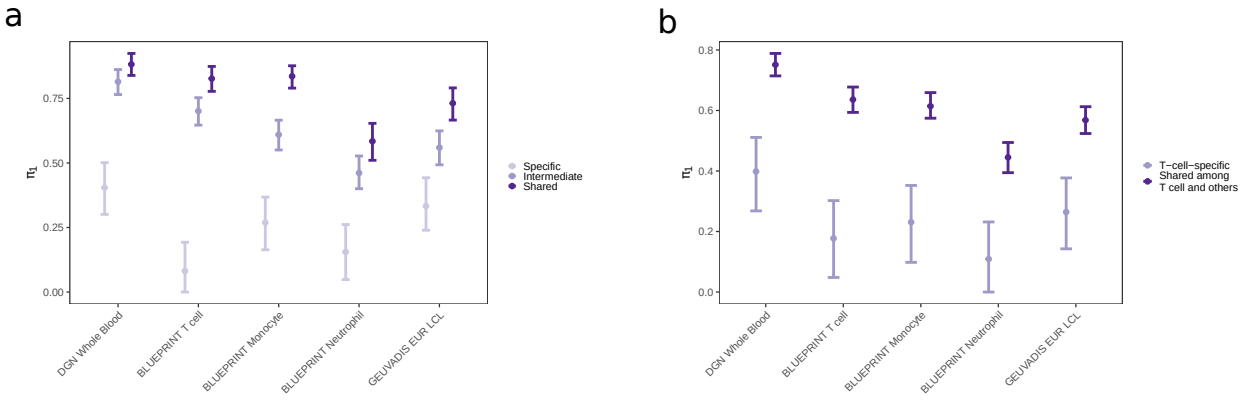

Figure 2.7. Validation of eQTL from 15 cell types or only T cells in DICE.
a, Storey's π_1 statistics measuring replication of shared and cell-type-specific eQTL in DICE across datasets. As expected, DICE eQTL that are shared across all six cell groups are more highly replicable in other datasets than eQTL that are specific to one cell group. Bars represent 95% confidence intervals calculated from 1,000 bootstrap samples. **b**, Similar to **a**, but only eQTL identified in T cells in DICE were validated in the other three studies.

The role of immune cells in non-immune traits

Immune cell types are relevant to many complex traits and diseases. Indeed, coding and regulatory variants in genes that primarily function in immune cells have been linked to autoimmune diseases, e.g. *PTPN22* in rheumatoid arthritis (RA), systemic lupus erythematosus (SLE) and type 1 diabetes (T1D)[118–120]. Interestingly, there are also many ways—perhaps less appreciated—by which immune cells impact non-immunological diseases, including coronary artery diseases (CAD)[121–123], metabolic diseases such as type 2 diabetes (T2D)[124], and neurological disorders[125,126]. For example, microglia cells have been shown to play an important role in the development of Alzheimer's disease[127,128]. We have also recently found that a highly significant Parkinson's disease risk locus is associated with the expression level of *LRRK2* in monocytes, but not in neuronal cells[126], suggestive of a specific effect on immune function. Thus, many, if not most, diseases are associated with a number of risk loci that function through immune cell types.

To identify trait-associated variants that likely act through immune cell types, we obtained eQTL p-values from relevant GTEx tissues for the DICE eQTL SNPs that colocalized with non-immune GWAS loci. Using these GTEx eQTL p-values, asked about the proportion of immune regulatory eQTL that colocalize with non-immune GWAS loci that do not show any effects on gene expression levels in GTEx tissues that are most relevant to each trait. We chose five GTEx heart tissues as the relevant tissues for our 11 heart-related GWAS. For breast cancer, we ascertained the effect of colocalized eQTL SNPs in breast and adipose tissues. For Parkinsons disease (PD), we used the 13 GTEx brain tissues. Overall, we found that 65 of 267 (24.3%) loci that colocalized in the 14 selected GWAS are specific in DICE and BLUEPRINT immune cell data (Supplementary File 5, Supplementary Notes). For example, we found that 8 of 36 eQTL that colocalized with diastolic blood pressure in DICE are not significant in any of the five heart related tissues (P- value > 0.05 / 5) (**Figure 2.8**).

Of note, two genes with colocalized eQTL, *FOLR3* and *LTBP4*, were significantly associated with gene expression levels in GTEx whole blood, suggesting that they indeed likely function through the immune cells. Interestingly, this possibility is further supported by studies that have shown that *FOLR3* is down-regulated in peripheral blood of patients with hypertension[129,130].

Colocalization of immune regulatory QTL with non-immune trait GWAS loci

We found that no more than 30% GWAS loci for T2D, eGFR and height colocalized with our QTL in the BLUEPRINT dataset, whereas more than 50% AD GWAS loci colocalized with a BLUEPRINT QTL. This is consistent with the known role of immune system in AD etiology. The number of genetic loci below a given p-value cutoff varies between our GWAS due to reasons including power and the genetic architecture of diseases. We observed that GWAS loci with lower p-values were more likely to colocalize with QTL. To rule out

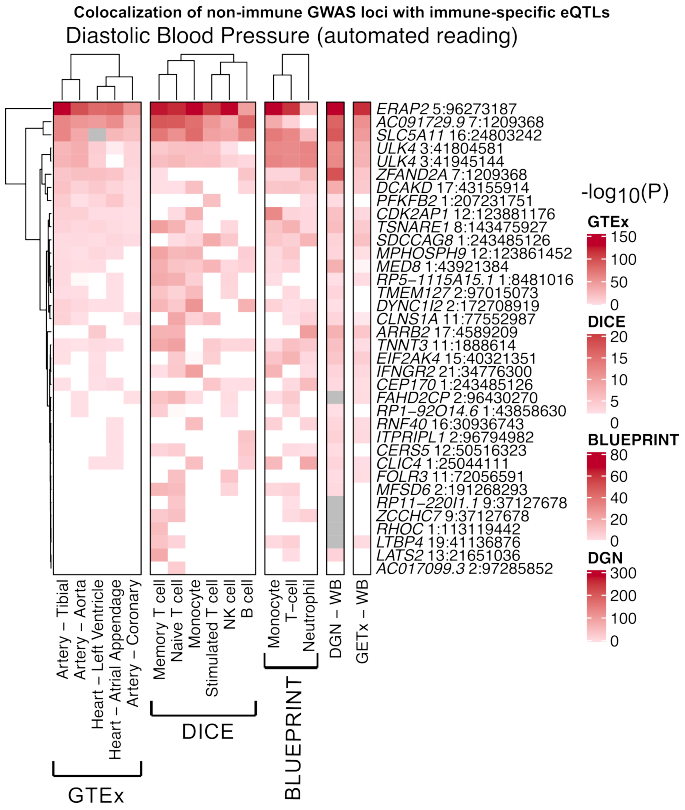

Figure 2.8. Validation of eQTL colocalized with non-immune GWAS in relevant GTEx tissues. Heatmap of eQTL association strengths in GTEx tissues and immune cell types for DICE eQTL that colocalize with diastolic blood pressure loci. Several GWAS loci colocalize with eQTL active in immune cell types but not in heart-related tissues.

the possibility that the difference in colocalization percentages between our autoimmune and non-autoimmune GWAS is due to differences in sample sizes (and therefore p-value distribution), we also calculated the proportion of colocalized loci binned by p-values of GWAS lead SNP. In this analysis we included GWAS lead SNPs with p-values below 10^{-5}. We found that at all p-value bins, the median of percentage of colocalized loci is higher for autoimmune diseases than non-autoimmune traits, and this difference is larger at higher p-value bins (**Figure 2.9**). Interestingly, while many colocalized genes in different autoimmune diseases were shared, indicating partially overlapping disease etiology, they rarely overlapped with colocalized genes in non-autoimmune traits.

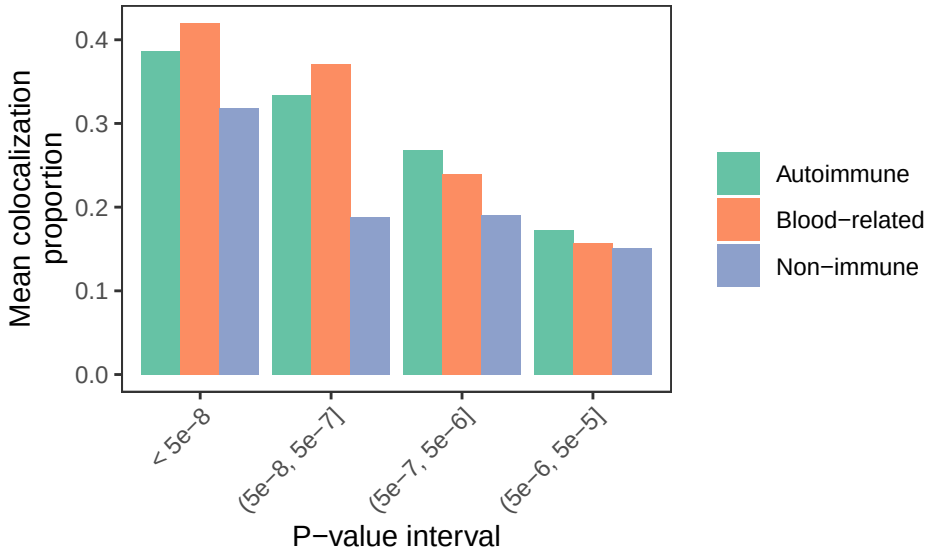

Figure 2.9. Mean colocalization rates for 72 GWAS stratified by P-value bins.

Underestimation of colocalization using COLOC

We found that the number of cell types in which a GWAS locus colocalizes with an eQTL is generally smaller than the number of cell types in which that same eQTL is inferred to be active. We speculated that this discrepancy results from the variation in the posterior probabilities of colocalization computed by COLOC, owing to inherent noise in estimat-

ing the effect sizes and statistical significance of eQTL. In support of this, we found a gene *RNASET2*, whose eQTL colocalized with a CD risk locus in 7 out of the 13 cell types analyzed (PP4 ranges between 0.79 and 0.99), but whose eQTL were inferred to be active across all 13 cell types (**Figure 2.10**).

We found that in the colocalized cell types, the lead eQTL SNP was also the lead GWAS SNP. In the 7 other cell types, the lead eQTL SNPs did not correspond exactly to the lead GWAS SNP, but were in strong LD ($r^2 > 0.6$). As a result of this variation, the posterior probabilities of colocalization (PP4 values) in these 6 cell types ranged from 0.58 to 0.69, which did not pass our cutoff of 0.75. Taken together, these observations suggest that *RNASET2* eQTL colocalize with the Crohns disease GWAS locus in all 13 immune cell types.

Robustness of colocalization

A recent study on COLOC reported that mis-specification of prior parameters can heavily impact the inferred posterior probability of colocalization[131]. To verify the robustness of our colocalization estimates, we performed the same analyses as above using `HyPrColoc` (Hypothesis Prioritisation in multi-trait Colocalization)[81] instead of `COLOC`. Unlike `COLOC`, `HyPrColoc` calculates both SNP-level alignment posterior probabilities and a regional posterior probability. `HyPrColoc` then uses the product of the SNP-level alignment posterior probabilities and the regional posterior probability as the colocalization posterior probability. It further applies non-uniform priors to SNPs in a given genomic locus, which was proposed to be more conservative than `COLOC`[81]. We found that the posterior probabilities calculated in `HyPrColoc` were highly correlated (Spearman's $\rho = 0.86$) to the posterior of colocalization (PP4) estimated using `COLOC`, but were consistently lower (**Figure 2.11a**).

Using the posterior probability cutoff recommended by the authors (0.25), we were able to replicate all colocalized signals identified using `COLOC`. Indeed, `HyPrColoc` found

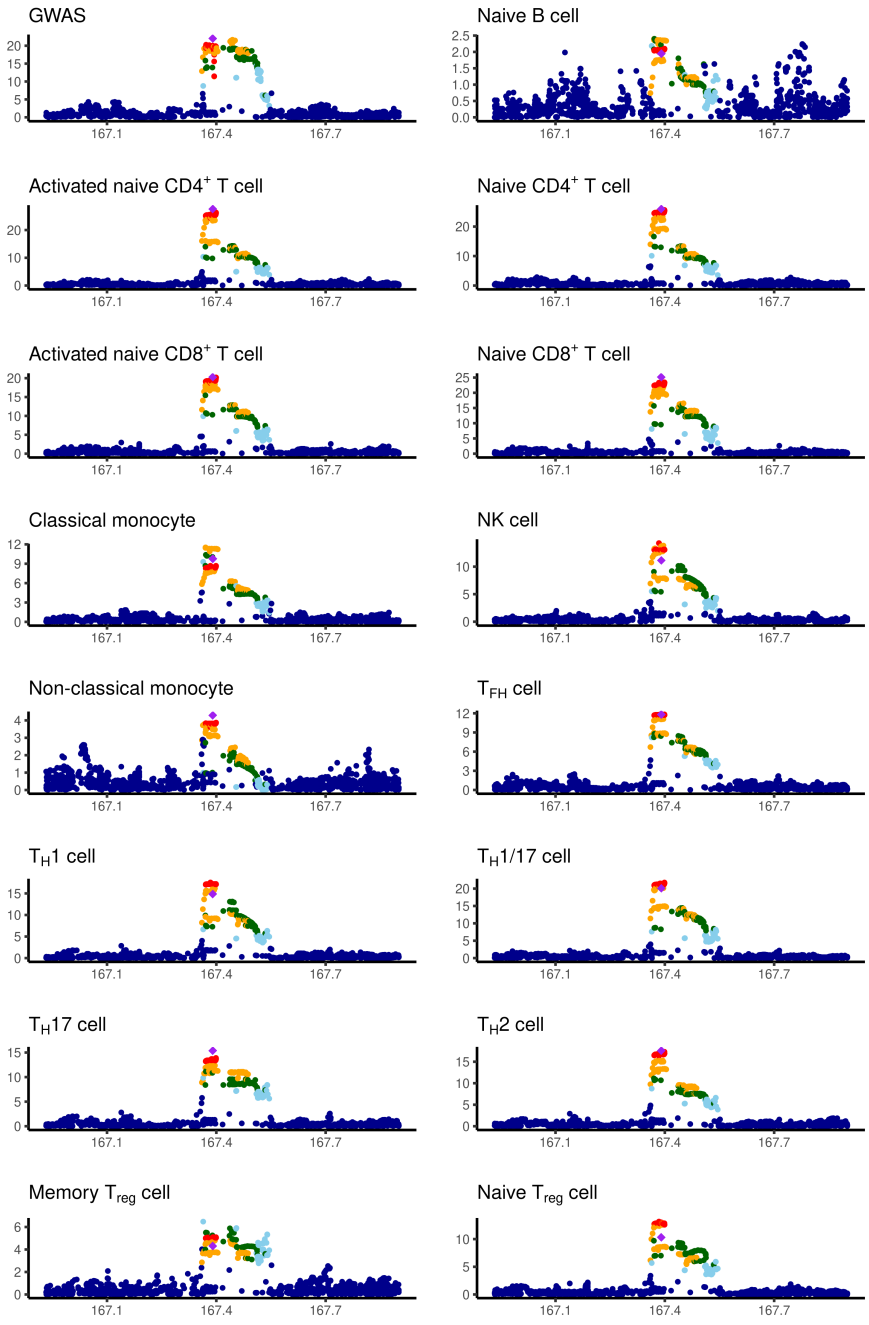

Figure 2.10. LocusZoom plot for an *RNASET2* eQTL and a CD GWAS locus.
The eQTL is shared among all twelve cell types in *mash* but only colocalized in six using COLOC.

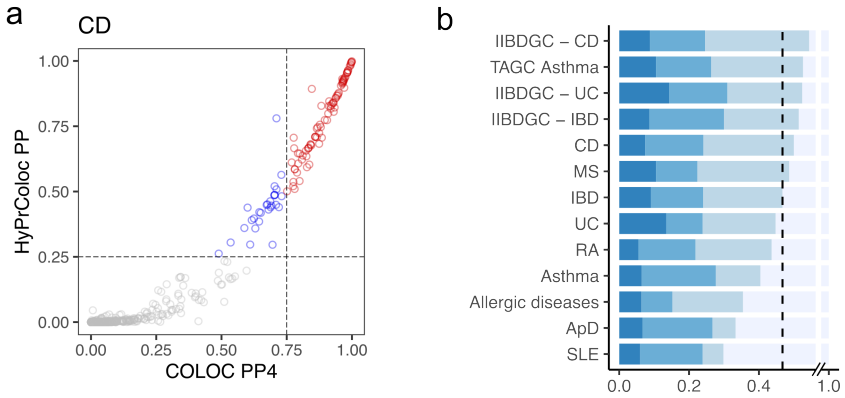

Figure 2.11. Comparing `COLOC` and `HyPrColoc`.
a, Comparison of posterior probabilities of `HyPrColoc` and `COLOC` for CD GWAS. Red: colocalized in both `COLOC` and `HyPrColoc`; blue: only colocalized in `HyPrColoc`; grey: not colocalized. Dashed lines: posterior probabilities used in each method. Similar patterns for other autoimmune were also observed. **b**, The rate of colocalization is on average 44.9% for the 13 IRD analyzed using `HyPrColoc`. AE GWAS was not included in this analysis because SNP effect sizes were not available in the summary statics.

the same number or slightly more colocalized loci when compared to `COLOC` (mean: 44% compared to 40%; **Figure 2.11b**). Interestingly, we found that `COLOC` and `HyPrColoc` yield identical results when the `COLOC` PP4 cutoff was lowered to 0.5 (**Figure 2.11a**). Our re-analysis of colocalization using `HyPrColoc` therefore suggests that our initial `COLOC` results are robust to assumptions on the prior distribution of colocalization probabilities. We thus performed all downstream analyses based on `COLOC` colocalization status.

To better understand the effect of a fixed PP4 cutoff on over-estimating cell-type-specificity of colocalization, we categorized all gene-cell pairs tested for colocalization by (i) whether the eGene colocalizes in at least one cell-type and (ii) whether the eGene is shared in at least four cell categories in DICE. We then compared the PP4 values of these genes in each cell-type. We found that when the eGenes are shared across cell types, the COLOC PP4 values are larger than when the eGenes are not shared for uncolocalized gene-cell pairs. By contrast, this difference was much smaller for colocalized gene-cell pairs (**Figure 2.12**).

While these observations suggests that our PP4 cutoff should perhaps be lowered, we

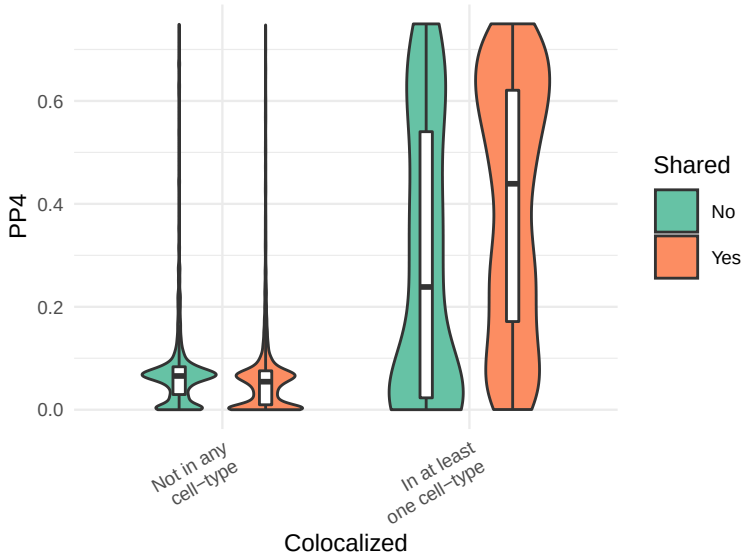

Figure 2.12. Shared eGenes that colocalized in at least one cell-type tend to have larger PP4 in cell types that do not colocalize.
Each entry in the plot represents the PP4 value of an eGene with a GWAS locus in a given cell-type in DICE. Not in any cell-type: the eGenes do not colocalized in any cell-type; In at least one cell-type: the eGenes colocalized in at least one cell-type, but may be uncolocalized in other cell types (PP4 < 0.75). Only uncolocalized tests were included in the plot. Shared: if an eGene is significant in 12 or more cell types. Data from all 14 immune-related GWAS were plotted together.

found that lowering the PP4 cutoff does not solve the overestimation issue as PP4 will be inevitably smaller than any reasonable cutoff in some cell-type. We interpret these findings to support the possibility that the cell-type-specificity of eQTL-GWAS loci colocalization is often overestimated.

2.7 Supplementary Figures for Chapter 2

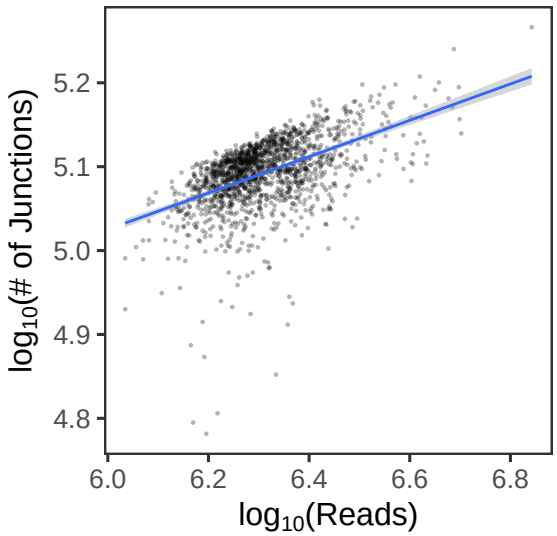

Figure S2.1. Number of exon-exon junctions in each sample is positively correlated with library sizes.
Blue line represents fitted line using a simple linear model.

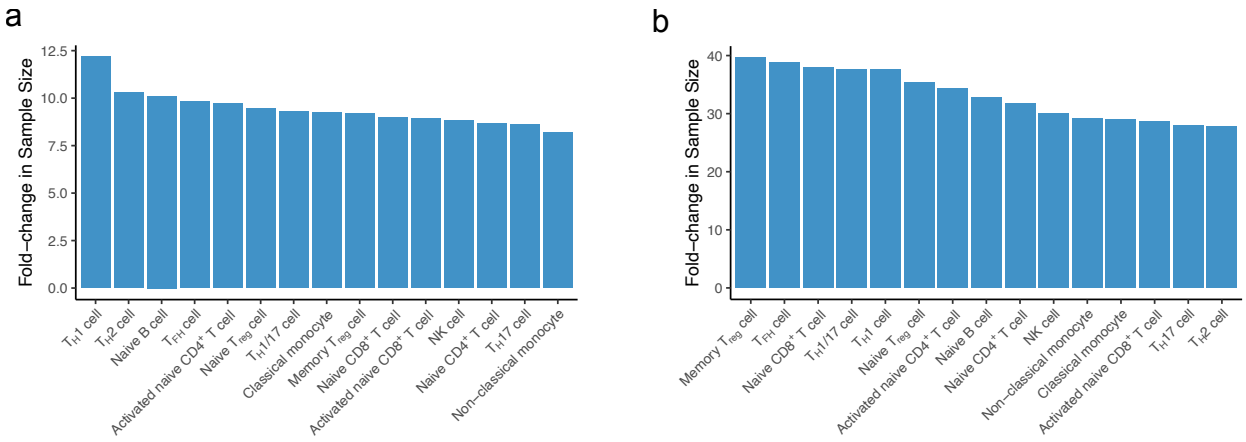

Figure S2.2. Fold-change of effective sample sizes as estimated by *mash*.
a, Fold-change in effective sample sizes for DICE eQTL. **b**, Fold-change in effective sample sizes for DICE sQTL.

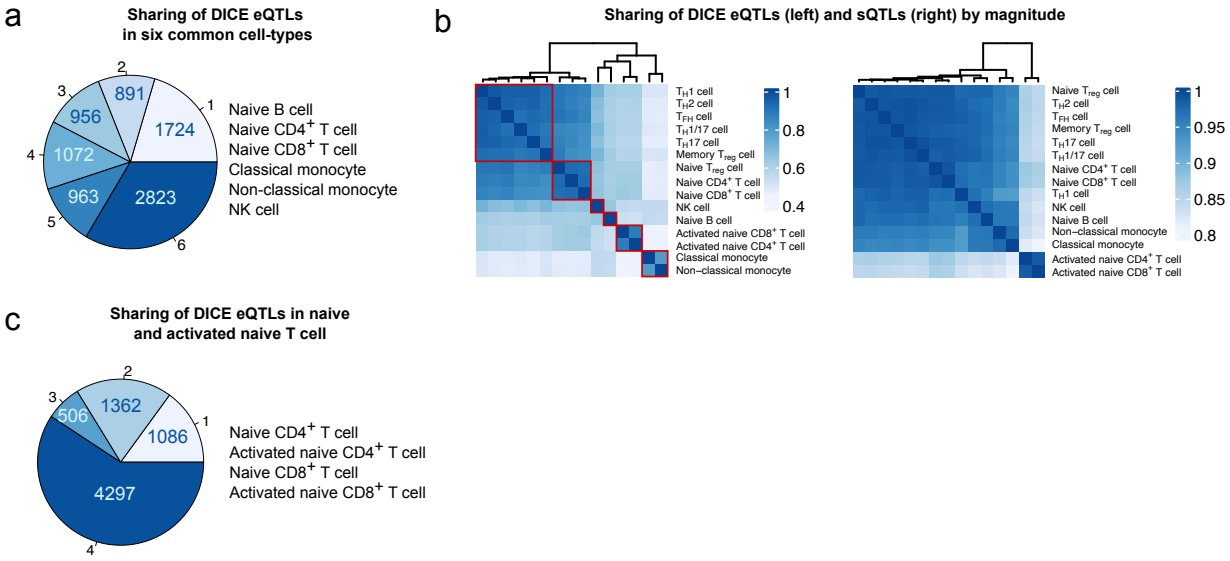

Figure S2.3. Sharing of eQTL and sQTL using *mash* excluding genes in the HLA locus. **a**, Sharing of eQTL among the six common immune cell types in DICE. **b**, Sharing of eQTL (left) and sQTL (right) by magnitude. **c**, Sharing of eQTL among naïve and activated $CD4^+$ and $CD8^+$ T cells.

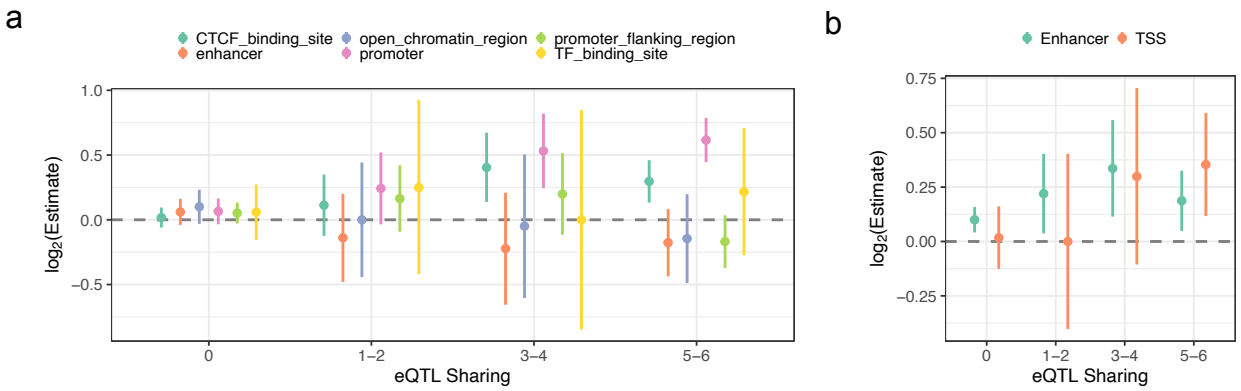

Figure S2.4. Enrichment of eQTL in regulatory elements.
a, Enrichment of DICE eQTL in UCSC Regulatory Build. **b**, Enrichment of DICE eQTL in TSS and enhancers from Calderon *et al.*[71] Bars represent 95% confidence intervals for $\log_2$(Odds ratio).

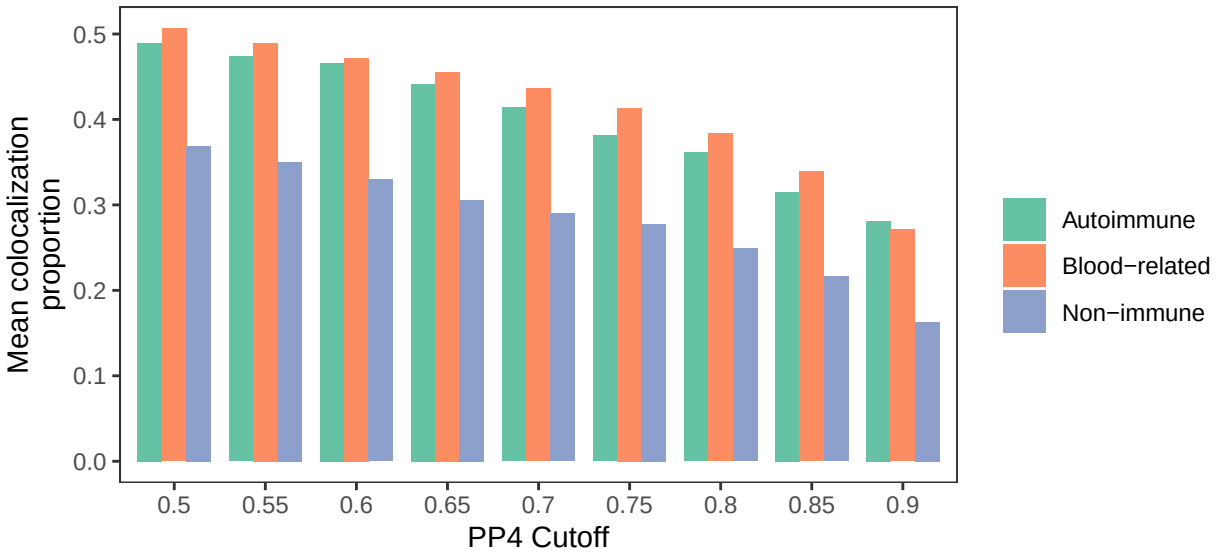

Figure S2.5. Mean colocalization rates as a function of PP4 cutoff in `COLOC`.

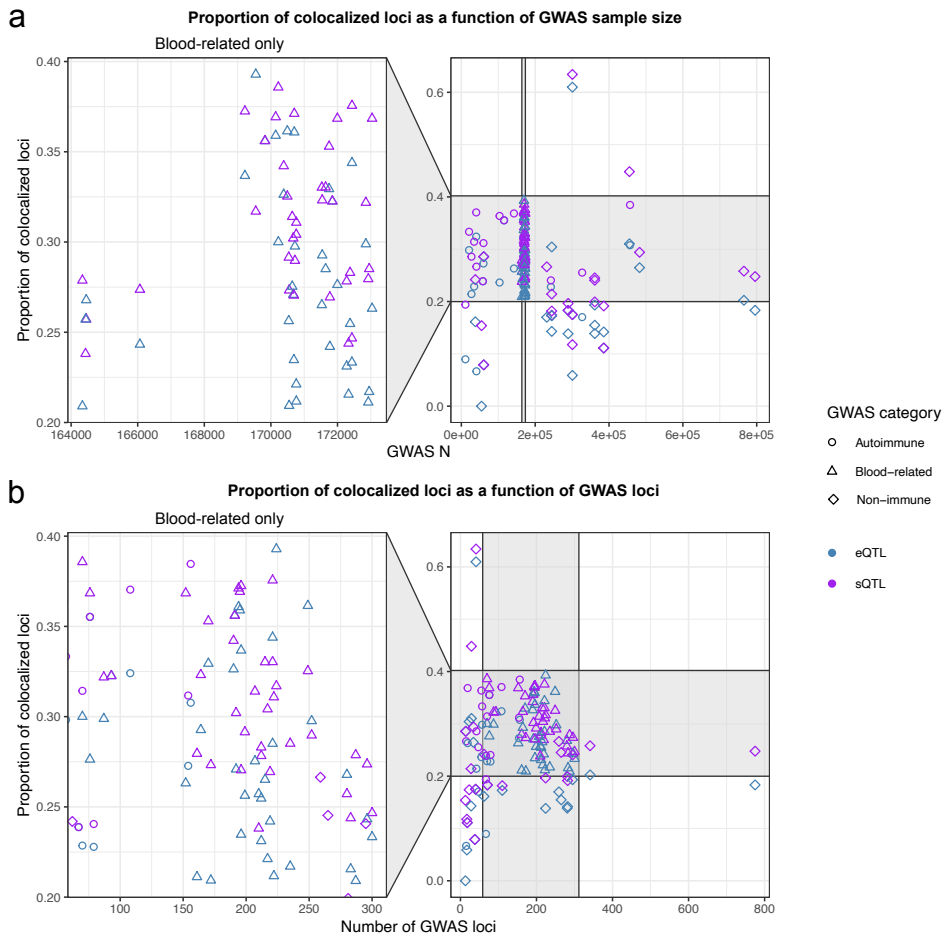

Figure S2.6. Mean colocalization rates as a function of GWAS sample sizes and number of GWAS loci.
Proportion of colocalized GWAS loci is not related to (a) GWAS sample sizes or (b) the number of GWAS loci.

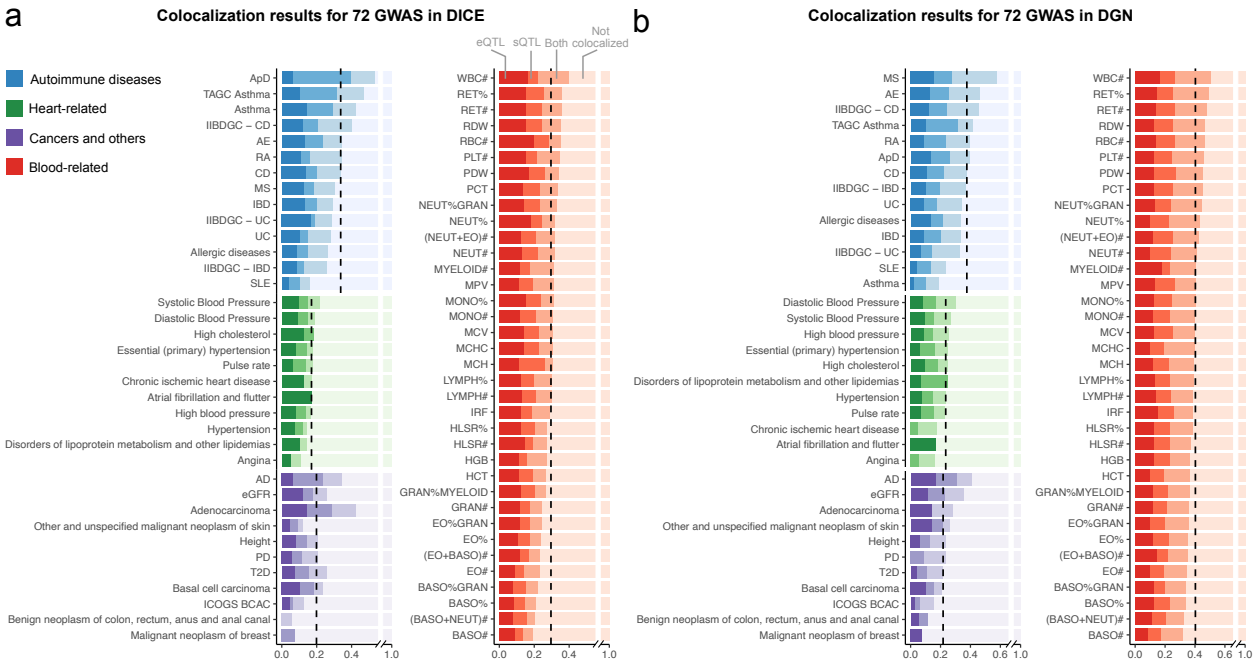

Figure S2.7. Colocalization rates for 72 GWAS in DICE and DGN consortium.
a, Stacking bar plots showing the proportion of GWAS loci colocalized with eQTL, sQTL, both or none in DICE dataset. **b**, Similar to **a**, Stacking bar plots showing the proportion of GWAS loci colocalized with eQTL, sQTL, both or none in DGN dataset.

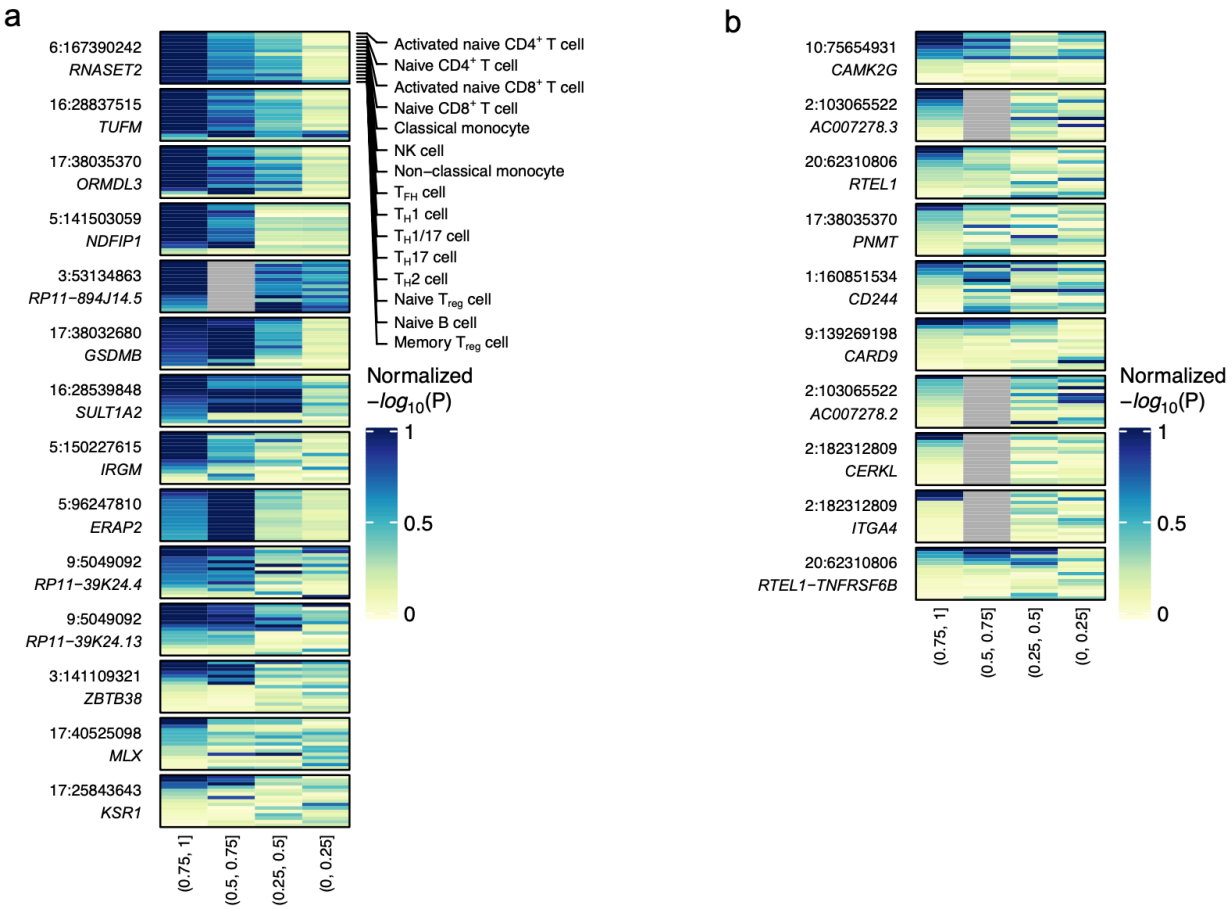

Figure S2.8. Many eGenes colocalized in CD GWAS are shared among the immune cells. **a**, eQTL p-values in different LD bins at GWAS loci with colocalized eQTL across all 15 cell types. **b**, Cell type-specific colocalized eGenes show low p-values across LD bins only in a small number of cell types.

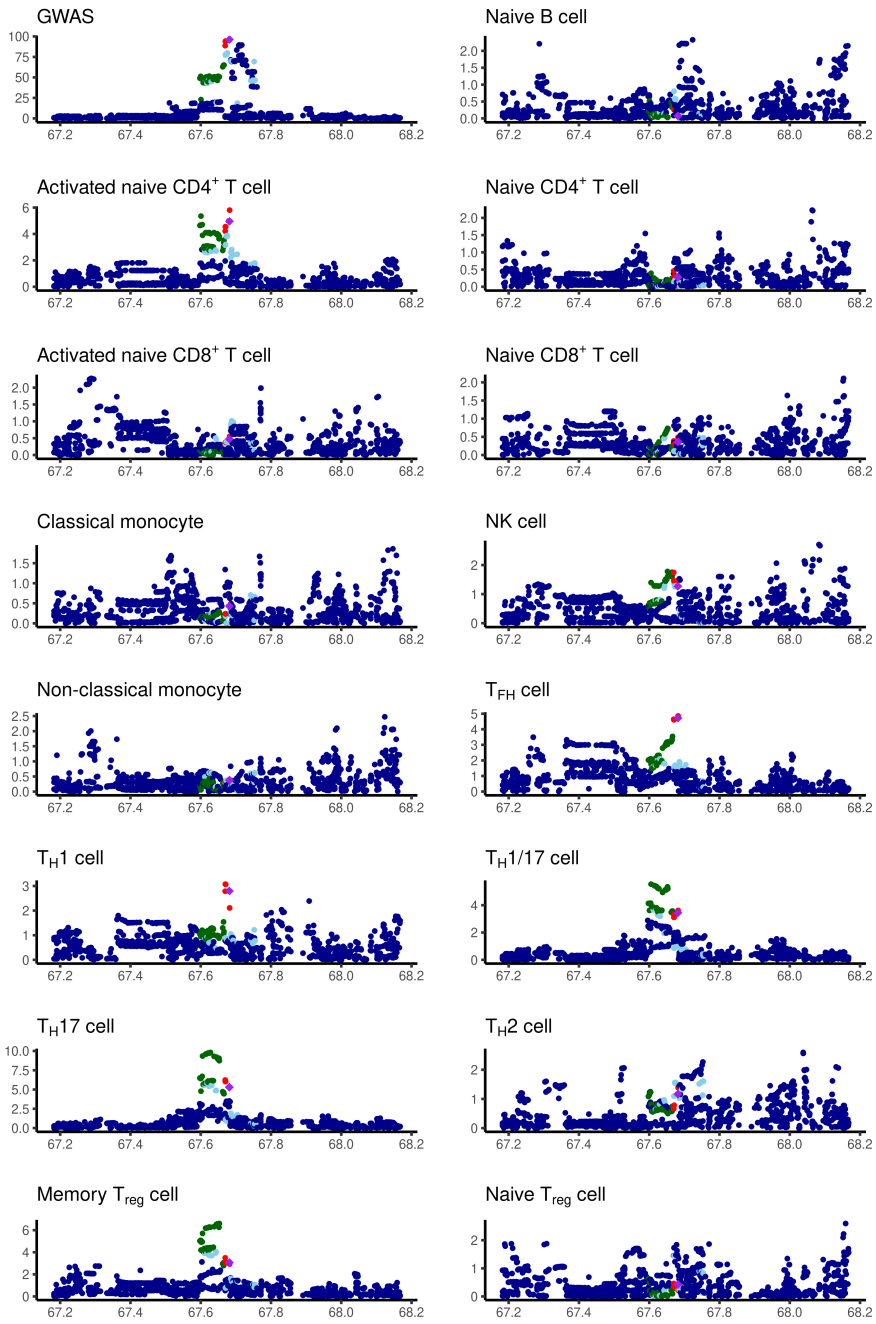

Figure S2.9. LocusZoom plot for *IL23R* eQTL and a CD GWAS locus.
IL23R eQTL colocalized with the CD GWAS locus only in activated naïve $CD4^+$ T cells.

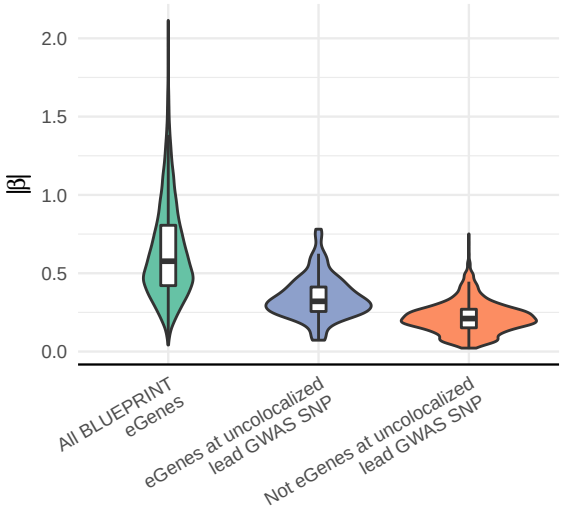

Figure S2.10. Ascertainment of eQTL effect sizes at uncolocalized lead GWAS SNP.
All BLUEPRINT eGenes: all eGenes from eQTL mapping that pass genome-wide multiple testing adjustment; eGenes at uncolocalized lead GWAS SNP: SNP-gene associations at lead SNPs of uncolocalized GWAS loci that were significant QTL after multiple testing adjustment; Not eGenes at uncolocalized lead GWAS SNP: SNP-gene associations at lead SNPs of uncolocalized GWAS loci that did not pass multiple testing adjustment.

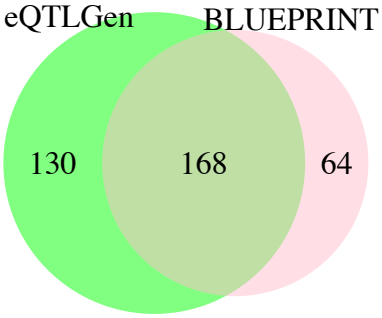

Figure S2.11. Comparison of colocalized loci between eQTLGen and BLURPRINT for 14 autoimmune GWAS.

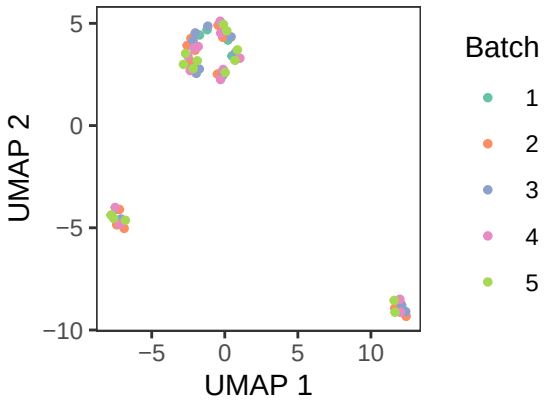

Figure S2.12. Two dimensional UMAP visualization of CUT&Tag read counts in the 30k most highly variable peaks across samples.
Coloring samples by batches show no clustering according to batch.

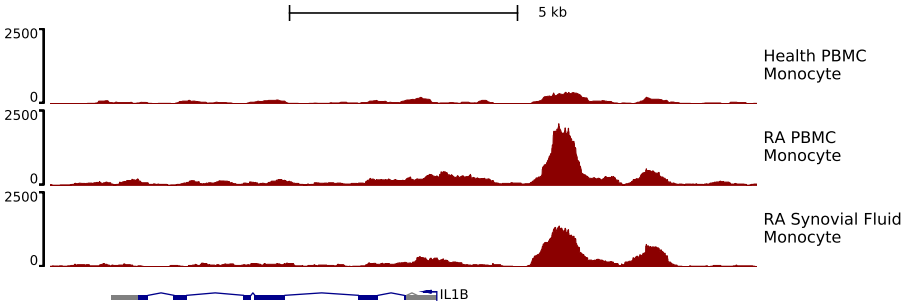

Figure S2.13. Genome tracks of H3K27Ac in monocytes near *IL1B* promoter.
The H3K27Ac profiles in monocytes from PBMC of RA patients near the *IL1B* promoter were more similar to that of monocytes from synovial fluids of patients than to healthy PBMC monocytes.

CHAPTER 3

DYNAMIC EFFECTS OF DISEASE-ASSOCIATED VARIANTS ON CHROMATIN ACCESSIBILITY ACROSS HUMAN IMMUNE CELLS

3.1 Abstract

A small fraction of autoimmunity-associated variants from genome-wide association studies (GWAS) colocalize with steady-state expression quantitative trait loci (eQTL) discovered from bulk RNA-seq studies. To learn more about the remaining associated loci, we created an extensive and unified single-cell chromatin accessibility (scATAC-seq) map in peripheral blood of rheumatoid arthritis (RA) patients and healthy controls, comprising a total of 218,934 cells from 56 individuals. Topic modeling of scATAC count data identified continuous cell states associated with RA risk and disrupted gene activities in RA that are masked in standard cluster analyses, including an effector $CD8^+$ T cell trajectory that tracks the activation of *LILRB1*. We identified 25,107 significant chromatin accessibility QTL (caQTL) at 10% FDR across eight cell groups. Twenty percent of these caQTL were dynamic along cell trajectories defined by our topic analysis. Remarkably, caQTL explained ~50% more GWAS loci compared to eQTL. Combining eQTL- and caQTL-GWAS colocalization allowed us to nominate potential causal genes and contexts at ~55% of colocalized GWAS loci. Importantly, we found evidence that GWAS loci colocalized with a caQTL but no eQTL may reflect cases where the causal effect is through nearby *cis*-regulatory elements but in an as yet uncharacterized context.

3.2 Introduction

Genome-wide association studies (GWAS) have identified thousands of noncoding variants associated with complex traits and chronic diseases like rheumatoid arthritis (RA)[45];

however, uncovering the molecular mechanisms—including the genes and biological contexts—that explain these associations remains a major challenge. To address this, molecular quantitative trait loci (molQTL) in relevant contexts can be integrated with GWAS summary statistics. While *cis*-QTL of transcriptomic phenotypes like gene expression (eQTL) and pre-mRNA splicing (sQTL) have been the focus of many studies, these *cis*-e/sQTL only explain ~40% of GWAS loci[33]. Additionally, *cis*-eQTL from healthy, post-mortem tissues in the GTEx Consortium only mediate ~11% of trait heritability[32], suggesting that current eQTL data are insufficient to fully capture regulatory pathways in complex traits.

These observations reveal two fundamental challenges in dissecting the causal mechanisms of GWAS variants using steady-state *cis*-eQTL. The first challenge is the lack of overlap between GWAS variants and *cis*-eQTL[31,132]. Possible explanations include: (1) insufficient power in eQTL studies; (2) contexts not studied in existing eQTL studies; and (3) genetic regulatory mechanisms independent of expression. Multiple studies have supported these possibilities at explaining specific GWAS loci, but still, the majority of GWAS hits remain unexplained[17,69,133]. The second challenge is—when a GWAS variant does colocalize with eQTL—to identify the causal gene and context through which the variant exerts its function. Due to widespread sharing of eQTL across contexts and horizontal pleiotropy, a GWAS locus typically colocalizes with many eQTL in more than one context[9]. Even if a colocalization is identified in a unique context, there is no guarantee that the context is causal as the same colocalization may be found in another uncharacterized context. Additionally, functionally important genes tend to be regulated by redundant enhancers to maintain a stable expression level under perturbations[37]. Consequently, the genetic effect of an enhancer on a gene can be masked in an important context, but instead revealed in a less important context where fewer enhancers are regulating the gene[134]. This indicates that context-specific eQTL-GWAS colocalization does not neces-

sarily pinpoint the true causal context. Moreover, a recent study demonstrates theoretically that primary, steady-state eQTL and GWAS largely pinpoint intrinsically different categories of genes[34], highlighting the need to study other types of regulatory variants.

Chromatin accessibility can potentially complement existing eQTL studies. For instance, chromatin accessibility QTL (caQTL) mediate roughly twice as much trait heritability as eQTL and colocalize with a larger fraction of GWAS loci compared with eQTL[133]. However, caQTL studies are still scarce and experimental validation is lagging; thus, the target genes and pathways affected by most caQTL remain elusive, hampering the interpretability and functional relevance of these discoveries. Understanding the regulatory consequences of caQTL is also crucial for correctly nominating the causal contexts for GWAS variants using caQTL. Therefore, deeper characterization of the complexity and pleiotropy of genetic effects on gene regulation is necessary for further understanding how current and future molQTL data can be used to elucidate mechanisms underlying GWAS associations.

Advances in novel single-cell genomic technologies now allow for the profiling of transcriptomes and epigenomes of tens to hundreds of thousands of single cells, providing an exciting opportunity to map the functional effects of disease-associated variants onto molecular phenotypes in diverse cell types. In particular, scATAC-seq is able to profile both active promoters and regulatory elements that cannot be studied using scRNA-seq. Although large-scale single-cell eQTL analyses have been performed[38,39,135–137], very few single-cell caQTL studies have been conducted so far[13,138].

To gain additional insights into the molecular mechanisms underlying genetic associations with complex traits, we built a resource of single-cell chromatin accessibility (scATAC) profile in peripheral blood mononuclear cells (PBMC) consisting of 218,934 cells from 56 individuals. We collected original data from 16 rheumatoid arthritis (RA) patients and 17 healthy donors and incorporated two public PBMC scATAC-seq datasets[13,139]. We

show that topic modeling of scATAC count data can reveal cell state continuum not represented by cell clusters and thus can be a powerful alternative to cluster-based analyses; for example, we found an effector phenotype trajectory leading to a rare $CD8^+$ T cells population in RA patients that implicates the activation of disease genes such as *LILRB1*. We next leveraged pseudobulk allelic imbalance and single-cell counts in a Poisson mixed-effects model to identify 25,107 significant caQTL (10% FDR) in common immune cell types, quadrupling the number of significant genetic effects on chromatin accessibility identified in previous studies[13]. Remarkably, our caQTL colocalized with ~50% more GWAS loci than bulk eQTL, and enabled us to better disentangle the genetic regulatory mechanisms that underpin autoimmunity-associated variants.

3.3 Results

Multiplexed single-cell chromatin accessibility profiling of peripheral blood mononuclear cells in healthy and RA patients

Severe technical batch effects in single-cell genomic experiments complicate differential expression and accessibility analyses across conditions when samples are profiled in different experiments[140]. To overcome this difficulty, we sought to pool samples from healthy and RA patients in a single library and to use naturally occurring genetic variation across the donors to identify each cell's donor identity. To test whether this strategy could work for scATAC-seq-as it has already been shown to work well for scRNA-seq-we obtained whole-blood scATAC-seq data from a pool of four individuals, and quantified the number of reads overlapping a common SNP from the 1000 Genomes Project for each individual cell. A median of 1,169 and 231 SNPs were covered by at least one or two reads, respectively. As expected, the number of SNPs covered by reads is positively correlated with the number of unique fragments in each cell (**Figure S3.1a**). Moreover, 90.5% of

the cells that have >1,000 fragments had at least 100 SNPs covered by at least two reads (Supplementary Notes). These results indicate we would obtain highly accurate data by demultiplexing cells from a typical pooled scATAC library comprised of samples from several individuals.

To obtain a comprehensive map of chromatin accessibility in peripheral blood mononuclear cells (PBMC) from healthy donors and rheumatoid arthritis (RA) patients, we collected scATAC-seq data from 16 RA patients and 17 healthy controls. We pooled samples from two to four donors, except for one healthy sample that was processed separately. We combined RA samples with at least one healthy PBMC sample to separate batch effects from potential disease effects (**Figure 3.1a**). To allow for accurate demultiplexing, low-pass whole genome sequencing (LP-WGS) was performed for all 33 individuals, followed by genotype imputation using GLIMPSE[141]. This allowed us to assign a donor to 73,980 (59.8%) of the filtered 123,714 cells with high confidence, as well as to remove 34,021 (27.5%) likely doublets.

Depending on whether two or four donors were pooled, the number of high-quality cells from each individual ranged from 718 to 4,441 (mean: 2,228). We projected all the cells onto 50 dimensions using Latent Semantic Indexing (LSI). To integrate data across experiments and remove batch effects, we used `fastMNN` to adjust the LSI reduced dimensions, because it is more resilient to erroneous removal of biological variations compared to other methods[142]. Next, we projected fastMNN-adjusted reduced dimensions to a two-dimensional UMAP embedding. Visualizing the harmonized data in the UMAP space, we found that cells from all individuals are well-mixed except for two healthy donors (see below; Supplementary Notes) and there is no separation between fresh and frozen samples (**Figure S3.1b-c**). To annotate cell types in scATAC-seq data, we reanalyzed scRNA-seq data we previously collected from PBMCs of three RA patients, one ankylosing spondylitis (AS) patient control, and one healthy control[55]. Azimuth was employed to label cell

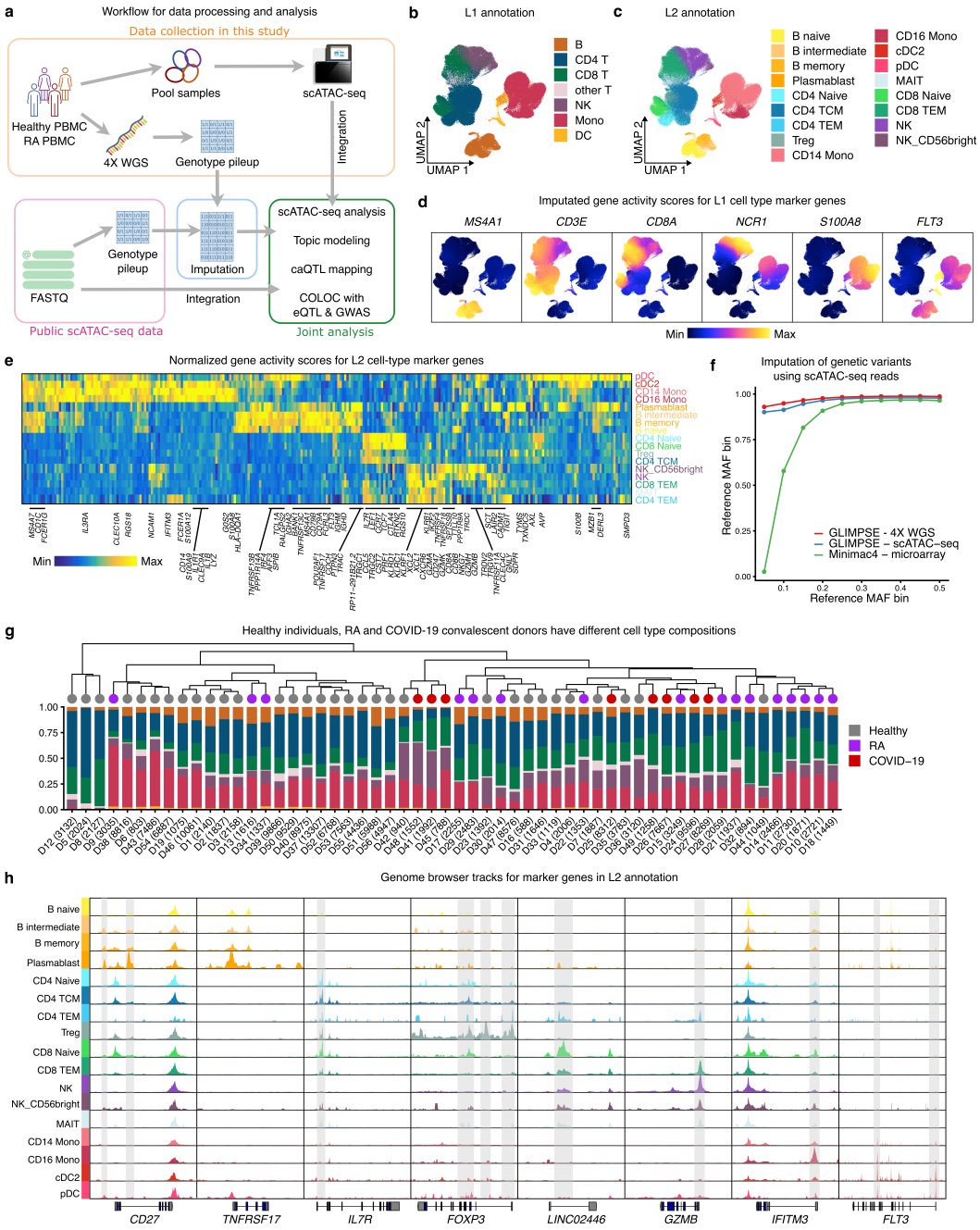

Figure 3.1. Integrated map of scATAC-seq of PBMC from three studies in 56 donors. **a,** Schematic of study design and analysis workflow. **b,** A UMAP embedding of all cells colored by `Azimuth` L1 annotation. **c,** The same UMAP embedding as in **b**, colored by `Azimuth` L2 annotation. **d,** Gene activity scores of marker genes in the common immune cell types. **e,** Heatmap for marker genes for cell subtypes in L2 annotation. **f,** Comparison of imputation quality (INFO) score from either low-pass WGS or aggregated scATAC-seq using GLIMPSE and from DNA microarray using Minimac4. **g**, L1 cell type compositions. Samples are clustered by distances in scaled proportions; bar colors are the same as **b**. **h,** Genome browser tracks of scATAC-seq reads around marker genes. Shaded regions highlight cell type-specific open chromatin regions.

types in the scRNA-seq data[143], and we transferred these labels onto cells in our integrated scATAC-seq dataset by integrating with scRNA-seq data[144]. Using this method, we identified a low number of cells for rare cell types such as Tregs (183), only accounting for 0.228% of all PBMCs in our data. We reasoned that low total cell number could limit the ability to confidently identify rare cell types (Supplementary Notes), we thus sought to expand our dataset by integrating additional datasets to our study.

Harmonization of chromatin accessibility profiles and genotypes across three studies

We supplemented our scATAC-seq dataset with two recently published PBMC scATAC-seq datasets to improve the power and utility of our study[13,139]. After assessing quality metrics such as TSS Enrichment scores and unique fragment counts per cell, we concluded that the three datasets were comparable in quality (**Figure S3.1d**). This prompted us to integrate the three datasets for joint analysis, thereby increasing the sample size from 33 to 56 donors (13 healthy donors, Benaglio et al.[13]; two healthy donors and eight convalescent COVID-19 donors; You et al.[139]).

To integrate the three datasets, we performed joint LSI to map cells from the three studies to a shared latent space and again used 'fastMNN' to remove batch effects. After integration and filtering, we retained 218,934 high-quality single-cells. We identified 28 distinct cell clusters in this integrated dataset. We called a unified peak set consisting of 287,567 *cis*-regulatory elements (cREs) 500 bp long on chr1-22 and chrX from the 28 cell clusters using the `ArchR` implementation of MACS2[76]. We again used `Azimuth` to annotate cell types in this integrated dataset. The final annotation had two levels of granularity: the L1 level contained seven common immune cell types, while the L2 level contained 17 cell types/subtypes that are well represented in our data (**Figure 3.1b-c**). Annotated cell types by `Azimuth` and Leiden clusters showed a high degree of agreement (**Figure S3.1e**;

adjusted Rand index: 0.42). In this annotation, we captured 3,961 Tregs, accounting for 1.81% of all PBMC, more consistent with known Treg compositions in peripheral blood (1-4%) (Supplementary Notes). This indicates accurate integration and cell type identification in our integrated dataset. To further validate our cell type annotation, we assessed gene activity (GA) scores of marker genes in each cell type and observed high GA scores for *MS4A1* in B cells, *CD3E* and *CD8A* in T cells, *NCR1* in NK cells, *S100A8* in monocytes, and *FLT3* in DC (**Figure 3.1d**). We also visualized genome browser tracks for markers specific to L2 cell annotations, which show patterns that are broadly consistent with the annotated cell types (**Figure 3.1e**).

We compared all three donor groups (healthy, RA, and COVID-19 convalescent) to determine if there were any differences in cell type compositions. While the overall cell type compositions were similar among all individuals, there were a few notable exceptions (**Figure 3.1f**). For example, three COVID-19 convalescent donors (D41, D45, D48) had expanded $CD8^+$ T cell and NK cell populations, consistent with the original study[139] (**Figure 3.1f**). In addition, three healthy subjects (D5, D8 and D12) showed significantly lower B cell and monocyte compositions but an increased proportion of $CD8^+$ T cells. These individuals also shared upregulated effector and pro-inflammatory genes including *IFNG*, *IL23R*, *IL17A* and *IL12RB2* (**Figure S3.1f**). We hypothesized that these healthy individuals may be enduring underlying chronic inflammation. Notably, two of these donors also formed their own clusters when the 33 individuals we collected in this study were analyzed together (**Figure S3.1b**). This demonstrates that our multiplexed experimental design is capable of disentangling individual effects from batch effects.

To identify chromatin accessibility quantitative trait loci (caQTL) in our harmonized dataset, we sought to obtain genotypes for scATAC-seq samples from the two public datasets that were not multiplexed. We hypothesized that aggregated scATAC-seq reads may provide sufficient coverage for genotype imputation[145]. Remarkably, our low-pass WGS li-

braries and aggregated scATAC-seq libraries have similar coverage profiles across the genome (mean: 4.08 for 4X-WGS, 6.75 for scATAC-seq; **Figure S3.1g**). Therefore, we adapted the GLIMPSE imputation pipeline to aggregated scATAC-seq reads and compared imputed genotypes from scATAC-seq reads using GLIMPSE with those imputed from microarray in the original study[13]. We found that GLIMPSE-imputed SNPs from low-pass WGS and scATAC-seq both had consistent high quality scores across all reference MAF bins, whereas Minimac4-imputed SNPs were of significantly lower quality for relatively rare SNPs (reference MAF < 0.2, **Figure 3.1g**). In addition, genotype dosages imputed from scATAC reads and microarray were highly correlated across all reference MAF bins (higher than 91%, **Figure S3.1h**), suggesting that scATAC-imputed genotypes are accurate and are not biased by allelic-imbalance signals in chromatin accessibility. Consequently, we merged genotype likelihood estimations from all three studies and performed joint imputation using GLIMPSE. This enabled us to generate a harmonized callset of 6.3 million high-quality SNPs for joint caQTL mapping (**Figure S3.1i**).

In summary, we constructed a map of accessible chromatin from 218,934 peripheral blood cells from 56 individuals and generated high-quality, harmonized genotype information for all the individuals, enabling fully-integrated downstream analysis. Surprisingly, we found that imputation using aggregated scATAC-seq reads offers high-quality genotype information and is superior to microarray-based methods; this workflow can be easily adopted for future population-scale scATAC studies.

Topic analysis of chromatin accessibility defines cell types and states

One benefit of single-cell data is that it can capture rare cell types as well as transitional cell states. But typical single-cell data analysis aggregates cells into discrete clusters, essentially ignoring the heterogeneity among cells belonging to the same cluster. On the other hand, topic modeling on the single-cell count matrix can represent each

cell as a grade of membership (referred to as "loadings" hereafter) to multiple topics, or "components"-akin to admixture analysis in population genetics. Each topic captures an axis of variation in the data, which can represent cell types, contexts, or biological processes, enabling straightforward interpretation. Topic modeling also estimates the probability of a cRE belonging to a topic (referred to as "scores" hereafter), which can be framed as a differential accessibility analysis for cREs between topics[146]. Therefore, cREs with the highest scores in each topic can be analyzed to reveal associated biological functions.

We applied a new topic modeling approach, fastTopics[147], to our scATAC count matrix with various numbers of total topics (referred to as "k" hereafter, Methods). As expected, the number of total topics greatly influences which biological processes are captured by different topics. For example, when k=6, the topics largely represented common immune cell types including B cells, $CD4^+$ T cells, $CD8^+$ T cells/NK cells, and monocytes; topic 3 (k3) was observed in both $CD8^+$ T cells and NK cells, likely capturing the cytotoxic signatures shared between these two cell types (**Figure 3.2a**). When we increased the number of total topics to 12, k3 became private to NK cells and was replaced by k9 in cytotoxic $CD8^+$ T cells (**Figure 3.2a**). We chose 20 topics for all downstream analyses as the majority of common cell types and states in our PBMC data were represented under this parameter (**Figure 3.2a**, **Figure S3.2**).

Both cell-level loadings and cRE-level scores offer rich information for biological interpretations. We first asked whether topics largely represent cell types annotated using Azimuth. We calculated the mean loading for cells belonging to the same annotated cell type and found multiple topics that were almost exclusive to given cell types, including B cells, monocytes, and DC (**Figure 3.2b**). We note that the same topic can represent both a cell type and a biological program (e.g. the topic for memory $CD4^+$ T cells also represents T cell memory program). We found that T cell subsets are often mapped to more than one

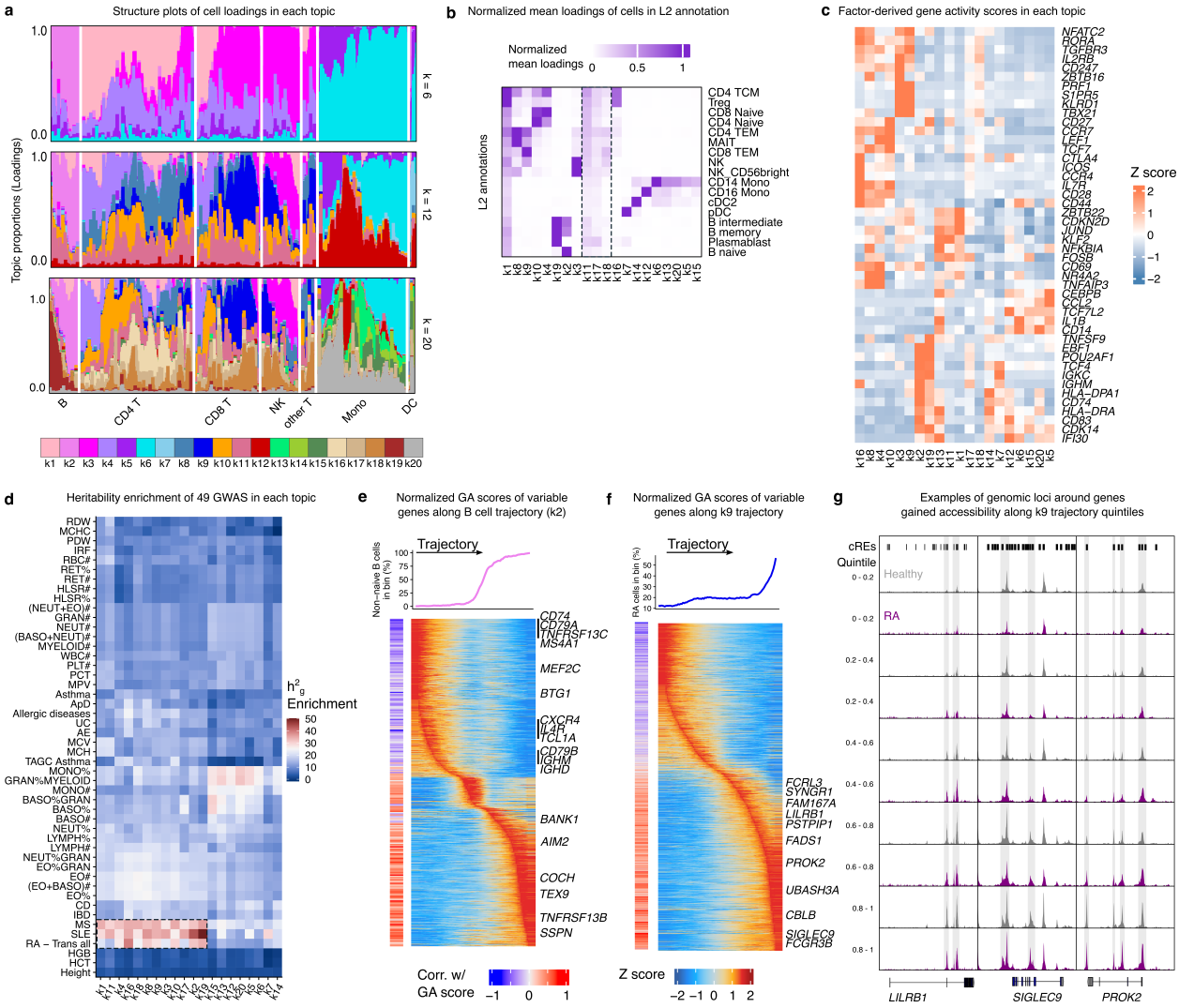

Figure 3.2. Topic modeling helps interpretation of inter-cellular and inter-individual variation in scATAC-seq profiles.

a, Structure plots of topic loadings in 2,000 randomly selected cells when fitting 6, 12 and 20 topics. Cells were grouped by seven common immune cell types to highlight coarse-grained differences in topic loading among cell types. **b,** Heatmap showing the average loading for each topic in each cell types in L2 annotation. **c,** Heatmap of scaled gene-level scores calculated from cRE-level scores in each topic. **d,** Heritability enrichment score of 49 GWAS in top 10% cREs for each topic. **e,** Top, smoothed percentage of non-naïve B cells in trajectory percentile. Heatmap on the left shows the Spearman's correlation between GA scores and memory B cell trajectory; heatmap on the right shows row-normalized GA score changing along the trajectory. **f,** Similar to **e**, showing the trajectory and relevant genes in RA-associated topic k9. **g,**, Genome browser tracks of the genetic region around three genes (*LILRB1*, *SIGLEC9* and *PROK2*) that progressively gained accessibility along k9 trajectory. Cells are grouped by disease status and k9 quintiles.

topic, highlighting the difficulty to precisely annotate T cells based on scATAC-seq data alone. Interestingly, several topics did not clearly correspond to any specific cell type. For example, topic 1 (k1) was ubiquitously present in all cell types. We found that k1 loadings are highly correlated with TSS Enrichment score and the top 1,500 cREs with the highest scores in k1 were over-represented in promoter regions (p-value = 0, hypergeometric test; **Figure S3.3a,b**), suggesting that k1 likely represents single-cell data quality rather than biological variation. Similarly, k11 and k18 showed weak, but widespread signals in B cells and T cells.

To facilitate functional annotation of the topics, we derived a gene-level score in each topic from the cRE-level scores. We tested four cRE-to-gene mapping strategies: (1) cREs directly overlapping with TSS; (2) the nearest cRE to a TSS; (3) cREs co-accessible with TSS, and (4) distance-based exponential weighting function similar to GA score calculation in `ArchR` (**Figure S3.3c**, Methods)[148]. We benchmarked the four methods using k4 as ground truth, as it indubitably represents naïve T cell programs. We found that the distance-based exponential weighting function from `ArchR` captured the most known marker genes for naïve T cells (**Figure S3.3c**). Thus, we used this strategy to identify a set of genes driving each topic (**Figure 3.2c**). We identified well-known cell type markers among topic-driving genes, including *EBF1* and *CD83* for naïve B cells (k2); *CD27* and *TNFSF9* for memory B cells (k19); *CD247* and *ZBTB16* for NK cells (k3); *S1PR5*, *KLRD1*, *PRF1* and *TBX21* for cytotoxic $CD8^+$ T cells (k9); and *ICOS* and *CTLA4* for Treg (k16) (**Figure 3.2c**). We then performed gene-set over-representation analysis on the top 500 genes driving each topic. Interestingly, k1 was enriched in mitochondrial-related pathways, while k11 was enriched in splicing pathways. We conclude that these topics capture specific biological processes, rather than cell type identities, which is consistent with the weak, but widespread loadings in multiple cell types (**Figure S3.2**). Finally, we tested the enrichment of transcription factor (TF) binding motifs in the top 3,000 cREs for each topic

(Methods). Again, TFs known for several immune cell types and states showed significant enrichment in the corresponding topics. Interestingly, the most enriched TF motif in topic k17 is BATF (**Figure S3.3d,e**). Although BATF mostly functions in monocytes, its specific enrichment in a small group of k17-associated $CD8^+$ T cells was intriguing. This group of cells were also enriched for the expanded $CD8^+$ T cells from the three outlier healthy donors discussed previously (D5, D8 and D12). Consistent with this, the BATF motif showed deviation in its binding activity in these $CD8^+$ T cells (**Figure S3.3f**). Indeed, BATF is responsible for mediating effector $CD8^+$ T cell differentiation and response during chronic viral infection[149–151], and ectopic overexpression of BATF in T cells drives proliferation[152]. This is in line with our hypothesis that these healthy individuals were experiencing chronic inflammation at the time of sampling (see **Figure 3.1f**, **Figure S3.1b,f**). These lines of evidence suggest that k17 may represent BATF-driven programs in a small subset of $CD8^+$ T cells.

We next studied the disease relevance of each topic. We used stratified LD Score regression (s-LDSC) to calculate heritability enrichment for 48 GWAS (11 immune-related diseases, 36 blood phenotypes and height) in the top 10% cREs in each topic (Methods)[74]. As expected, we observed higher h2g enrichment in many immune-related diseases and blood phenotypes compared to height, a negative control trait (**Figure 3.2d**). In particular, we found large h2g enrichment for three autoimmune diseases (RA, systemic lupus erythematosus [SLE] and multiple sclerosis [MS]) in the lymphoid-related topics (**Figure 3.2d**). For SLE, B cell topics (k19 and k2) were the most enriched, consistent with a recent bulk ATAC-seq study and the known crucial role of B cells in SLE etiology[153,154]. In addition, several monocyte and myeloid-related GWAS (MONO%, MONO#, GRAN%MYELOID, BASO%GRAN, BASO%) were enriched in monocyte-associated topics (k5, k6, k12, k13, k15, k20; **Figure 3.2d**), suggesting that cellular programs in these topics may causally regulate myeloid cell numbers and proportions.

Topic-derived cell trajectories identify cell differentiation and disease-associated continuums

Recent single-cell genomics studies have shown that cell states and disease states often form a continuum rather than distinct clusters; and this continuum can reflect cell differentiation or disease progression[155,156]. We reasoned that if a topic represents a cellular program or state, the loading for this topic should capture the cell trajectory along that program. As a proof-of-concept, we derived a memory B cell trajectory based on topic k2 loadings (Methods). As anticipated, we observed progressive enrichment of non-naïve B cells (includes memory B cells and plasmablasts) along the trajectory (**Figure 3.2e**, top). We correlated gene activity (GA) scores with the memory B cell trajectory and observed decreasing GA scores for naïve B cell marker genes (*IL4R*, *TCL1A*) and increasing GA scores for memory B cells marker genes (*CD27*, *COCH*, *AIM2*) (**Figure 3.2e**, bottom; Supplementary Notes). Therefore, cell loadings from our topic modeling results can effectively capture biologically meaningful continuum along cell states.

The capability to identify memory B cell trajectory directly from topic loadings prompted us to assess if any topic is associated with RA states. In this analysis, we removed cells from COVID-19 convalescent donors to avoid confounding from COVID-19-associated inflammatory signatures. We tested the association between topic loadings and RA cells using logistic regressions and assessed the statistical significance by likelihood-ratio tests (LRT, Methods). We found that k9 and k4 are enriched for RA cells (nominal p-value < 0.05, **Figure S3.3g**). K9 largely represented $CD8^+$ TEM (and $CD4^+$ TEM to a lesser extent) and accessible regions in k9 were significantly enriched for RA heritability (**Figure 3.2b,d**). We observed that cells with high k9 loadings-especially those at the top 20%-are disproportionately RA cells, suggesting an expanded TEM population in RA patients (**Figure 3.2f**, top). To further study the relevance of k9 to RA, we built a k9 trajectory and divided the cells into five equal loading bins (quintiles, Methods). We then tested for differential GA

scores between RA cells in the different quintiles against all cells (RA and healthy) in the first quintile and we found that RA cells in high quintiles have more differentially accessible genesespecially the fifththan RA cells in lower quintiles (**Figure S3.3h**). We identified 352 and 256 genes with significantly ($|\log_2 FC| > 1$, 1% FDR) and progressively increased or decreased accessibility along the k9 trajectory, respectively and confirmed their changes in accessibility by visualizing the GA scores along k9 trajectory (Methods). Among these genes we found well-established RA risk genes including *IL2*, *TNFSF4*, *CCL18*, and *CXCL13* (**Figure 3.2f**)[157,158]. In contrast, we only identified ~200 up-regulated genes when comparing all RA cells against all healthy cells in k9, despite this test being better-powered as it used more cells. This demonstrates that the open chromatin landscape of cells from RA patients form a continuum from almost healthy to diseased state, which is masked by "healthy" RA cells in conventional cluster-based differential analysis.

We highlight several genes that are up-regulated along the k9 trajectory, all of which are understudied and may play a role in RA pathogenesis. For example, we observed increased accessibility around the LILRB1 gene body along k9 trajectory (**Figure 3.2g**); accordingly, *LILRB1* is up-regulated in peripheral $CD8^+$ T cells from seropositive RA patients from an earlier study[52]. We also observed increased accessibility around gene *SIGLEC9* (encodes Sialic acid-binding Ig-like lectin-9, or Siglec-9, **Figure 3.2g**); it has been shown that soluble Siglec-9 is up-regulated in the synovial fluids of RA patients[159]. Another gene PROK2 also showed increased accessibility along k9 trajectory (**Figure 3.2g**). PROK2 encodes prokineticin-2 (PK2), which is elevated in RA synovial fluid and the synovial membrane of collagen-induced arthritis (CIA, an animal model for RA) in mice[160,161]. These findings suggest that k9 represents a rare population of $CD8^+$ T cells in RA peripheral blood that may reflect activated cellular programs in the inflamed synovium, and can be used to study RA pathobiology.

Harmonized scATAC-seq dataset reveals high-resolution shared and cell type specific caQTL

Our harmonized scATAC-seq dataset with genotype information for all 56 individuals allows us to study the impact of genetic variation on chromatin accessibility in multiple cell types and contexts. To map the genetic effects of chromatin accessibility, we first used `RASQUAL` to model both intra-individual allelic-imbalance and inter-individual phenotype variations for SNPs in a 10 Kb window flanking the cRE center in whole blood (WB, aggregating all cells from a donor) as well as the seven immune cell types in L1 annotation. Phenotypic principal components (PCs) were calculated to assess technical and batch effects across our datasets; as anticipated, the top 2 PCs largely separated samples by study and library (**Figure S3.4a**). We chose the number of PCs in the model empirically by maximizing the number of significant caQTL detected at 10% FDR. We also included five genotype PCs, the number of cells within each group, library depths and GC content of each peak as covariates[40]. We identified the most number of caQTL (15,962, 10% FDR) in WB, owing to the largest number of cells used for pseudobulk aggregation. The number of caQTL in each L1 cell type varied greatly and were proportional to the number of cells, ranging from 147 in DC to 9,200 in monocytes (10% FDR; **Figure 3.3a**). In total, 25,107 caQTL (8.90% of all tested cREs) were found, 6,782 of which were only discovered in L1 annotations, but not in WB. Compared to the most recent work on mapping caQTL using scATAC-seq by Benaglio et al.[13], we quadrupled the number of identified caQTL.

We next verified that the large number of caQTL we found are very likely true positives. To do this, we mapped caQTL in WB in the three studies separately and found multiple lines of evidence indicating that the caQTL uniquely identified in our harmonized data ("novel caQTL") are bona fide caQTL. First, novel caQTL showed allelic imbalance in the smaller dataset from Benaglio et al.[13], despite being nonsignificant (**Figure 3.3b**). Second, caQTL we found are strongly enriched in bulk caQTL from lymphoblastoid cell lines

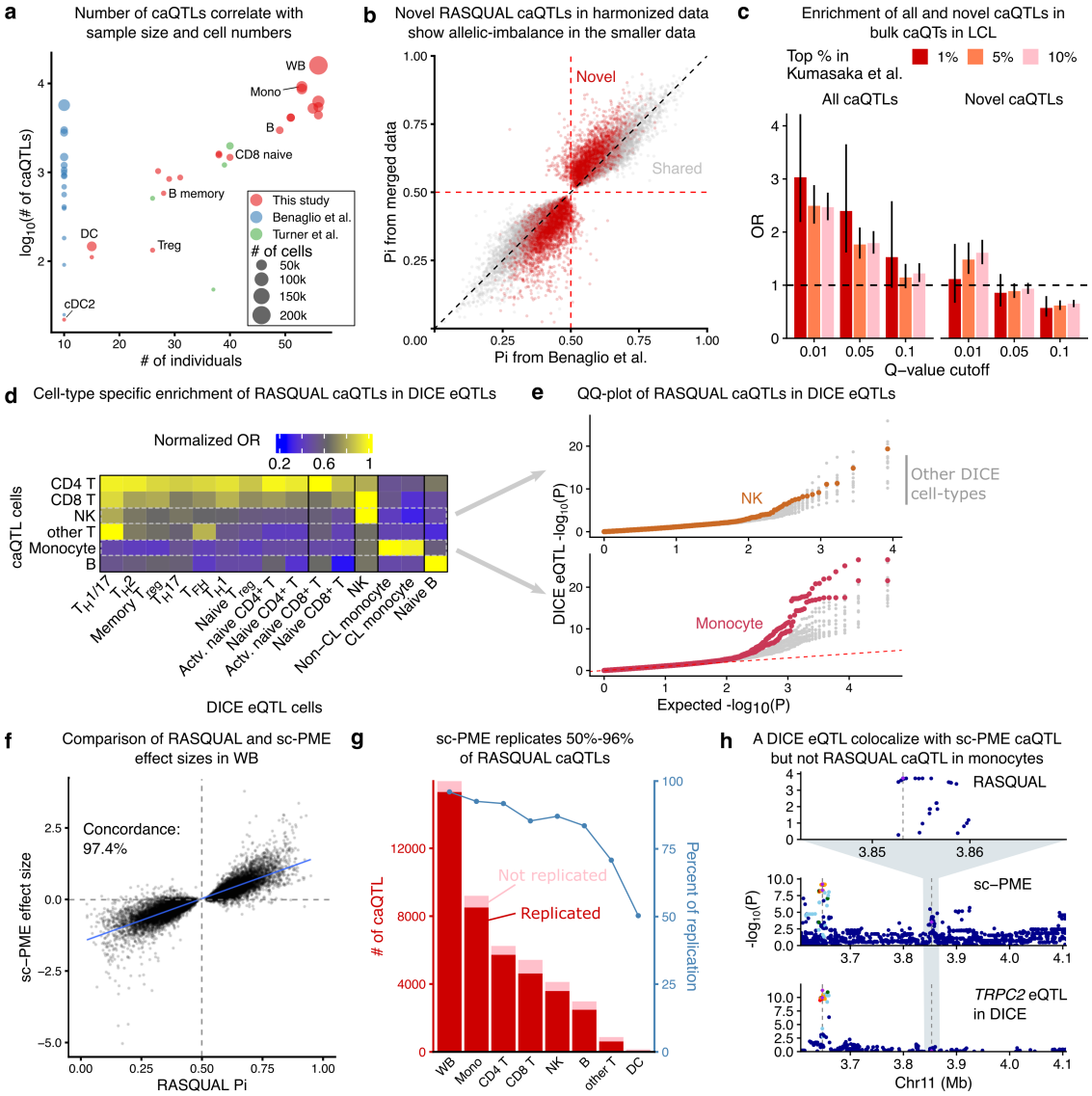

Figure 3.3. Mapping of caQTL with RASQUAL and sc-PME model.
a, The number of RASQUAL caQTL positively correlates with cell number and sample size. **b,** RASQUAL caQTL only found in merged data also show allelic imbalance in smaller data (Benaglio et al.), and have correlated effect sizes. **c,** Enrichment of WB caQTL in bulk caQTL in LCL for all significant ones (left) and those only found in the merged data (right). **d,** Cell type-specific enrichment of caQTL in DICE eQTL. Odd ratios are normalized by the maximum in each row. **e,** QQ-plot of RASQUAL caQTL in DICE eQTL. caQTL in NK cells (top) and monocytes (bottom) show elevated signals only in eQTL from corresponding cell types. All other DICE cell types were in grey. **f,** RASQUAL and sc-PME caQTL effect sizes are highly correlated. **g,**, Replication of RASQUAL caQTL in sc-PME model. Barplot shows the number of RASQUAL caQTL that are replicated or not in sc-PME model; line chart shows the percentage of RASQUAL caQTL replicated. **h,**, A DICE eQTL in monocytes colocalizing with sc-PME caQTL that is different from the RASQUAL lead SNP. Vertical dashed lines highlight the genomic coordinations of lead SNPs in RASQUAL, sc-PME and DICE. The shaded region highlights the mapping window of RASQUAL and its position relative to the mapping window of sc-PME. SNPs are colored by LD to the lead SNP in sc-PME.

(LCL; **Figure 3.3c**)[162]. Moreover, the novel caQTL showed higher enrichment in the top 5% (OR: 1.48, p-value: 1.34e-4, Fisher's exact test) and 10% (OR: 1.61, p-value: 1.43e-10) of LCL caQTL compared to the top 1% (OR: 1.12, p-value: 0.62; **Figure 3.3c**), indicating that novel caQTL in our dataset tend to have smaller effect sizes and could only be detected in our larger, harmonized dataset. Finally, caQTL in L1 cell types showed cell-typespecific enrichment in eQTL from 15 immune cell types in the DICE consortium[29,33]. For instance, caQTL in monocytes and NK cells are the most enriched for eQTL in classical/non-classical monocytes and NK cells in DICE, respectively (**Figure 3.3d,e**). Similarly, caQTL in $CD4^+$ T cells were broadly enriched in eQTL across various T cell subtypes in DICE. When visualizing p-value distribution of caQTL SNPs in DICE eQTL, the same consistent trend emerged (**Figure 3.3d**). In conclusion, our analysis captured high-quality cell-typespecific effects on chromatin accessibility and significantly increased the number of significant caQTL compared to previous studies.

Modeling single-cell count data with Poisson regression improves estimation of caQTL effect size

`RASQUAL` increases detection power by combining between-individual and within-individual (allele-specific) signals but does not produce statistical measurements (i.e. effect size and its standard error) that are required in downstream analysis. Moreover, `RASQUAL` is biased toward heterozygous SNPs with higher read coverage; SNPs outside cREs have weaker or no allele-specific signal and lower accuracy in haplotype phasing, making their effects difficult to model in `RASQUAL`. To overcome these limitations and enhance the utility of our caQTL results, we sought to produce standard summary statistics of caQTL effects, thereby allowing for popular downstream analysis, such as colocalization[21], TWAS, mashr[57] and meta-analysis. To this end, we modeled single-cell count data with a Poisson mixed-effects model (sc-PME) to re-map `RASQUAL` caQTL[39]. We focused on the 25,107

significant caQTL peaks identified by RASQUAL in WB and L1 cell types to reduce computation time and avoid spurious associations for cREs with high read dropout rates (Supplementary Notes). We fitted the sc-PME model in each study separately, and performed meta-analysis to combine the summary statistics from the three studies in WB and each cell type.

We verified that our sc-PME model produced accurate effect sizes and p-value estimates. First, the effect sizes from both methods were highly concordant for lead SNPs (97.4% in WB; **Figure 3.3e**). We then asked whether RASQUAL caQTL could be generally replicated by sc-PME. In WB, 96% of RASQUAL caQTL was replicated; across L1 cell types, an average of 82.1% of RASQUAL caQTL was replicated. As anticipated, the rate of replication was lowest among rare cell types like DC (**Figure 3.3f**). Since sc-PME was fitted in each study separately and then combined using meta-analysis, the high rate of replication of RASQUAL caQTL again demonstrates that our caQTL mapping pipeline is able to control for possible type-1 error due to inter-study variations. Finally, compared to a standard linear mixed-effects model, effect sizes from sc-PME models had higher reproducibility and correlation with RASQUAL caQTL (Supplementary Notes, **Figure S3.4b**).

Another potential benefit of the sc-PME model is that the larger mapping window (500 Kb v.s. 10 Kb in RASQUAL) can capture significant caQTL that are not tested in RASQUAL at all. Indeed, RASQUAL caQTL were highly enriched toward the cRE centers, whereas those only found in sc-PME tend to locate farther away (**Figure S3.4c**). To test this possibility, we focused on 2,224 caQTL whose sc-PME lead SNPs resided in another cRE that lies outside the 10 Kb window used in RASQUAL, as they are more likely to capture the causal mechanism. Of those 2,224 caQTL-containing distal cREs, 434 (19.5%) also had a significant caQTL, showing a strong enrichment of cREs with a caQTL compared to genome-wide average (8.73%, p-value=1.39e-56; hypergeometric test). This suggests that sc-PME captures bona fide distant caQTL that cannot be effectively detected in RASQUAL. We

highlight a cRE (chr11:3856237-3856737) whose lead caQTL from RASQUAL in monocytes lies within itself, but the lead sc-PME caQTL is ~250 Kb upstream and colocalized with an eQTL for gene TRPC2 in classical monocytes from DICE (PP4=0.81, **Figure 3.3h**). These lines of evidence support the notion that sc-PME is better than RASQUAL at capturing putative causal SNPs.

In summary, we identified a set of 25,107 caQTL in WB and separate immune cell types, for which summary statistics can be readily used in standard and popular downstream analyses.

Sharing and specificity of caQTL and eQTL across cell types and states

Previous single-cell eQTL studies in immune cells find that genetic effects on gene expression are largely cell-typespecific[38]. Similarly, caQTL studies using scATAC-seq also reveal a large number of cell-typespecific caQTL[13,138]. The high prevalence of cell-typespecific QTL implies that most genetic variants have cell-typespecific effects that can only be uncovered using single-cell genomic datasets. However, this contradicts observations from previous QTL analyses in immune cells[33,135] and other cell types[137,163], which instead suggests that most QTL effects are shared across cell and tissue types. Another explanation to the sharing of eQTL among bulk tissues is the sharing of similar cell types across tissues, and accordingly tissue-specific eQTL can largely be attributed to tissue-specific cell types[27]. However, this cannot explain the sharing of eQTL between FACS sorted immune cell types and LCLs. We thus sought to systematically evaluate the specificity of caQTL across eight contexts (WB and the seven common immune cell types in L1 annotation), and to understand their impact on gene expression levels.

We first compared caQTL sharing from RASQUAL and sc-PME results. Interestingly, we identified more shared caQTL in sc-PME than in RASQUAL results. Of those 18,746 RASQUAL caQTL identified in L1 cell types, 13,308 (71.0%) were unique to one cell type, and merely

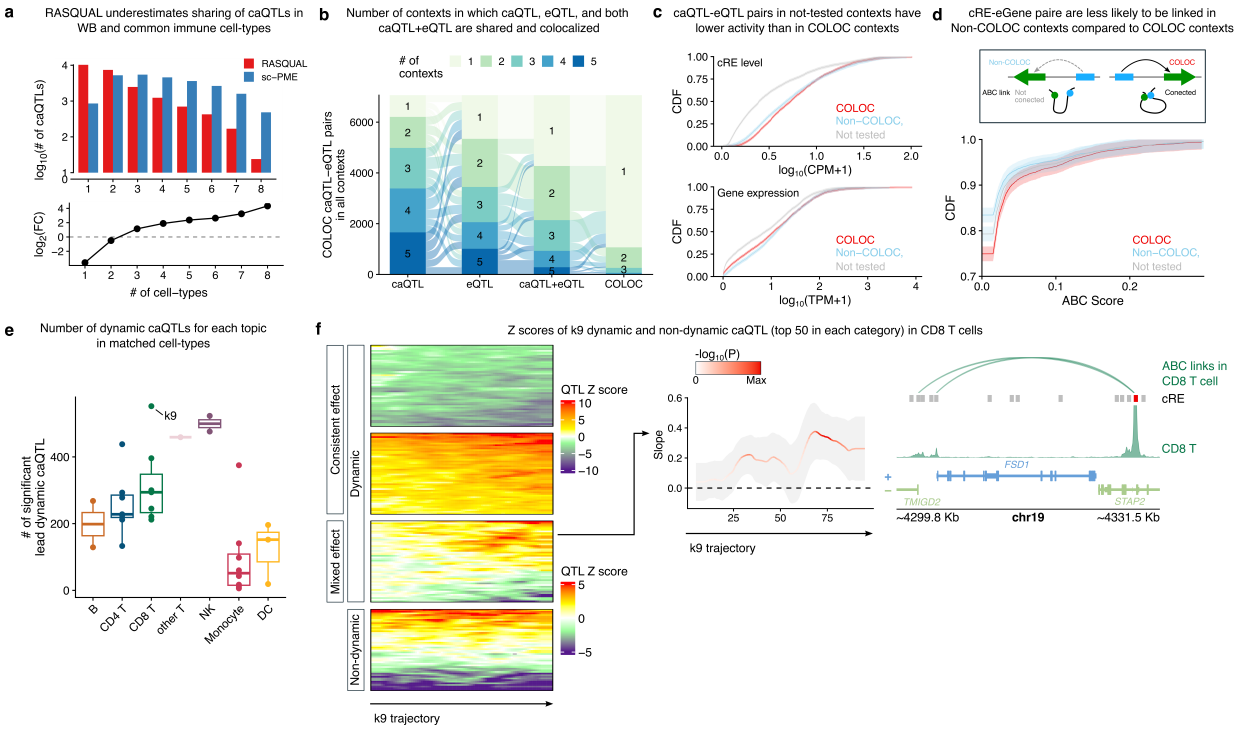

Figure 3.4. caQTL-eQTL colocalization and dynamic caQTL mapping.
a, Sharing of caQTL in seven immune cell types for both RASQUAL and sc-PME results. Top: bar plot of number of caQTL shared in a given number of contexts. Bottom: $\log_2$ fold-change in the number of caQTL between sc-PME and RASQUAL, highlighting the increased level of sharing for sc-PME caQTL. **b,** Barplot for the number of contexts in which an eQTL, caQTL, or both are significant and the number of contexts in which they colocalize. **c,** Comparison of cRE levels (top) and gene expression levels (bottom) for COLOC caQTL-eQTL pairs in COLOC contexts, non-COLOC contexts and contexts where colocalization was not tested (No tested). **d,** Cumulative distribution of ABC scores between all the COLOC caQTL-eQTL pairs in COLOC contexts, non-COLOC contexts and contexts where colocalization was not tested (No tested). ABC scores of caQTL-eQTL pairs without observed links in Nasser et al. were set to 0. **e,** The number of significant dynamic caQTL in each cell type. Each point represents a pair of cell type and topic with which the interaction between loadings and genotypes were tested. **f,** Left: Z scores of dynamic and non-dynamic caQTL from $CD8^+$ T cells in rolling windows along k9 trajectory. Dynamic caQTL were further categorized by whether their effects are consistently significant along k9 trajectory or only in part of the trajectory. The most significant 50 caQTL from each category were plotted. Right: effect sizes of one dynamic caQTL, where the underlying cRE (chr19:4328325-4328825) is linked to an immune-related gene *TMIGD2* (also known as *CD28H*). Shaded region represents standard error of effect sizes.

24 caQTL were shared in all seven (**Figure 3.4a**). Although these findings are consistent with analyses from previous work17, we found that the estimated levels of caQTL sharing were much higher in sc-PME results. In fact, only 5,324 (22.2%) caQTL are found to be unique in one cell type, and an average of 4,132 caQTL were identified to act in two to five cell types (**Figure 3.4a**). Furthermore, sc-PME identified 878 caQTL that are only significant in L1 cell types but not in WB, whereas `RASQUAL` identified 9,145 caQTL that are specific to L1 cell types (**Figure S3.5a,b**). These results indicate that the sc-PME model is better powered than RASUAL and the prevalence of cell-typespecific QTL in single-cell genomics data could be explained by low QTL detection power rather than cell type specificity.

Many caQTL should regulate promoters or enhancers with an effect on the expression level of a nearby gene; therefore, we asked whether caQTL are also eQTL in the DICE dataset47 in caQTL-eQTL colocalization analysis [21]. We matched 15 cell types/subtypes to five cell types in our scATAC data (excluding DC and other T cells in L1 cell types) and these are referred to as "contexts" hereafter (**Figure S3.5c**). In total, we identified 7,063 unique caQTL-eQTL pairs that colocalized in at least one matching context (referred to as COLOC caQTL-eQTL pairs hereafter), including 3,169 eGenes (24.5% of all tested) and 4,409 cREs (21.7% of all tested). In each context, 17.6-29.8% (average: 22.9%) of test eGenes and 8.43-15.3% (average: 12.1%) of tested cREs colocalized, respectively. Moreover, caQTL-eQTL colocalizations are highly context-specific, with 5,995 (84.5%) being colocalized in only one context. In stark contrast, sharing of caQTL-eQTL pairs across all contexts were much higher (**Figure 3.4b**). Therefore, a caQTL-eQTL pair might colocalize in one context but not another simply because the eQTL is absent.

To better understand the identified caQTL-eQTL colocalization-or the lack thereof-we focused on the 7,063 COLOC caQTL-eQTL pairs and, for each pair, grouped all the contexts into three scenarios: (1) contexts where a caQTL-eQTL pair colocalizes (COLOC

contexts); (2) contexts where the caQTL-eQTL pair was tested for colocalization but does not colocalize (non-COLOC contexts); (3) contexts where the caQTL-eQTL pair was tested for colocalization due to the lack of significant caQTL/eQTL (not-tested contexts). First, we hypothesized that the lack of colocalization could be simply attributed to the lack of activity, namely, low enhancer or gene expression levels. We thus compared cRE levels and gene expression levels across the three types of contexts above. We found that in not-tested contexts, cREs/eGenes have significantly lower activity than in COLOC contexts, suggesting that the lack of caQTL/eQTL due to low activity can explain the lack of caQTL-eQTL colocalization in these contexts (**Figure 3.4c**). Similar trends were also observed for eQTL and caQTL effect sizes, further supporting the idea that low activity genes/cREs require higher power to call significant QTL (Supplementary Notes). Notably, we found cases where the cRE activity is shared across multiple contexts but the eQTL is context-specific. For example, gene PADI2 is highly and specifically expressed in classical monocytes and has a classical monocyte-specific eQTL, which colocalized with caQTL chr1:17336114-17336614 (**Figure S3.5d-g**). Although this caQTL is shared in $CD8^+$ T cells and B cells, it was not tested for colocalization with PADI2 because the eQTL is absent in these cell types due to low expression (**Figure S3.5e**). This indicates that cREs can share activity and genetic control in many contexts, but they are not necessarily functionally regulating the target gene in all these contexts. Although a larger sample size would increase the statistical power to identify significant eQTL for lowly expressed genes, our analysis underpin the biological reasons for the lack of eQTL sharing. Moreover, we observed no difference in cRE and gene expression levels between COLOC contexts and non-COLOC contexts, suggesting that other mechanisms exist. Since enhancer-promoter looping is often necessary for an active enhancer to impact gene expression, we hypothesized that in non-COLOC contexts, a cRE may not be physically contacting a TSS, even though the caQTL and eQTL are significant. We compared predicted enhancer-TSS links

from the Activity-by-Contact (ABC) model for our caQTL-eQTL pairs in COLOC, non-COLOC and no-test contexts[164]. We found that caQTL-eQTL pairs are more likely to be linked and have significantly higher ABC scores in COLOC contexts than in non-COLOC contexts (**Figure 3.4d**). This implies that even when an eQTL and a caQTL colocalizes in one context, they may not colocalize in another context because the cRE is not physically interacting with the same TSS. This offers a mechanistic explanation for context-specific caQTL-eQTL COLOC and highlights the importance of distinguishing "merely active" cREs from "functional" cREs in caQTL analysis.

Dynamic effects of caQTL along topic model-derived trajectories

Shared QTL often have quantitatively different effects across contexts. Similarly, although cell-typespecific QTL are typically defined by a hard cutoff, their true effect sizes are continuous rather than "all-or-none" across contexts. Measured along a certain trajectory, such as differentiation or time after stimulation, many QTL effect sizes appear to be tracking the trajectory; these QTL are termed "dynamic QTL'[165,166]. Single-cell data offers a unique opportunity to understand the dynamic effects of QTL, because each single cell is a measurement that can be placed along a latent trajectory that can be statistically defined. To identify caQTL that have dynamic effects, we tested for the linear interaction between lead caQTL in seven cell types and the loadings of relevant topics and identified 4,577 (19.1%) unique cREs that have at least one dynamic caQTL in at least one cell-typetopic pair (q-value < 0.01). On average, we detected 233 significant dynamic caQTL in each cell-typetopic pair, suggesting that dynamic effects are often cell-type and topic-specific. As expected, we found that the number of cells in each topic greatly influenced the statistical power for calling dynamic caQTL; widespread topics represented in multiple cell types including k11 had the most dynamic caQTL, whereas topics only represented in a small number of cells, like k20 (active in a subset of monocytes) had the smallest number of

dynamic caQTL (**Figure 3.4f**).

To get a deeper understanding of our dynamic caQTL, we focused on k9-interacting caQTL in $CD8^+$ T cells, which represent the largest group of dynamic caQTL in $CD8^+$ T cells (**Figure 3.4f**). To visualize the changes in effect sizes of these caQTL along k9 trajectory, we fitted the sc-PME model without an interaction term in k9 rolling windows. Examining the caQTL Z scores in each rolling window, we manually divided dynamic caQTL into two groups (consistently significant and partially significant) and contrasted their effects with randomly chosen non-dynamic caQTL (likelihood-ratio test p-value > 0.5) (**Figure 3.4g**, left). Consistently significant dynamic caQTL are significant along the k9 trajectory, with varying effect sizes. Partially significant caQTL tend to have negligible effects in part of the trajectory and progressively gained (or lost) genetic effects along the trajectory. Thus, partially significant dynamic caQTL are more likely to be masked when all cells are analyzed as a group (**Figure 3.4g**, right). To understand the functional relevance of k9-interacting dynamic caQTL, we mapped these cREs to potential target genes through the ABC model and tested for gene sets enrichment. Many genes were enriched in immune and disease-related pathways, including the immune response-regulating signaling pathway (GO:0002764, p-value = 3.01e-4; *BLK*, *CD8A*, *CD8B*, *THEMIS2*, *ICOSLG*, *NOD2*, and *KLHL6*) and RA development (Wikipathways: WP5033, p-value = 5.20×10^{-3}; *BLK*, *CCR6*, and *CIITA*) (**Figure S3.5h**). Several target genes of these dynamic caQTL were also implicated in the k9-associated RA genes we identified in topic analysis, including *TMIGD2*, *SPRY1*, *CD8A*, *CD8B*, *SOCS3*, and *LIF*. Gene TMIGD2 was down-regulated in RA cells along k9 trajectory and was regulated by an upstream cRE chr19:4328325-4328825 that had a dynamic caQTL (**Figure 3.4g**). TMIGD2 (also known as *CD28H* or *IGPR-1*) encodes a recently identified T cell costimulatory receptor and can promote cell proliferation and cytokine secretion upon engagement with its ligand B7H7[167]. Although the role of *TMIGD2* in autoimmunity is elusive, it remains a promising candidate for

future research and a drug target[168]. Another gene SPRY1 was under the regulation of cRE chr4:124521047-124521547, which lies ~203 Kb downstream of SPRY1, and may impact RA etiology in fibroblast-like synoviocytes[169]. Finally, we colocalized sc-PME caQTL with a trans-ethnic RA GWAS1 (see below) and found that 24 out 69 (34.8%) colocalized GWAS loci have an underlying dynamic caQTL. Moreover, we found that colocalized caQTL are also enriched for dynamic caQTL (OR=2.27, p-value=1.70×10^{-5}; Fisher's exact test). These results demonstrate that dynamic caQTL may have important regulatory functions in immune and disease-related pathways. Taken together, our analyses suggest that mapping dynamic caQTL is a crucial step for understanding how genetic variation may impact disease-relevant genes.

Identification and interpretation of autoimmune GWAS loci on chromatin accessibility across cell types and states

To understand the relationship between genetically controlled cRE and immune-related complex traits, we colocalized sc-PME caQTL with GWAS of 11 immune-related diseases and 36 blood phenotypes, and compared the results with eQTL-GWAS colocalization[33]. We first calculated the proportion of colocalized GWAS loci in a context-agnostic fashion. In total, 60.9% (4,981 out of 8,176) GWAS loci (GWAS-SNP pairs) colocalized with either a caQTL or an eQTL, of which 1,743 colocalized with both a caQTL and an eQTL (caQTL+eQTL), 2,337 colocalized with caQTL only (caQTL-only), and 901 with eQTL only (eQTL-only). For each GWAS, caQTL explained 6.0% to 33.3% more loci than eQTL (**Figure 3.5a**), increasing the percentage of colocalized loci to an average of 50.7% from 32.6% for eQTL alone, consistent with previous observations that caQTL could offer novel biological insights on GWAS mechanisms not captured by eQTL[133,170].

To compare which cell type contexts were implicated at the caQTL+eQTL loci, we mapped 15 DICE cell types to the five common immune cell type contexts as we did above

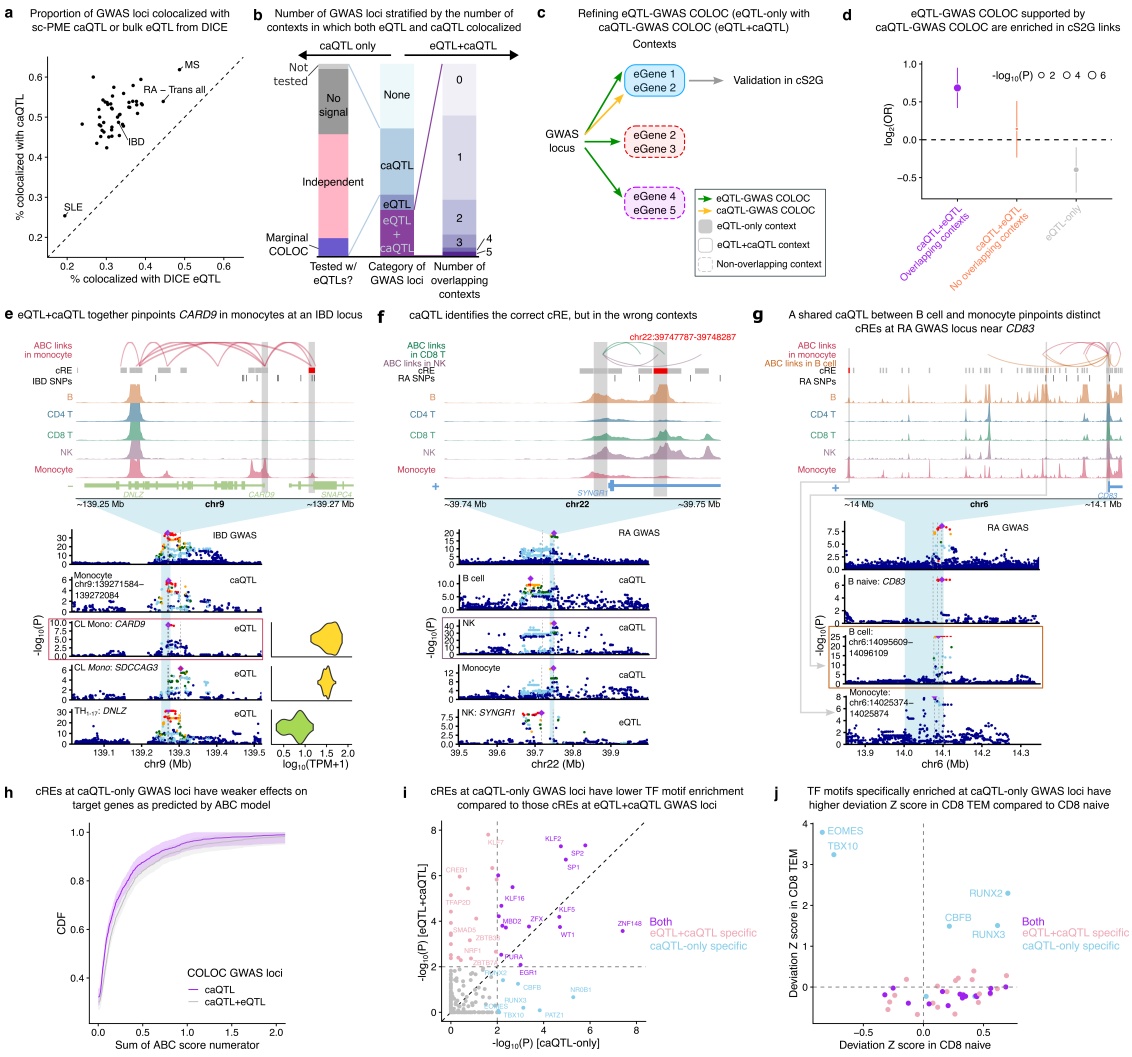

Figure 3.5. caQTL and GWAS analysis.
a, Comparison of the percentage of colocalized GWAS loci with caQTL and DICE eQTL. **b,** Number of contexts in which GWAS loci colocalize with either eQTL, caQTL or both. **c,** Schematic showing that restricting eQTL-GWAS COLOC by caQTL-GWAS COLOC in the same context to narrow down potential causal genes and contexts. **d,** Restricting eQTL-GWAS pairs to contexts also supported by caQTL-GWAS COLOC increases enrichment in causal S2G (cS2G) links. Error bars represent 95% confidence intervals of $\log_2$(OR) estimates. **e,** An example showing that eQTL- and caQTL-GWAS COLOC together narrows down *CARD9* gene in monocytes as the causal gene and context for an IBD locus. **f,** Similar to **e**; an example showing a caQTL shared across many cell types that colocalized with RA GWAS near gene *SYNGR1*. However, the *SYNGR1* eQTL only colocalized in NK cells. **g,** Similar to **e**; an example showing a caQTL shared by two cREs in B cells and monocytes, respectively, that colocalized with RA GWAS near gene *CD83*. **h,** Cumulative distribution of the sum of ABC score numerator for all TSS connected to a cRE, colored by whether the cRE colocalize at GWAS locus also colocalize with eQTL. **i,** Comparison of TF binding site enrichment in cREs underlying caQTL-only or caQTL+eQTL GWAS loci in CD8$^+$ T cells. **j,** Comparison of TF motif deviation Z scores in CD8$^+$ naïve CD8$^+$ TEM for TFs enriched at colocalized cREs in CD8$^+$ T cells.

for caQTL-eQTL colocalization analysis (**Figure S3.5a**). Of these 1,743 caQTL+eQTL loci, 1,252 had colocalization with both eQTL and caQTL in at least one overlapping context; but this only constituted 25.1% of all colocalized GWAS loci. The remaining 491 (27.4%) caQTL+eQTL loci colocalized with caQTL and eQTL in distinct contexts, raising the question of which contexts are likely causal. Notably, GWAS loci colocalized in more contexts tend to colocalize with multiple cREs and eGenes, suggesting widespread pleiotropy and higher uncertainty in determining a single causal gene and context (**Figure S3.6a,b**).

By restricting our investigation of GWAS loci to caQTL+eQTL contexts, we can reduce the number of identified eGenes and contexts in eQTL-GWAS colocalization results (**Figure 3.5c**). We hypothesized that this shortlist of colocalized eGenes might be enriched in causal genes. To test this, we separated all eQTL-GWAS colocalization (V2G) pairs to three groups depending on whether they are supported by caQTL-GWAS colocalization: (1) have caQTL-GWAS colocalization in at least one overlapping context (caQTL+eQTL, overlapping contexts); (2) have caQTL-GWAS colocalization but not in overlapping contexts (caQTL+eQTL, no overlapping contexts); (3) no caQTL-GWAS colocalization in any context (eQTL-only) (**Figure 3.5d**). We validated these V2G pairs with the combined SNP-to-Gene (cS2G) resource, which has been shown to identify causal SNP-gene links with high confidence[171]. We found that caQTL+eQTL V2G pairs in overlapping contexts are significantly enriched for cS2G links ($\log_2$OR = 0.68, p-value = 2.68e-07; Fisher's exact test), whereas caQTL+eQTL V2G pairs in non-overlapping contexts show no significant enrichment ($\log_2$OR = 0.14, p-value = 0.433). In contrast, eQTL-only V2G pairs were moderately depleted with cS2G links ($\log_2$OR = -0.40, p-value = 7.43e-3), suggesting that more evidence is needed to nominate causal genes from this set (**Figure 3.5d**). We also stratified caQTL+eQTL colocalization by the number of contexts in which they both colocalize and found that the number of overlapping contexts does not have an impact on the enrichment in cS2G (Supplementary Notes). This demonstrates the

importance of caQTL-GWAS COLOC in the interpretation of eQTL-GWAS COLOC results.

Using this approach, 54.2% of caQTL+eQTL GWAS loci can be narrowed down to no more than two eGenes in less than two contexts. Many of these candidate eGenes and contexts have known roles in disease etiology. For instance, the inflammatory bowel disease (IBD) risk locus 9:139269198 colocalized with eQTL of CARD9 and SDCCAG3 in monocytes, and that of DNLZ in TH1-17 cells. The same locus also colocalized with two cREs in monocytes and NK cells. By overlapping caQTL- and eQTL-GWAS colocalization, we identified CARD9 and monocytes as the putative causal gene and context for this GWAS locus. In support of this, we found that: (1) CARD9 is the only gene linked to the GWAS SNP 9:139269198 in cS2G data (cS2G score=1); (2) the colocalized cRE is predicted to be connected to CARD9 TSS in monocytes in the ABC model, whereas none enhancer-TSS links were found in T cells; (3) CARD9 is highly expressed in classical and non-classical monocytes, whereas DNLZ is lowly expressed in TH1-17 cells; (4) the identified cRE (chr9:139271584-139272084, ~13.3Kb upstream of CARD9 TSS) is monocyte-specific and harbors fine-mapped IBD GWAS SNPs (**Figure 3.5e**). Indeed, CARD9 is one of the established causal genes in IBD[172–174].

While it is promising to combine eQTL- and caQTL-GWAS colocalization to nominate causal genes, we next turned our focus to the caQTL-only GWAS loci, which consisted of 57.3% of all caQTL-GWAS colocalization. In addition, 720 caQTL+eQTL GWAS loci also colocalized with caQTL in some contexts where there is no eQTL-GWAS colocalization. Previous studies suggested that these caQTL without eQTL effects represent poised enhancers[170]. Similarly, we propose that in caQTL-only contexts, we can identify the correct cRE and its genetic effects, but the underlying contexts are not causal ones.

To better illustrate this idea, we highlight two example GWAS loci where the putative causal context can be anchored by eQTL data and known biological knowledge. For example, we found a cRE chr22:39747787-39748287 with a shared caQTL that colocal-

ized with an RA GWAS locus 22:39747780 in five contexts (B cell, $CD4^+$ and $CD8^+$ T cell, NK and monocyte). However, the same RA locus only colocalized with SYNGR1 eQTL in two contexts ($CD8^+$ T cell and NK). This cRE lies within the first intron of SYNGR1 gene and is linked to its promoter in the ABC model, likely regulating SYNGR1 expression (**Figure 3.5f**). SYNGR1 is not expressed in the three contexts where it does not have an eQTL, although the promoter region is accessible (**Figure S3.6c,d**). This suggests that a caQTL can be active in some contexts but does not have functional effect on gene expression. Therefore, although caQTL-GWAS colocalization reveals that this RA GWAS locus has genetic effects on the nearby enhancers in all contexts, only two of the contexts (i.e. $CD8^+$ T cell and NK cell) are likely causal. Hypothetically, this would lead to the wrong conclusion if a study did not include $CD8^+$ T cells or NK cells.

In more complicated cases, the same GWAS locus can colocalize with distinct cREs in distinct contexts. For instance, an RA GWAS locus (6:14107197) colocalized with two cREs sharing the same caQTL SNP in B cells (chr6:14095609-14096109) and monocytes (chr6:14025374-14025874); and it also colocalized with a *CD83* eQTL in B cells (**Figure 3.5g**). In the ABC model, the cRE in B cells (chr6:14095609-14096109) targets the *CD83* promoter, both of which overlap with fine-mapped RA GWAS SNPs. CD83 is crucial for B cell maturation and antigen presentation[175]. A previous scATAC study on germinal-center B cells has also reported the relationship between this B cell cRE (chr6:14095609-14096109) and *CD83* promoter at this RA risk locus[176]. On the contrary, the colocalizing cRE in monocytes (chr6:14025374-14025874) lies farther upstream and is not linked to *CD83* promoter (**Figure 3.5g**). We concluded that the causal gene and context for this GWAS locus is *CD83* in B cells. In monocytes, the caQTL-GWAS colocalization indicates the effect of this GWAS locus is mediated by chromatin accessibility, but failed to identify the correct regulatory element. Therefore, in cases like this, caQTL-GWAS colocalization not only leads to the likely wrong contexts, but also the likely wrong cRE.

Finally, to further understand the regulatory roles of caQTL-only and caQTL+eQTL GWAS loci, we studied their "regulatory potential" as the sum of ABC score numerators of all connected TSS. We found that caQTL-only GWAS loci are less connected to TSS compared with caQTL+eQTL loci (two sided KS-test p-value: 1e-16, **Figure 3.5h**); suggesting that caQTL-only GWAS loci could be regulating fewer genes or having weaker effects on their target genes, which can explain why these loci are not detectable as eQTL and is consistent with our observation in the locus around *CD83* gene (**Figure 3.5g**). We next compared the enrichment of TF motifs in colocalized cREs at caQTL-only versus caQTL+eQTL GWAS loci. We found that cREs at caQTL+eQTL loci have a greater number of enriched TF motifs and stronger enrichment p-values than caQTL-only cREs in all contexts except for B cells, which has the fewest number of colocalized cREs (**Figure 3.5i**, **Figure S3.6e**). This indicates that cREs at caQTL-only GWAS loci may not have clear regulatory functions in the contexts included in our data. Nevertheless, some TF motifs were specifically and strongly enriched in cREs at caQTL-only GWAS loci, including EOMES, RUNX2 and RUNX3 in $CD8^+$ T cells. EOMES is typically associated with T cell activation and exhaustion, while RUNX factors and their companion CBFB are associated with $CD8^+$ T effector/memory (TEM) populations. But DICE only included CD3/CD28-activated $CD8^+$ T cells that resemble short-term, acute activation more than exhaustion or long-term memory; this may explain the lack of eQTL-GWAS colocalization at these loci. Indeed, when comparing the TF motif deviation Z scores of all the enriched TFs between $CD8^+$ naïve and $CD8^+$ TEM cells in our scATAC data, we found that TF motifs that were specifically enriched at caQTL-only GWAS loci have much higher deviation Z scores than caQTL+eQTL loci in $CD8^+$ TEM, including for *EOMES*, *RUNX2* and *RUNX3* (**Figure 3.5j**). This observation is consistent with the notion that for some caQTL-only GWAS loci, the caQTL captures the genetic effects of these GWAS variants, but in a non-causal context. These genetic variants are likely to impact disease risk in a context that

has yet to be studied.

To conclude, we colocalized nearly 50% more GWAS loci with our caQTL compared to eQTL alone, and discovered that combining caQTL-GWAS with eQTL-GWAS colocalization enabled us to better nominate potentially causal genes and contexts at roughly 50% of colocalized GWAS loci. Furthermore, we showed that GWAS loci only colocalized with caQTL are likely to pinpoint the correct genetic effects on candidate cREs, but in the wrong context. We concluded that our caQTL-only GWAS colocalization results may have causal function in as yet unprofiled contexts.

3.4 Discussion

In this study, we constructed a large, harmonized map of single-cell chromatin accessibility accompanied with high-quality genotype information. We demonstrated that it is feasible to use a multiplexed library preparation regime and confidently assign cells to each donor using common genetic variants represented in scATAC-seq reads. While this allowed us to profile a large number of individuals with fewer libraries, it may not capture all the rare cell types from each individual. Yet we believe that future larger-scale scATAC studies could benefit from this multiplexed experiment design to increase power and alleviate batch effects.

Through our topic analysis directly on the cell-level count data, we showed that we can capture disease effects only present in a small proportion of RA cells, which are usually masked in cell clusters. One limitation of topic analysis is that it does not allow for explicit removal of potential batch effects from scATAC count data. We found that topic analysis is much less affected by batch effects compared to LSI dimension reduction and clustering, and batch-associated topics can always be removed in downstream analysis. Nevertheless, more principled ways to remove batch effects either before or after topic analysis could prove useful for future research.

Although sc-eQTL studies have become common, sc-caQTL studies are still rare. By applying both allelic-imbalance modeling in RASQUAL and the sc-PME model, we were able to identify 25,107 caQTL. We not only quadrupled the number of caQTL from a previous study17, but we also significantly improved the summary statistics to allow for downstream analysis, allowing for better utilization and interpretation of sc-caQTL data. We also utilized the continuous nature of cell loadings in topic analysis to map dynamic caQTL. We argue that topics are more straightforward to interpret compared to principal components[39,177].

One of the overarching goals in the post-GWAS era is to understand the biological mechanisms of associated variants in the noncoding genome. Here we sought to gain novel insight into this question by utilizing both caQTL-eQTL and caQTL-GWAS colocalization. We reported that adding sc-caQTL increases colocalized GWAS loci by 50% compared to eQTL alone. While this finding is promising and reflects discoveries in bulk caQTL studies[133,170], there has been no straightforward explanation as to why, in a given context, many GWAS loci only have effects on chromatin, but not gene expression. Here we offer a more cautionary tale in the interpretation of these caQTL-GWAS colocalization results. We showed that combining caQTL- and eQTL-GWAS colocalization can help us better nominate putative causal genes and contexts; this suggests that caQTL-GWAS colocalization not corroborated by eQTL are not necessarily functional in the given colocalizing context. In some simple cases, we found convincing examples where caQTL are shared across contexts, but eQTL are cell type-specific due to cell type-specific gene expression. This raises the possibility that many caQTL-GWAS colocalizations only reveal the correct genetic effects on the cRE, but not the causal context. This notion is intimately related to the concept of poised or primed enhancers, which play crucial roles in immune activation and development[53]. In general, we propose that even when the idea of poised enhancer does not readily apply, such as between B cells and monocytes, open

chromatin regions tend to be shared across those contexts. Indeed, a recent multi-modal study shows that open chromatin states are shared among various RA-associated cell-states, whereas the gene expression profiles are distinct[178]. We believe that this layer of complexity in cRE functions is worth experimental validation. Taken together, we demonstrate the utility and importance of chromatin accessibility studies in the functional study of regulatory elements, but we argue that caQTL data-especially those without clear effects on gene expression-defy over-simplified explanations. Further experimental validation in diverse disease-relevant contexts is necessary to understand this layer of complexity in cRE functions.

3.5 Methods

Sample collection and library preparation.

Clinical patient samples collection. The Institutional Ethics Review Board of the Fourth Military Medical University approved this study. We collected RA and healthy samples from Xijing Hospital, Fourth Military Medical University. We obtained written informed consent from all donors. All of the RA patients fulfilled both the 1987 revised American College of Rheumatology criteria and the ACR 2010 Rheumatoid Arthritis classification criteria[112]. We isolate peripheral blood mononuclear cells from blood through gradient centrifugation using lymphocyte separation medium and either processed directly or cryopreserved using CryoStorő CS10 and stored in liquid nitrogen.

scATAC-seq library preparation. We constructed scATAC-seq libraries using the Chromium Next GEM single-cell ATAC reagent kit v1.1 according to the manufacturer's instructions. Before nuclei isolation, the cryopreserved cells were thawed at 37řC for 5 min and viable cells were enriched by Dead Cell Removal Kit (MACS) following the manu-

facturer's instructions. To remove the ambient DNA released by fragile cells, DNase I was used according to the 10x Genomics Protocol: Nuclei Isolation for Single Cell ATAC Sequencing. For fresh samples, PBMCs were processed immediately. To construct a scATAC-seq library with pooled samples, the same amount of cells from two to four distinct samples were combined in a 2 mL microcentrifuge tube to extract nucle, following the 10x Genomics Protocol: Nuclei Isolation for Single Cell ATAC Sequencing. Single-cell ATAC-seq libraries were generated following the Chromium Single Cell ATAC Reagent Kits v1.1 User Guide (10x Genomics; CG000209 Rev G). Library concentrations were measured by a Qubit 2.0 fluorometer. The quality of the final library was analyzed on an Agilent Bioanalyzer. Libraries were sequenced on a NovaSeq 6000 Illumina sequencer.

Low-pass whole genome sequencing. For WGS, one million PBMCs per sample were collected. Samples were sent to Berry Genomics (Beijing, P. R. China) for library construction and sequencing, targeting 4X coverage for all samples.

Preprocessing of scATAC-seq data.

We processed data in all three studies from FASTQ files using the following pipeline. We analyzed each library separately to identify high-quality barcodes and remove doublets. Reads were processed using cellranger-atac v2.1.0 with an in-house GRCh37 reference genome generated using scripts from 10X Genomic documentations. We removed reads that were unmapped, did not have primary alignment, failed platform/vendor quality checks, and had duplicated or supplementary alignment; we only kept reads that were paired and mapped in proper pairs (`samtool view -f 3 -F 3844`). We then removed allelic-biased reads using the WASP workflow implemented in Hornet, using a list of imputed bi-allelic SNPs (see below). We converted each resulting BAM file from each library into a fragment file using sinto v0.7.5 and loaded into an `ArchR` project separately. As a first pass,

we excluded cell barcodes with fewer than 1,000 and more than 50,000 unique fragments, with a TSS enrichment score lower than six for all libraries, and those with high ratios of reads mapping to nucleosomes, mitochondrial genome or ENCODE blacklist regions in a library-specific manner.

Genotype imputation from low-pass WGS and aggregated scATAC-seq data.

FASTQ reads from low-pass WGS were aligned to the human reference genome GRCh37 primary assembly from GENCODE v19. Duplicated reads were removed using `MarkDuplicates` in the `picard` tool. Read coverage across the genome was visualized using `plotCoverage` from `deeptools`, excluding blacklist regions from ENCODE. Genotype likelihood calculation and imputation were performed following GLIMPSE documentation[141]. Basically, we first inferred genotype likelihoods across all SNPs from all individuals in the 1000 Genome Project using `bcftools mpileup`. In this step, we only included sites with sequencing depth below 15 to avoid regions with unreasonably high read coverage. For non-multiplexed scATAC-seq libraries, we calculated genotype likelihood from `cellranger-atac` generated BAM files after removing reads that are unmapped, not primary alignment, failed platform/vendor quality checks, duplicated or supplementary alignment and keeping reads that are paired and mapped in proper pair (`samtool view -f 3 -F 3844`). Genotype likelihood was then estimated using `bcftools mpileup` with the same parameters for WGS data above. Next, we merged genotype likelihood from all individuals from the three studies, and performed GLIMPSE genotype imputation jointly. Imputed genotypes were phased with eagle v2.4.1[179].

To confirm GLIMPSE-imputed genotypes from scATAC-seq reads are of high quality and are not biased by strong allele-specific signals in accessible chromatin regions, we compared Minimac4 imputation from microarray with GLIMPSE results in the 13 individuals with microarray data from Benaglio et al.[13] We first imputed microarray genotyped

SNPs using the pipeline documented in Michigan Imputation Server (`https://imputationserver.readthedocs.io/en/latest/pipeline`). We used the same reference panel as in our GLIMPSE pipeline. We then calculated mean imputation quality score (INFO score) for SNPs stratified by reference MAF bins. We also calculated correlation between genotype dosages for SNPs imputed using GLIMPSE and Minimac4 and derived mean correlation across the 13 individuals in each reference MAF bin using `vcf-stats`.

Demultiplexing of scATAC-seq libraries.

We used the imputed genotype from GLIMPSE to demultiplex scATAC-seq libraries collected in our cohort. We genotyped each cell in a scATAC-seq library at SNPs with MAF > 0.05 in 1000G Project v3 using `cellsnp-lite` v1.2.2. In this step we included cell barcodes with at least 500 unique fragments. Note that this cutoff is more lenient than the 1,000 unique fragments we used in the preprocessing step above. This is mainly to remove the majority of barcodes with very low unique fragments to speed up the `cellsnp-lite` run. These SNPs were then used to subset the high-quality, non-monomorphic imputed SNPs (`RAF>0.05 & RAF<0.95 & AF>0 & AF<1 & INFO>0.7`) from low-pass WGS, which formed the reference data for demultiplexing. Finally, we ran Vireo v0.5.7 to assign an individual to each cell barcode [145]. We only kept cell barcodes that can be confidently assigned to an individual ID in downstream analysis.

Preprocessing and cell type annotation of scRNA-seq data.

We re-analyzed previously published PBMC scRNA-seq data from three RA patients, one ankylosing spondylitis (AS) control and one healthy control. Each scRNA-seq library was first processed separately. Gene count matrix was used to create a Seurat object, keeping genes expressed in at least five cells and cells with at least 20 expressed genes (`min.cells=5, min.features=20`). Percentage of mitochondrial reads was calculated with

`PercentageFeatureSet` function. We further kept cells with at least 500 mRNA, percent_MT below 10% and removed the top 2% cells with the most number of read counts. These cells were analyzed using the Seurat pipeline. Briefly, SCTransform was performed using percent_MT, total RNA counts and number of features as variables to regress out. We then performed PCA with 30 PCs, found neighbors with default parameters and identified cell clusters with a resolution of 1.5. After these preprocessing steps, we used the `paramSweep_v3` function, `summarizeSweep` function and `find.pK` functions to find suitable parameters for `doubletFinder`. We used `modelHomotypic` functions with Seurat clusters to calculate homotypic proportions and used `doubletFinder_v3` with parameters estimated above to remove potential doublets.

To annotate cell types with Azimuth, the five scRNA-seq libraries were integrated into one Seurat object by first identifying top 3,000 features with `SelectIntegrationFeatures` function, and applying `PrepSCTIntegration`, `FindIntegrationAnchors` and `IntegrateData` functions. We then ran `Azimuth` on the integrated dataset with its built-in PBMC reference data. Cells were annotated on two granularities: L1 consisted of eight common immune cell types: B, $CD4^+$ T, $CD8^+$ T, dendritic cells (DC), monocytes, natural killer cells (NK), other T cells and other ambiguous cell types; L2 consisted of 17 cell types and subtypes. We required cells to have consistent annotations in L1 and L2. For example, cells annotated as NK need to belong to one of NK, NK Proliferating or NK_CD56bright in L2 annotation.

Basic analysis of scATAC-seq data.

After keeping cell barcodes with a valid individual ID from Vireo, we also used AMULET to flag and remove potential doublets (AMULET q-value < 0.1). This is because Vireo only identifies doublets that contain multiple cells from distinct individuals, but not for the same individual.

Through the processing steps above, we identified a list of barcodes that represent high-quality single-cells with individual ID for each library. We then loaded the fragment files containing these barcodes from all libraries to one `ArchR` project for integrated analysis. Dimension reduction on this full dataset was performed on the binary tile matrix, selecting the top 30,000 variable tiles and outputting 50 reduced dimensions with 'addIterativeLSI()' function in `ArchR`. We then feed this LSI projection to the 'reducedMNN()' function in R package 'batchelor' to remove batch effects across libraries22. We implemented a wrapper function to add MNN-adjusted dimensions to the `ArchR` project, enabling seamless downstream analysis. Clusters were identified with a resolution of 0.8. For visualization, the fastMNN-adjusted reduced dimensions were used to derive a UMAP embedding with `minDist=0.8` and `spread=1`.

To calculate gene-activity scores (GA scores) from scATAC-seq profiles, we generated an in-house gene reference set from the GENCODE v19 annotation. Basically, we started from all the gene symbols in the full GENCODE annotation and removed those whose `gene_type` map to one of snRNA, misc_RNA, snoRNA, rRNA, miRNA, pseudogene, polymorphic_pseudogene, IG_V_pseudogene, TR_V_pseudogene, IG_C_pseudogene, TR_J_pseudogene, IG_J_pseudogene, processed_transcript, sense_intronic, 3prime_overlapping_ncrna and sense_overlapping, keeping 32,885 genes on chr1-22 and chrX. We then extracted the transcript start sites (TSS) and exons for these genes and constructed a gene annotation object that was added into our `ArchR` project. Our custom annotation included important marker genes that are missed in the default hg19 annotation used by `ArchR`, such as gene *RP11-291B21.2* (also known as *LINC02446*, **Figure 3.1h**), a long noncoding RNA that marks activated $CD8^+$ T cells. We next calculated the GA-score using default parameters.

To better annotate cell types in our scATAC-seq data, we integrated it with our Azimuth-annotated scRNA-seq data and transferred the annotation labels to scATAC-seq cells. We

first performed unconstrained integration using the `addGeneIntegrationMatrix` function in the `ArchR` package. We then examined the confusion matrix between cell clusters and annotated cell types. Several clusters contained mixed cell types from the reference dataset. Upon further speculation, we found these clusters tend to have higher rates of mitochondrial DNA and lie between well-defined cell types in the UMAP, suggesting these cells are of lower quality or are potential unremoved doublets. We excluded these cells from the dataset and performed constrained integration, where we restricted cells to three groups: T/NK cells, monocytes/DC, B cells and performed integration separately in each group. After this round of constrained integration, we found several T/NK cell subtypes in L2 have very low cell numbers in scATAC-seq data, we therefore only kept labels with sufficient cell number ($CD4^+$ naïve, $CD4^+$ TCM, $CD4^+$ TEM, $CD8^+$ naïve, $CD8^+$ TEM, MAIT, NK, NK_CD56bright, Treg) and performed another round of constrained integration. Finally, we re-calculataed the LSI, MNN-adjusted dimensions and the UMAP embedding and Leiden clustering on the remaining cells.

To identify candidate cREs, we first produced pseudo-bulk group coverages in each Leiden cluster and used the three studies as sample labels in `addGroupCoverages(sampleLabels="Study")`. We then called reproducible cREs by setting `reproducibility=2` in `addReproduciblePeakSet`. In this way, we were able to identify cREs that are called in at least two of the three studies in our data.

Topic modeling on scATAC count data.

Fitting the topic model. Topic modeling was performed using the R package `fastTopics`. We retrieved the cell-by-cRE count matrix from the `ArchR` object. In practice, we considered two aspects in fitting Poisson NMF to our count data. First, fitting the topic model on the full data is computationally expensive. Second, the NMF problem is non-convex, meaning that each model fit returns slightly different results, making it difficult to com-

pare the output using different parameters even on exactly the same data. To speed up the model fitting process, we randomly downsampled 10,000 cells. For cREs that have zero counts in the 10,000 sampled cells, instead of removing them from the matrix, we further sampled cells where they have non-zero counts. This ensures that we can project the fitted model to the full count matrix. To make sure we can easily compare multiple model fits on the same data, we first performed NMF using a small number of total topics (k), and then fit NMF with more topics conditioning on the previous model fit. We started by fitting the topic model with k=6 using the `fit_topic_model` function, using 100 main iterations and 200 refining iterations (`numiter.main=100, numiter.refine=200`). This returned a multinomial topic model fitting, which was then projected to the full count data using the `predict` function implemented in `fastTopics`. To fit a model with eight topics (k=8), we propagated the loading matrix and the factor matrix from k=6 with two more columns of uniformly distributed values (1/k for loading matrix and 1/(number of cREs) for factor matrix). We then applied the steps adapted from the `fit_topic_model` function. The expanded loading matrix and factor matrix were passed into `init_poisson_nmf` together with the downsampled count matrix to initialize a new Poisson model. Then, the model was fitted with the EM method for 100 iterations (main fitting) and updated with SCD method for 200 iterations in two consecutive runs of the `fit_poisson_nmf` function. The fitted Poisson model was converted to a multinomial NMF model with `poisson2multinom` function. The output of this final step is a cell-by-eight-topic loading matrix. We iterated this process with 10, 12, 14, 16, 18 and 20 topics. This framework allowed us to keep the order of topics constant, making it easier to compare across model fits, while also updating them when more topics are added. To visualize topic modeling results in a Structure plot, we performed PCA on the loading matrix (after centering and scaling) and used the rotated data matrix for K-means clustering (K=30). Then, we randomly selected 2.5% of cells from each cluster, resulting in a subset of ~5,500 cells for plotting.

Calculating gene scores. To define a molecular program underlying each topic, we relied on the factor matrix. We selected the top 10% of cREs with the highest score in each topic. To calculate gene-level scores from cRE-level scores, we applied `ArchR`'s exponential-weighting strategy to calculate gene activity scores. Briefly, scores of cREs within the gene body are directly summed up, and scores of cREs up to 5 Kb upstream of the gene TSS are weighted by distance-based power-law. We then calculated the Z score of each gene across all the topics.

Stratified-LDSC analysis on top cREs in each topic. For s-LDSC analysis, each cRE was extended to 1,500 bp around its center. We then selected cREs with the highest 10% factor scores from each topic for s-LDSC analysis. We used GWAS summary statistics of 11 immune-related diseases and 36 blood cell type traits, and height as a negative control. GWAS summary statistics were munged by `munge_sumstats.py`. LD score calculation and s-LDSC analysis were carried out according to LDSC documentation.

Trajectory analysis in topic model. We defined the cell trajectory directly from topic loadings with slight modifications to accommodate the analysis workflow of the `ArchR` package. As a proof-of-concept, we first scrutinized the B cell trajectory. Since k2 represents naïve B cells, its loadings are the highest in naïve B cells and decreases in memory B cells and plasmablast. To construct a trajectory that represents B cell maturation, we used the reverse order of k2 loadings, such that the trajectory value increases as naïve B cells transit into memory B cells. The trajectory was restricted to L1-annotated B cells, and we set the value of all other cell types to NA, so that `ArchR` does not use these cells in the analyses. The trajectory values were then scaled to the range of 0-100 for downstream analysis. To study the change in the proportion of memory B cells or plasmablasts along the trajectory, the trajectory was divided into percentiles, and we calculated the proportion of non-naïve B cells in each percentile according to L1 annotation. When building a

trajectory for k9, we removed cells whose k9 loadings are below 0.1, as these likely represent background noise rather than biologically meaningful variations.

After deriving the trajectory values from cell loadings, we added the trajectories to the `ArchR` project as a metadata column. To visualize the changes of gene activity scores along the trajectory, we used the `getTrajectory` function, followed by `plotTrajectoryHeatmap` functions with options `varCutOff=0.8, returnMatrix=TRUE`. The heatmap was visualized using ComplexHeatmap[180].

To assess the relevance of k9 trajectory to RA, we first asked in a cluster-based analysis, how many differentially active genes can be found between healthy and RA cells. To do so, we compared all RA cells with all healthy cells in k9 trajectory regardless of k9 loadings using the `getMarkerFeatures` function. We then grouped cells into quintiles according to their k9 loadings, where higher quintiles were enriched for more RA cells. We next tested for differential gene activity between all cells in the first quintiles and RA cells in the higher quintiles (second and above). Note in this test we used healthy and RA cells in the first quintile as control, following the idea that RA cells in the first quintile are epigenetically similar to healthy cells.

Chromatin accessibility QTL mapping.

RASQUAL caQTL mapping. Chromatin accessibility QTL (caQTL) were first mapped using `RASQUAL` on three grouping levels, whole blood-like (WB-like), L1 and L2 annotations. We generated pseudobulk counts by summing single-cell counts across cell barcodes within each group. For the WB-like group, we included all cREs and all individuals. For caQTL mapping in L1 and L2 cell types, we only included cell types with at least 50 cells in at least 10 individuals. From the pseudobulk count table, we calculated library sizes and phenotype PCs (after scaling and centering; using 'prcomp()' function in R). To get allelic-specific read counts, we extracted reads from each group using 'fil-

terbarcodes' command from sinto v0.7.5 and counted allelic-specific reads using 'createASVCF.sh' from RASQUAL. We only kept bi-allelic SNPs with at least four minor allele counts across tested individuals. We included library size as offsets and five genotype PCs, the number of cells, and the GC content for each cRE as covariates. RASQUAL was run in nominal mode and permutation mode. We extracted the lead SNP for each tested cRE and used nominal and permuted $\log_{10}$(q-value) from `RASQUAL` to calculate empirical p-values with `empPvals` function from the `qvalue` R package, and then derived q-values from the empirical p-values. We used q-value below 0.1 as the cutoff for significant caQTL.

To test the enrichment of `RASQUAL` caQTL from WB in bulk caQTL from LCL, we obtained summary statistics from a previous study46. We extracted lead SNP for each cRE in the LCL dataset and ranked them by their significance. We then tested the enrichment of `RASQUAL` lead SNPs in the top 1%, 5%, and 10% of the most significant LCL caQTL lead SNPs using Fisher's exact test.

sc-PME caQTL mapping. Single-cell caQTL mapping with the PME model was performed in three studies separately. For continuous covariates, we included top five genotype PCs, top five LSI dimensions, TSS enrichment scores, fraction of mitochondrial reads, $\log_{10}$ of number of unique fragments, all of which are scaled and fitted as fixed effects. We also included libraries and donors as random effects. The Poisson mixed effect model was fitted using the `glmer` function (`family=poisson`) in the `lme4` R package. We set the additional options as `nAGQ=0,control=glmerControl(optimizer="bobyqa",calc.derivs=F)` to save computational time for model fitting. We performed meta-analysis using effect sizes and standard errors from all SNPs in the three datasets and ran Metasoft without genomic control. For downstream analysis, we used effect sizes and standard errors from the random effects model and p-values from the Han and Eskin's Random Effects model (RE2)69. To call significant caQTL in the meta-analyzed sc-PME results, we first applied Bonferroni correction for all SNPs in a given cRE, extracted the lead SNP for each cRE,

and then calculated q-values from the Bonferroni-adjusted p-values across all lead SNPs. We used a q-value below 0.1 as a cutoff for significant lead caQTL.

Colocalization of caQTL with bulk eQTL and GWAS.

To perform colocalization between caQTL and eQTL, we used the eQTL summary statistics from the DICE study we published before2. We tested for colocalization between a caQTL and an eQTL when their corresponding lead SNPs are among the overlapping SNPs and there are more than 150 overlapping SNPs.

To perform colocalization between caQTL and GWAS, we used GWAS summary statistics of 11 immune-related diseases and 36 blood cell type GWAS that we have accessed and processed previously2. Briefly, we defined a GWAS locus as a 1 Mb window centered around a SNP with a p-value below 1e-7, starting from the SNP with the smallest p-value, removing all SNP lies within 1 Mb, and iteratively identified all GWAS loci until no SNPs with p-value below 1e-7 are remained. Similar to eQTL, we tested for colocalization between a caQTL and a GWAS locus when their corresponding lead SNPs are among the overlapping SNPs and there are more than 150 overlapping SNPs. All colocalization analyses were performed with the `coloc.abf` function from R package `coloc` v5.2.148.

3.6 Supplementary Notes for Chapter 3

Genome-wide coverage of scATAC-seq reads

We sought to test whether the number of scATAC-seq reads in each individual cell would be sufficient for demultiplexing pooled libraries. We quantified the number SNPs in the 1000G data that overlap with at least one or two unique reads in one scATAC-seq library. Since donors we collected are all East Asians, we first selected bi-allelic SNPs with a minor allele frequency (MAF) above 5% among the 493 East Asians (EAS) in the 1000G dataset,

keeping ~5.6 million SNPs[181]. For the scATAC-seq library, we kept cells with at least 500 unique fragments and a TSS Enrichment score of at least four. We then randomly sampled 5,000 cells, extracted all reads from these cells using sinto, and calculated the coverages over the 1000G SNPs using `samtools depth`. We found that an average of 188 SNPs were covered with at least two reads (median: 158) across all cells. Because scATAC-seq is a readout of genomic DNA in a single cell, any given single nucleotide position in the genomic can only be covered by at most two unique reads. In fact, only 368 (7.36%) cells have at least ten sites covered by more than three reads. We concluded that these reads may be due to misalignment and they constitute a negligible portion of the data.

Identifying healthy individuals with expanded T cell population

In our multiplexed scATAC-seq libraries, we expect cells from all individuals in a single library to be well-mixed when visualized on a 2-D UMAP embedding. We observed for all libraries except for one that contains four healthy individuals. Intriguingly, two of the four samples form their own clusters when this library was analyzed separately. For instance, cluster 4, 5, and 7 were exclusively donor PBMC003, while clusters 3 and 6 were exclusively PBMC004. Moreover, PBMC003 and PBMC004 are depleted with monocytes and B cells, represented by cluster 1 and cluster 2, respectively (**Figure 3.6a**). Upon scrutinizing marker genes for clusters that are unique to PBMC003 and PBMC004, we found markers for activated T cells like *IFNG*, *CD28* and *IL2RA* as well as cell proliferation like *CDK6* (**Figure 3.6b,c**), suggesting that PBMC003 and PBMC004 are two seemingly healthy individuals that have expanded T cell populations that set them apart from the other two healthy individuals in the same pool. Given our multiplexed experimental design, we concluded that the separation of cells from these two individuals is due to biological differences rather than batch effects.

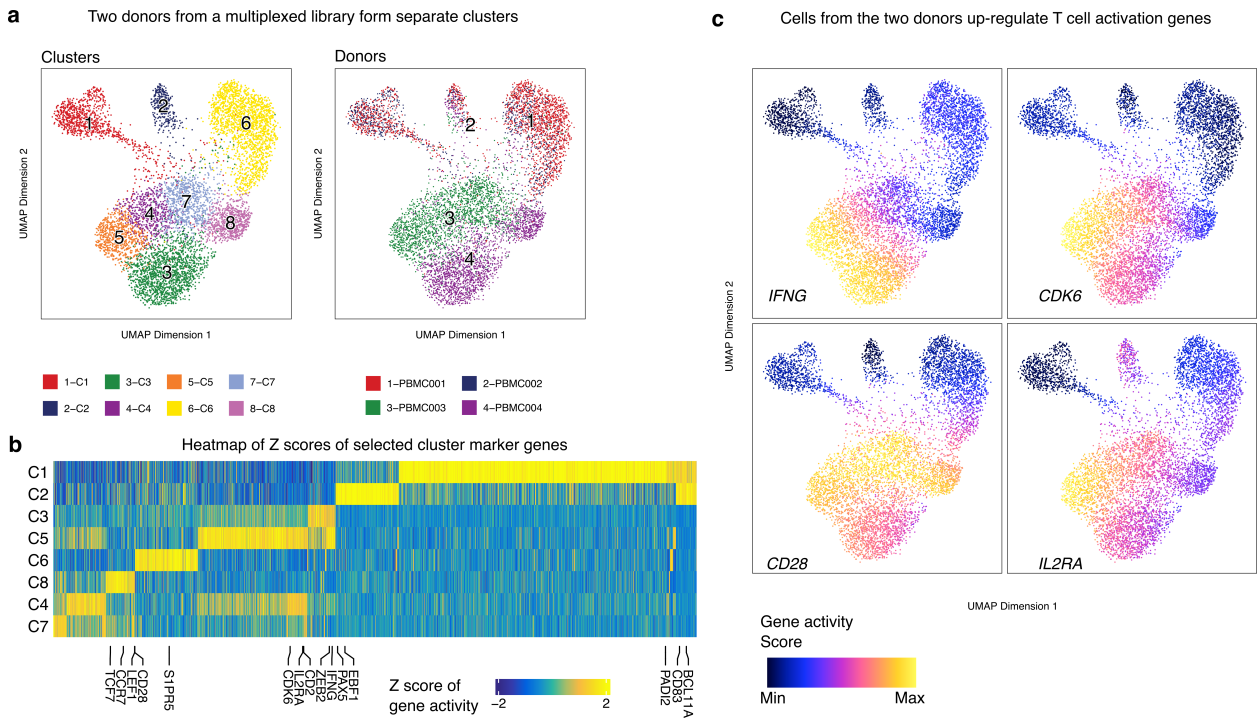

Figure 3.6. Two donors from a multiplexed library show T cell activation signatures.
a, Left: UMAP of the multiplexed library containing four donors colored by cell clusters. Right: UMAP colored by donors. **b,** Heatmap showing gene activity Z scores of cell cluster markers. **c,** UMAP colored by imputed gene activity scores of genes representing activated T cells.

Cell type annotation using scRNA-seq data

To annotate cell types and subtypes, we relied on the scRNA-seq dataset that is annotated by `Azimuth` and then integrated it with our scATAC-seq data to transfer cell type labels (Methods). We first tested this for all cells from the 33 individuals we collected in this study. To assess the quality of cell type annotation, we examined the number of Tregs identified and concluded that both the number (183) and percentage (0.228%) of annotated Tregs in PBMC are underestimated in our dataset. We reasoned that the relatively small total cell numbers in our dataset makes it difficult to confidently identify rare cell types. This prompted us to integrate our dataset with previously published scATAC-seq PBMC datasets, doubling the total number of cells. In this integrated dataset, we identified much more Tregs (2,390), and their proportion in PBMC is also higher, which is more consistent with known estimations. Notably, identified Tregs in our data increased to 2,390 from 183, consistent with the idea that increasing the total number of cells not only increases the absolute number of rare cell types, but also makes it easier to identify these rare cell types using existing computational methods.

`ArchR` offers two strategies for integrating scATAC-seq data with scRNA-seq data: unconstrained and constrained integration. In unconstrained integration, all cells in scATAC-seq are matched to all cells in scRNA-seq; in constrained integration, only cells from a certain group are used for integration, such as B cells and monocytes. We found that constrained integration is better at capturing rare cell types. For example, we identified 16 plasmablasts in unconstrained integration, while 472 plasmablasts were identified in constrained integration. Similarly, the number of Treg increased from 2,390 in unconstrained integration to 3,961 in constrained integration. Therefore, we used cell type labels from constrained integration for downstream analysis.

Optimizing estimates of correlation coefficients between topic loadings and gene activity scores

To enhance the utility and interpretability of continuous cell trajectories constructed from topic loadings, we sought to calculate the correlation between a trajectory and gene activity scores. To get better estimations of the correlation coefficients, we need to overcome the sparsity in scATAC data, which can cause underestimated correlation coefficients. Therefore, we adapted the `ArchR` method to generate pseudo-bulk replicates by grouping similar single-cells together. By default, `ArchR` creates a K-nearest neighbor graph in a latent space to identify similar cells[148]; in our implementation, we used the scaled loading matrix from `fastTopics` as the latent space to group similar cells, aiming for around 4,000 pseudo-bulk replicates with 50 cells in each replicate (`knnIteration=4000`, `k=50`). As a result, we identified 3,989 pseudo-bulk replicates consisting of 124,858 unique single-cells. Having identified the pseudo-bulk replicates, we calculated the mean gene activity scores across all cells belonging to each replicate. Accordingly, we also calculated mean cell loadings across all cells in each replicate. We further removed genes with mean activity score below 0.3 in all pseudo-bulk replicates, as these genes are likely to have very low expression levels. When calculating Spearman's correlation coefficients between a given topic and gene activity scores, we only included pseudo-bulk replicates with a loading score bigger than 0.01. Using this method, we found that Spearman's correlation coefficient between k19 (memory B cells) loadings and the memory B cell marker gene *TNFRSF13B* is 0.852, compared to the drastically lower estimation of 0.291 when single-cell measurements were used (**Figure 3.7**).

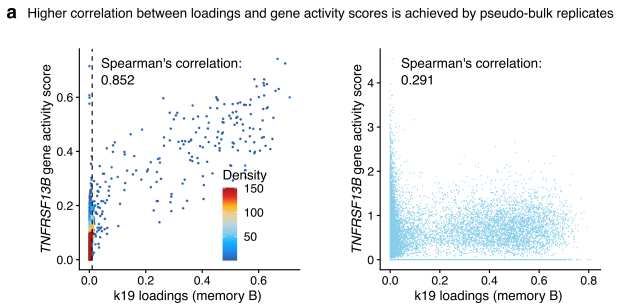

Figure 3.7. Pseudo-bulk replicates greatly enhance correlation estimation between loadings and gene activity scores.
Left: scatter plot showing the correlation between k19 loading and *TNFRSF13B* gene score. Dashed line highlights the cutoff (0.01) we used to exclude pseudo-bulk replicates with extremely low loadings from correlation estimation. Each data point represents a pseudo-bulk replicate consisting of 50 single-cells. Right: scatter plot showing the correlation between k19 loading and *TNFRSF13B* gene score in all single-cells. Note that cell density on the 2-D plot cannot be calculated because the majority of cells have a gene score of zero.

caQTL mapping with linear model and pseudo-bulk Poisson model

We also compared caQTL mapping results from our sc-PME model with the more commonly used pseudobulk linear model. Briefly, we mapped caQTL in pseudobulk whole blood (WB) using `QTLtools`[91]. Count data was normalized to a total of 1e4 in each sample, which is similar to CPM, followed by scaling across individuals and quantile-quantile normalization across peaks. We then calculated the phenotypic principal components (PC) from this scaled and normalized matrix. To be consistent with our `RASQUAL` caQTL calling, we used five genotypic PCs and seven phenotypic PCs as covariates. We ran `QTLtools` in both nominal mode (`-nominal`) and permutation mode (`-permute`). Using permutation output, we calculated q-values from the adjusted empirical p-value given by the fitted beta distribution and identified 3,417 cREs with significant caQTL across the genome. Notably, 3007 (88.0%) of these caQTL were also identified in `RASQUAL` WB caQTL; and 2,999 (87.8%) of these were identified in sc-PME WB caQTL.

We next compared the effect sizes estimated by the linear model with the effect sizes from `RASQUAL` and the sc-PME model. As expected, we observed high correlations between the effect size estimates, with Spearman's rho 0.86 and 0.75 for `RASQUAL` and sc-PME,

respectively (**Figure 3.8a**).

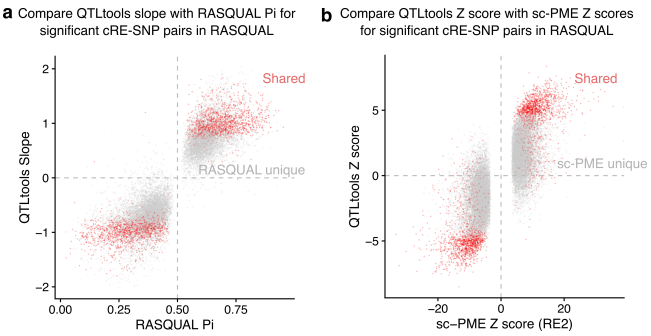

Figure 3.8. Linear model identifies fewer caQTL, although effect sizes are highly correlated with `RASQUAL` and the sc-PME.
a, A scatter plot comparing the effect size estimation from linear model in `QTLtools` and `RASQUAL` (Pi). Each point represents a significant cRE-SNP pair identified in `RASQUAL` output. Shared caQTL between `QTLtools` and `RASQUAL` are colored in red, while RASQUAL-unique caQTL are colored in gray. **b**, Similar to **a**, a scatter plot comparing the Z score estimation from linear model in `QTLtools` and meta-analyzed sc-PME output. Z scores for sc-PME models are calculated from p-values of the random effects method (RE2) in Metasoft. Each point represents a significant cRE-SNP pair identified in sc-PME output.

Unique caQTL in `RASQUAL` and sc-PME have smaller effect sizes in linear models compared to those shared between the two groups of methods, suggesting that the linear model is under powered to detect caQTL with smaller effect sizes (**Figure 3.8b**). Note the Z scores for sc-PME models are calculated from p-values of the random effects method (RE2) in Metasoft [182], because we used this p-value for identifying significant caQTL.

Higher dropouts lead to higher false discovery rates in sc-PME model

We sought to understand whether the single-cell Poisson mixed-effects model can lead to spurious associations when modeling single-cell count data. To this end, we performed a permutation analysis by changing the order between genotype and individual labels, thereby preserving the correlation structure between phenotype and covariates in our sc-PME model. We ran the permutation analysis on all the 27,117 cREs on chromosome 1 in the cohort of 33 individuals we collected. We calculated the q-value for each lead SNP in the same way as we did for non-permutation mapping. In this analysis, 6,371 cREs had a q-

value below 0.1, suggesting a false discovery rate (FDR) of 30%. In particular, the smallest p-values are concentrated in cREs with more than 85% of zero counts (**Figure 3.9**).

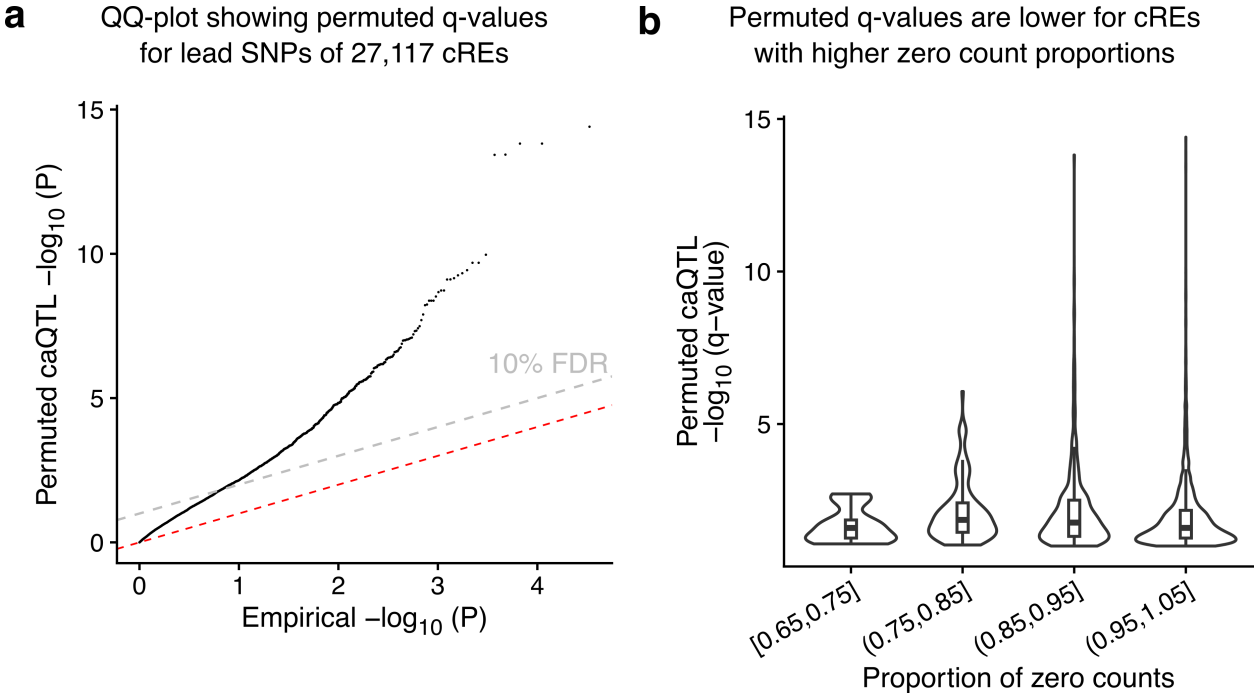

Figure 3.9. Higher dropouts lead to higher false discovery rates in sc-PME model.
a, QQ-plot showing the permuted q-values for lead SNPs of 27,117 cREs. Grey line represents the 10% FDR cutoff. **b**, Violin plots showing the relationship between permuted q-values and zero count proportions of cREs across all cells. The lowest q-values are particularly enriched in cREs with more than 85% of zero counts.

Comparing Z scores of caQTL-eQTL pairs in colocalization analysis

To further understand how a given caQTL-eQTL pair colocalizes in one context but not another, we compared effect sizes of all 7,063 colocalized caQTL-eQTL pairs in each of the five contexts. Similar to our observation on gene/cRE activity levels, caQTL and eQTL have the largest Z scores in colocalized contexts, followed by in non-colocalize contexts, and the smallest Z scores are observed in not-tested contexts (**Figure 3.10a,b**). This supports our argument that even when a caQTL-eQTL pair colocalizes in one context, it often does not colocalize in all other contexts because the eQTL/caQTL may be weaker due to

the molecular phenotype having lower activity in another context. Therefore, a functional caQTL is often also an eQTL in the same context, but this does not guarantee that the same caQTL in another context is also an eQTL. On the contrary, the genetic effects on chromatin accessibility can be widely shared across contexts, but not all caQTL are functionally regulating downstream genes. We believe this conclusion also has great implications for the proper interpretation of caQTL-GWAS colocalization results, especially in the absence of eQTL.

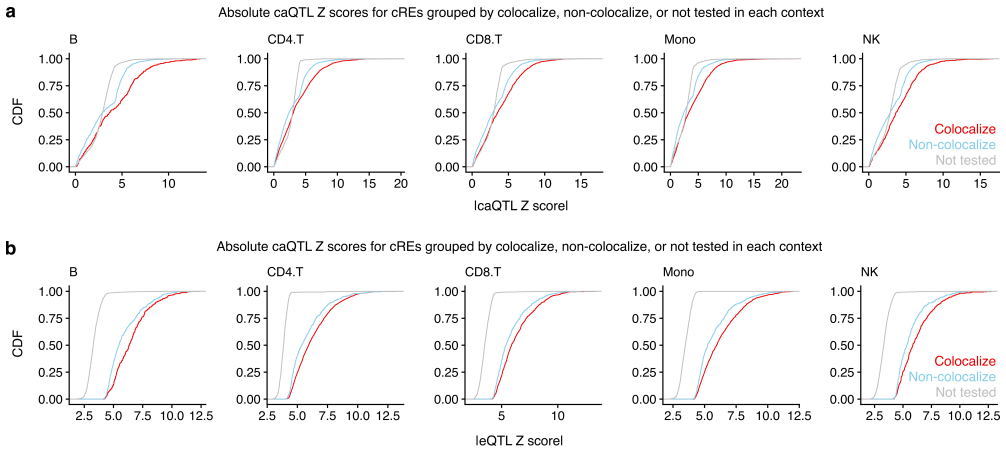

Figure 3.10. Weaker QTL explain the lack of sharing of caQTL-eQTL colocalization.
a, Cumulative distribution of absolute caQTL Z scores in five contexts for 7,063 colocalized caQTL-eQTL pairs in at least one context. Z scores in each context are grouped by whether a given caQTL is colocalized, tested for colocalization but does not colocalize (Non-colocalize), or not tested for colocalization (Not tested) due to the absence of eQTL or caQTL. **b**, Similar to **a**, showing the cumulative distribution of absolute eQTL Z scores. To allow for better comparison with caQTL, the 15 DICE cell types were mapped to the five common contexts by taking the maximum Z score across the cell types belonging to each context.

caQTL-eQTL colocalization across all pairwise contexts

Through our caQTL-eQTL colocalization analysis, we found that many caQTL are shared across many contexts, but they only colocalized with eQTL in a few specific contexts, mostly due to the context-specificity of eQTL. Therefore, we proposed that it is important to distinguish a cRE being merely "active" from it being actually "functional". Func-

tional cREs are actively regulating the expression of their downstream genes, whereas active cREs are open chromatin regions that are not actively regulating gene expression in a given context. Crucially, both functional and active cREs can have significant caQTL. This implies that even though a caQTL does not colocalize with eQTL in one context, it might be an eQTL in another context, where it has causal functions.

To test this hypothesis, we performed cross-contexts colocalization analysis, where we tested for the colocalization between the caQTL in contexts A with the eQTL in contexts B for all pairs of contexts. In this expanded colocalization analysis, 42% of eGene colocalized with at least one caQTL in any context; and the number of colocalized eGenes increased 1.78 fold on average across contexts. Similarly, 18% of caQTL colocalized with at least one eQTL in any context; although this increase is not as dramatic as that of eQTL, the number of colocalized caQTL increased by 2.2 fold on average across contexts. Thus, when all contexts are considered, the proportion of caQTL that are also eQTL almost doubled compared to only using matched cell types.

To understand the cross-contexts colocalization in better detail, we first "fixed" the caQTL to one context and scrutinized its colocalization with eQTL in all other contexts. By doing so, we can alleviate the problem of caQTL sharing and potential noise in caQTL p-values across contexts (even though the lead SNP is shared). We then compared the properties of eQTL and eGenes across contexts while caQTL remains constant. To illustrate this process, we use B cells as an example. For each caQTL in B cell, we identified all the colocalizing eQTL across contexts (B, $CD4^+$ T, $CD8^+$T, Monocyte, NK). These eQTL were divided into three categories: (1) the same eGene is also colocalized in B cells (i.e. shared colocalization; COLOC), (2) eQTL is significant but does not colocalize in B cells (Non-COLOC), and (3) eQTL is not significant in B cells thus not tested for colocalization with the caQTL (not-tested). We then scrutinized the effect sizes of these eGenes in B cells, stratified across the three categories. Consistent with our previous analysis for

colocalization in matching contexts, many eGenes were not tested for colocalization in B cells because of weaker eQTL. The effects are the weakest in the "not-tested" group, consistent with these eQTL being not significant in B cells. However, these genes have significant eQTL in another context that colocalized with the caQTL in B cells, raising the possibility that these caQTL do not have causal functions in B cells, although they are statistically significant. We performed the same analysis for all other contexts and observed the same trend (Supplementary Fig. 6a). We also compared the effect sizes of eQTL and reached the same conclusion, although the differences between the three categories are smaller (Supplementary Fig. 6b).

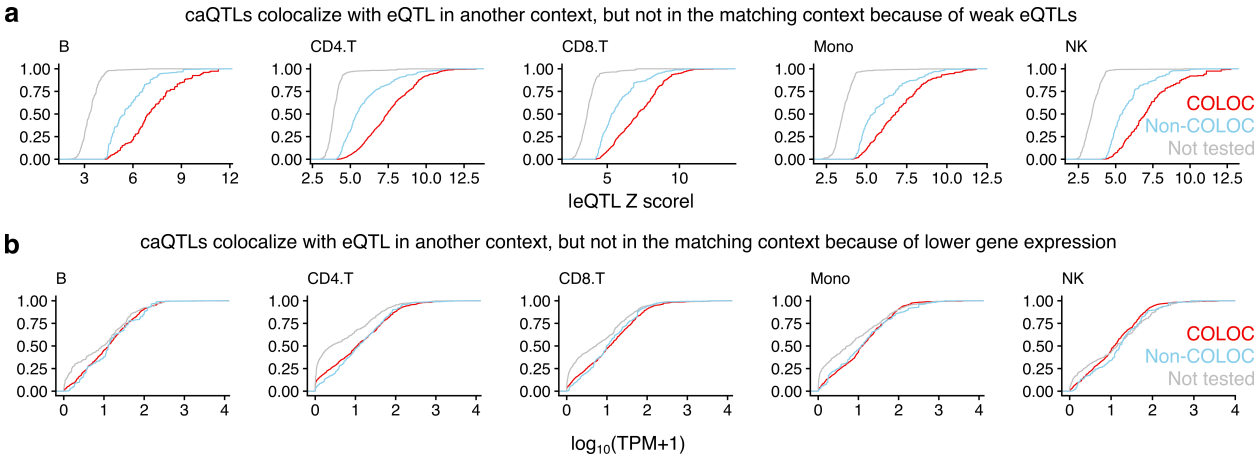

Figure 3.11. Cross-context comparison of QTL effects.
a, Cumulative distribution of absolute eQTL Z scores for colocalized eGenes across contexts. Z scores in each context are grouped by whether a given eQTL is colocalized (COLOC), tested for colocalization but does not colocalize (Non-COLOC), or not tested for colocalization (Not tested) due to the absence of eQTL in a given context. **b**, Similar to **a**, showing the cumulative distribution of gene expression levels. To allow for better comparison with caQTL, the 15 DICE cell types were mapped to the five common contexts by taking the maximum Z score and mean gene expression across the cell types belonging to each context.

3.7 Supplementary Figures for Chapter 3

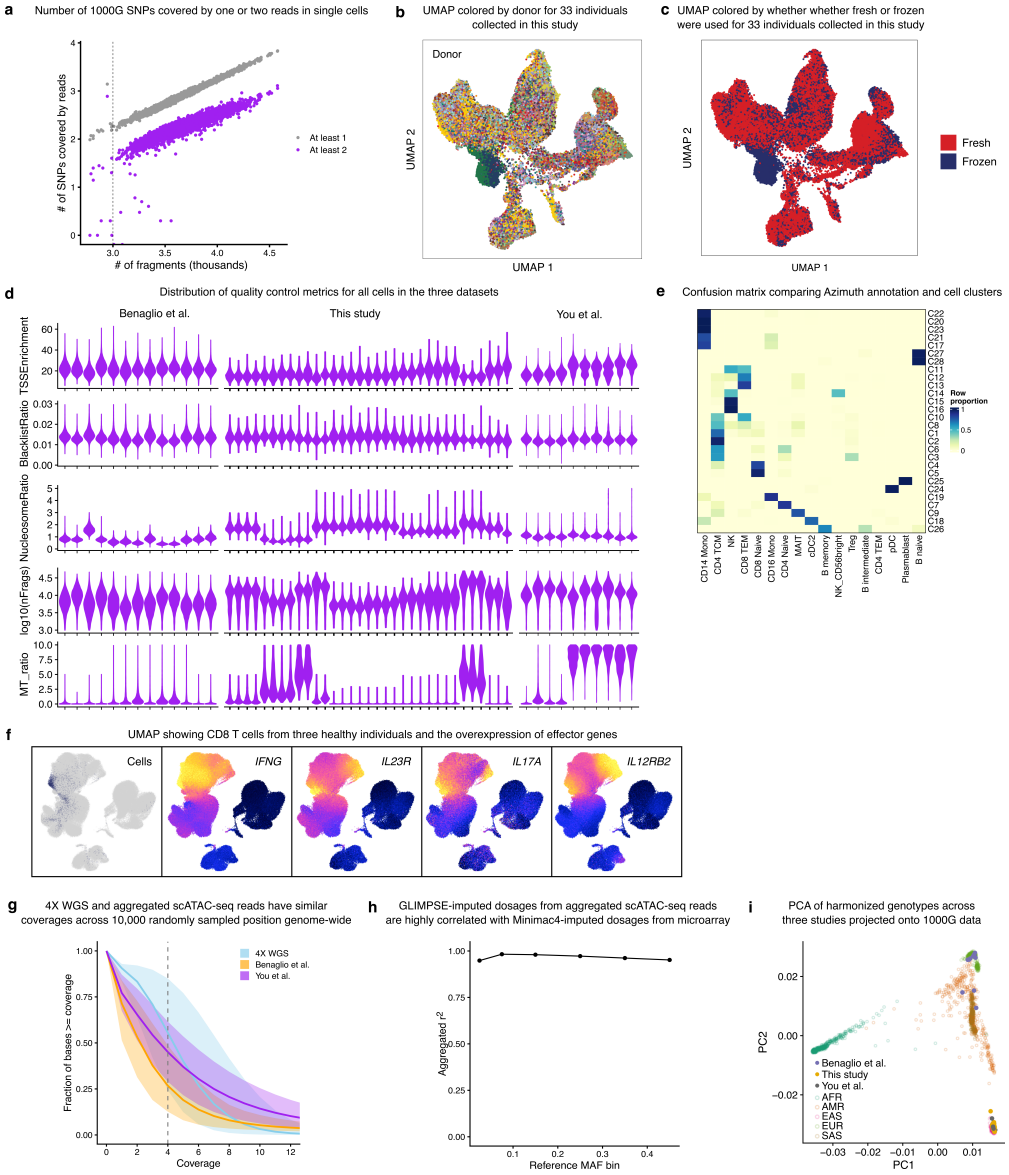

Figure S3.1. Quality control on scATAC-seq data and genotyping.
a, Number of 1000G SNPs covered by at least one or two reads correlates with the number of unique fragments in each cell. Dashed line marks 1,000 unique fragments. **b**, UMAP for 33 individuals collected in this study, colored by donor. **c**, UMAP for 33 individuals collected in this study, colored by whether fresh or frozen PBMC were used. **d**, Distribution of quality control metrics in the three datasets. **e**, Comparison `Azimuth` L2 annotation and cell clusters. Heatmap is colored by row normalized proportions, such that the cell type compositions sum up to one on each row. **f**, UMAP for $CD8^+$ T cells from three healthy individuals with expanded $CD8^+$ T cell populations and gene activity scores for effector genes. **g**, 4X WGS and aggregated scATAC-seq reads have very similar read coverages. **h**, Line chart showing mean correlation between GLIMPSE-imputed genotype dosages from aggregated scATAC-seq reads and those imputed from microarray data using `Minimac4`. **i**, PCA analysis of individuals in this study with 1000G samples.

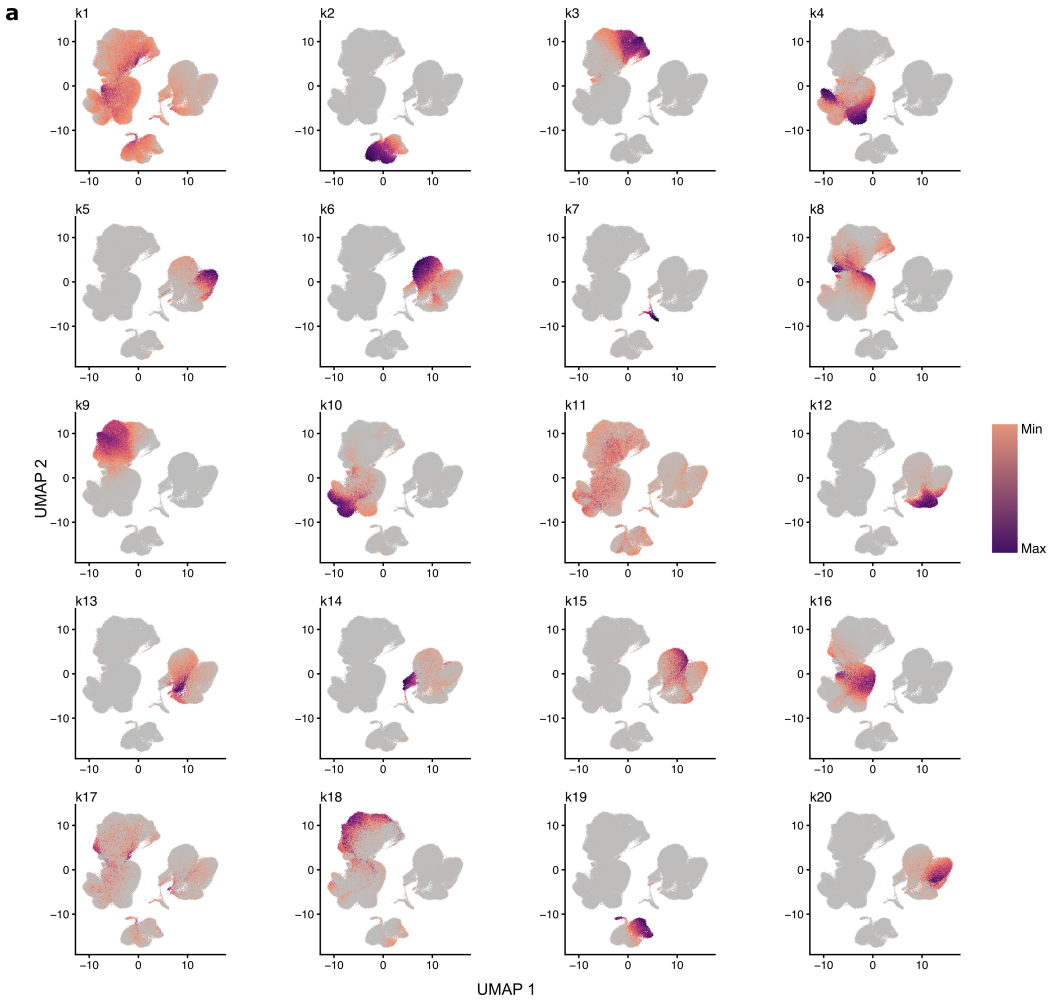

Figure S3.2. Visualization of cell loadings for the 20 topics in UMAP embedding.
a, UMAP plots showing the spacial distribution and quantity of loadings for the 20 topics. Loadings below 0.1 in a cell were set to 0 for visualization purposes (grey).

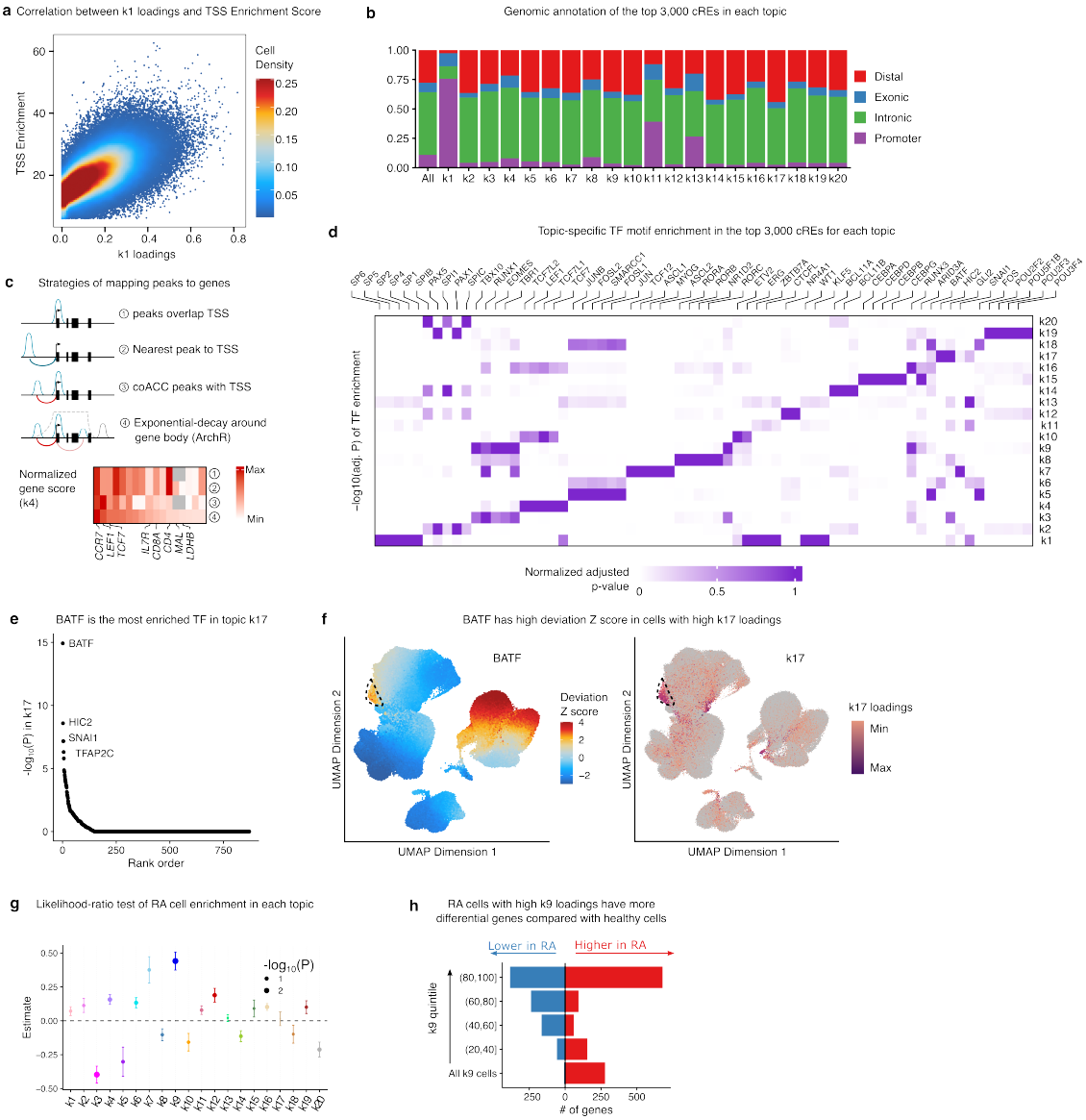

Figure S3.3. Topic analysis.
a, Correlation between k1 loadings and TSS Enrichment scores. **b**, Bar plot showing the genomic annotation of the top 3,000 cREs in each topic, compared to all cREs. **c**, top: schematic showing four strategies to calculate gene-level score from cRE-level scores. Bottom: gene-level scores calculated from the four strategies for cREs in topic k4, a naïve T cell topic. **d**, Adjusted p-values for TF motif enrichment in top 3,000 cREs in each topic. -$\log_{10}$(P) values are normalized for each topic. Top five enriched TFs are shown for each topic. **e**, Ordered adjusted p-values for TF motif enrichment in topic k17. **f**, UMAP colored by BATF deviation Z score (left) and k17 loadings (right). Circulated region represents cells with high k17 loadings and high BATF activity. **g**, Likelihood-ratio test of RA cell enrichment in each topic. Y-axis shows the effect sizes of RA status on each topic in the generalized linear model. Error bars represent 95% confidence intervals. **h**, The number of up-regulated and down-regulated genes between RA cells in the top four k9 quintiles and all cells in the first k9 quintiles, compared with the number of differential genes when all RA cells in topic k9 were compared with all healthy cells.

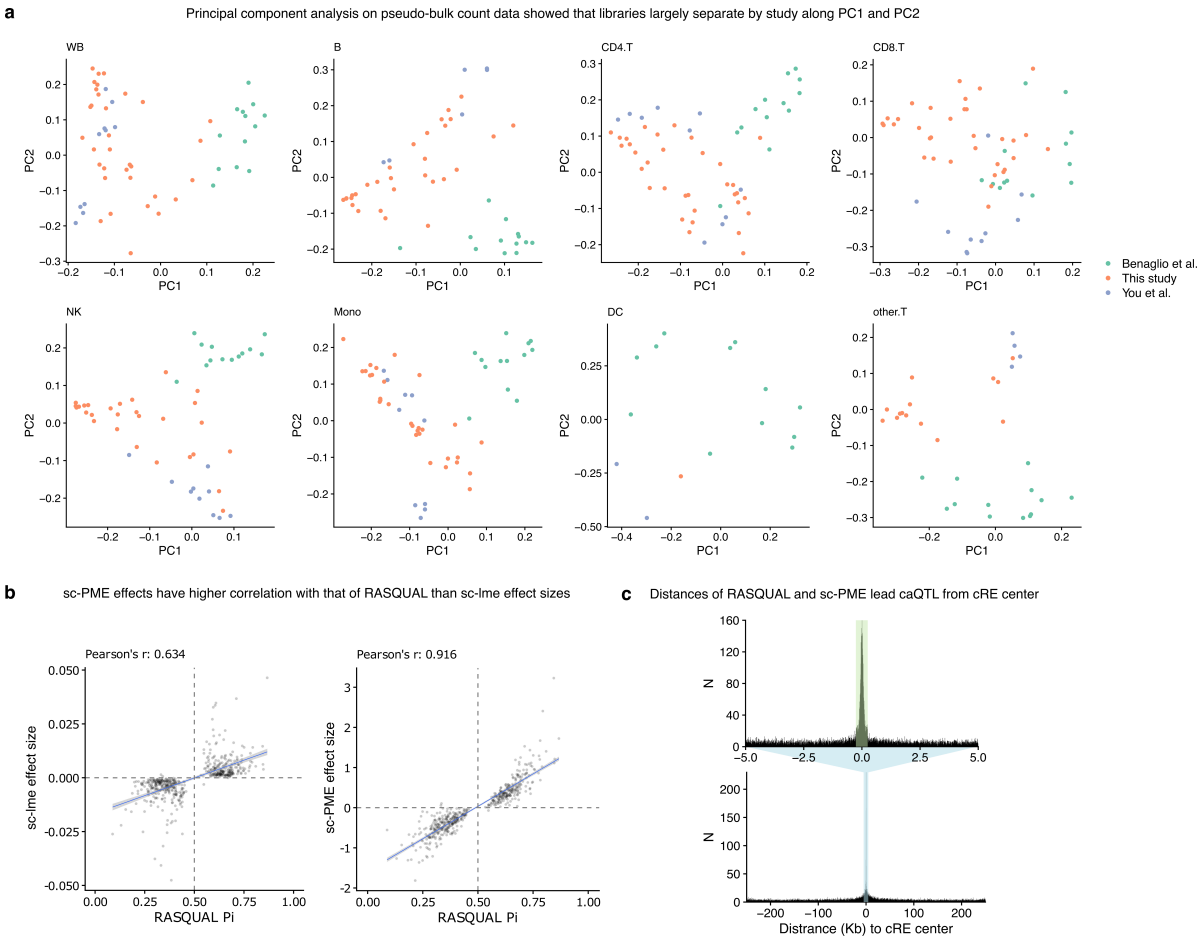

Figure S3.4. caQTL mapping using RASQUAL and sc-PME model in harmonized data.
a, Principal component analysis (PCA) on pseudo-bulk count data for WB and seven common immune cell types. Each sample is colored by study. **b**, Scatter plots comparing RASQUAL effect sizes (Pi) with sc-PME effect sizes (left) and sc-lme effect sizes (right). Only significant RASQUAL caQTL on chromosome one identified in 33 individuals collected in our study were plotted. **c**, Histogram showing distances from cRE centers to significant lead caQTL in WB from RASQUAL (top) and sc-PME (bottom). Green shaded region highlights cRE size (500 bp); blue shaded region highlights RASQUAL mapping window (10 Kb) relative to sc-PME mapping window (500 Kb).

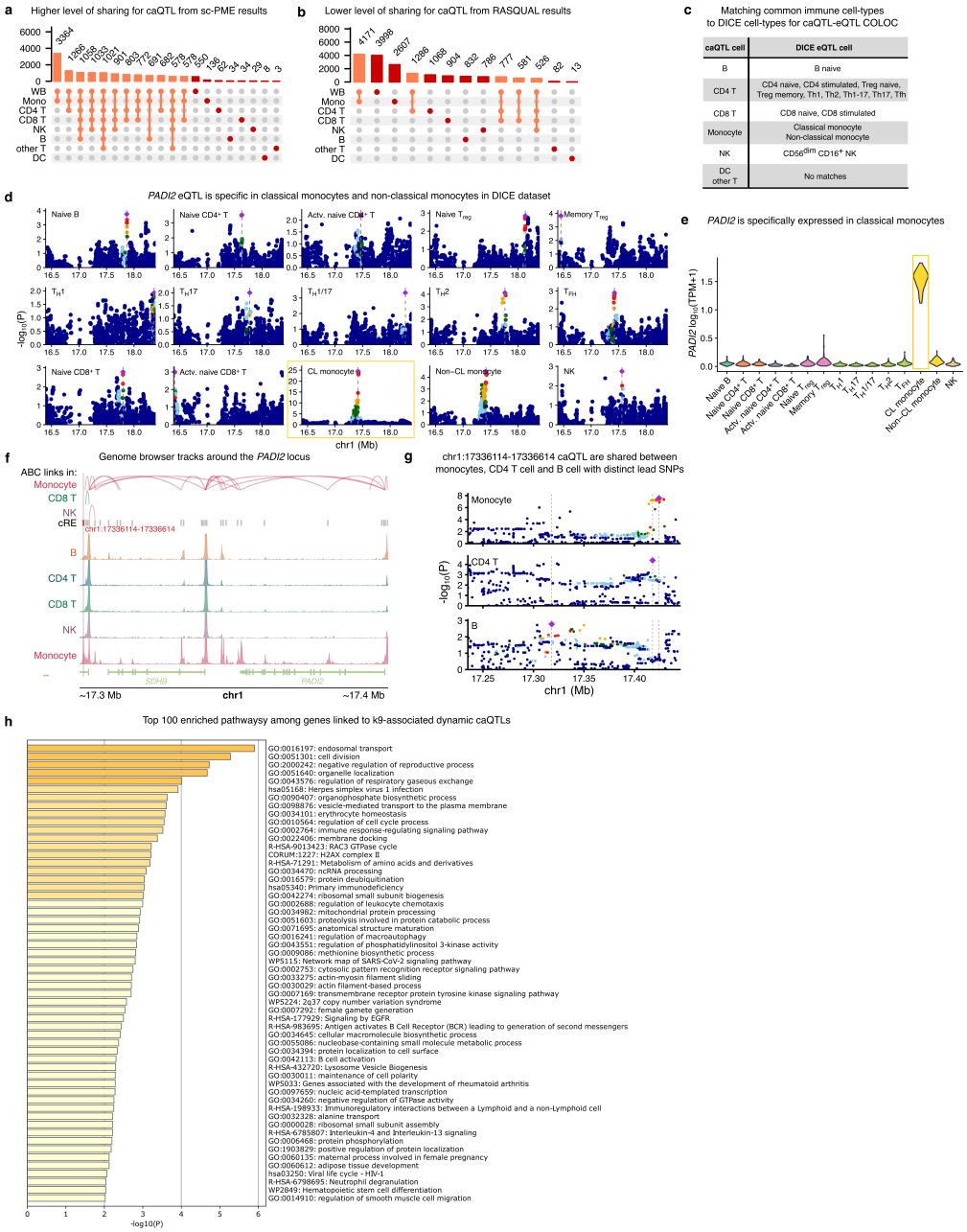

Figure S3.5. Comparative analysis of eQTL and caQTL.
a, Sharing of sc-PME caQTL in WB and seven common immune cell types. **b**, Sharing of `RASQUAL` caQTL in WB and seven common immune cell types. **c**, Mapping of 15 immune cell types in DICE data to the seven common cell types used in caQTL mapping. **d**, Manhattan plots showing that *PADI2* eQTL is only significant in classical monocytes (highlighted in colored box). **e**, Violin plot showing that *PADI2* gene is only expressed in classical monocytes. **f**, Genome browser tracks for the genomic locus around *PADI2*. Two cREs whose caQTL colocalized with *PADI2* eQTL were colored in red. **g**, Manhattan plots showing that cRE chr1:17336114-17336614 has shared caQTL in monocyte, CD8$^+$ T cell and B cell, but with different lead SNPs. **h**, Top 100 pathways enriched in genes linked to cREs with k9-interacting dynamic caQTL in CD8$^+$ T cells.

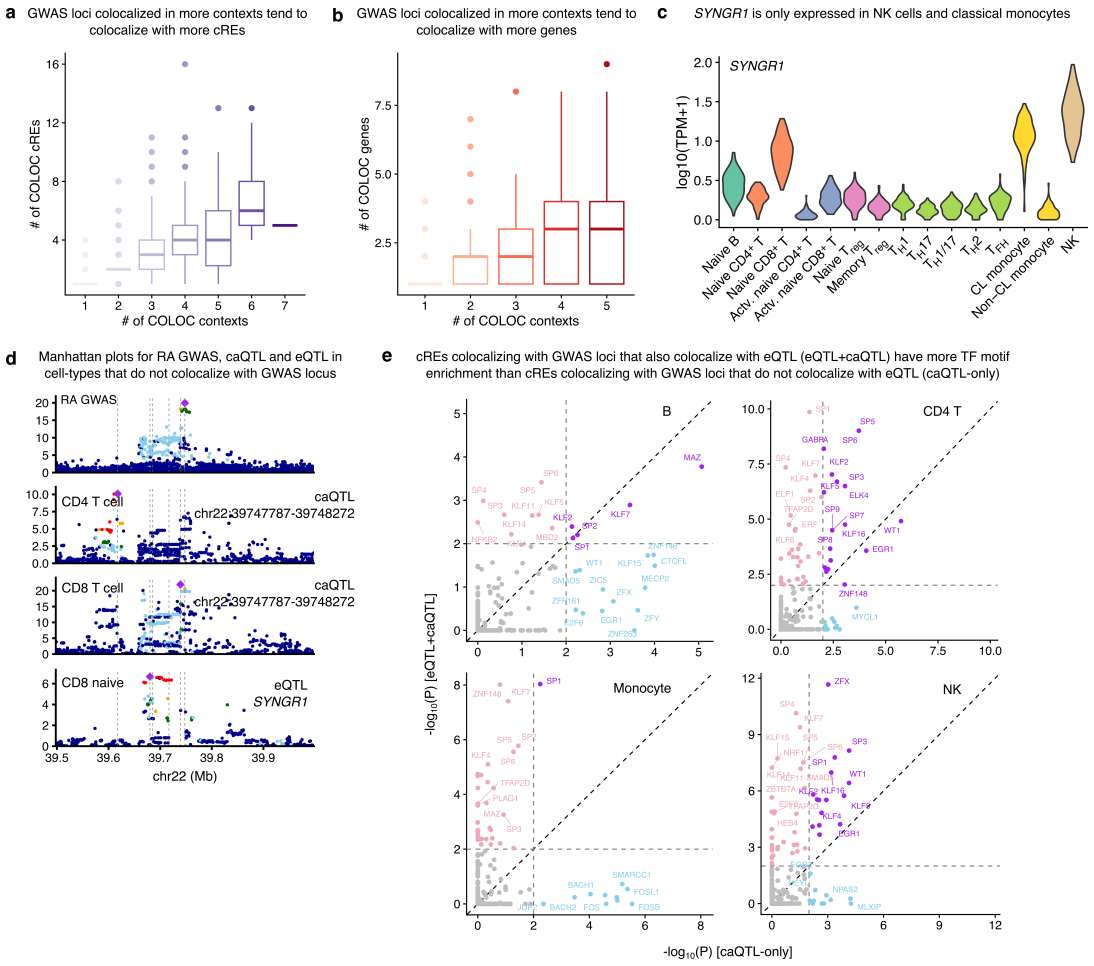

Figure S3.6. Comparative analysis of GWAS and caQTL.
a, Box plot showing that a GWAS locus that colocalized in more contexts also tends to colocalize with more distinct cREs. **b**, Box plot showing that a GWAS locus that colocalized in more contexts also tends to colocalize with more distinct eGenes. **c**, Violin plot showing that gene *SYNGR1* is only expressed in NK cells, classical monocytes, and to a lesser extent in naïve $CD8^+$ T cells. **d**, Manhattan plots showing GWAS, caQTL and eQTL around an RA GWAS locus. **e**, Comparison of TF motif enrichment p-values for cREs colocalizing with GWAS loci that also colocalize with eQTL (eQTL+caQTL) versus cREs colocalizing with GWAS loci that have no eQTL colocalization (caQTL-only).

CHAPTER 4

CONCLUSION AND DISCUSSION

4.1 Significance and limitations

The central goal of this dissertation is to evaluate how molecular QTL can be utilized to nominate potential causal genes and contexts for disease GWAS variants, and to better address the question of why studies so far have had very limited success in doing so, which has been proposed by some researchers as the *missing interpretation* problem[132].

In **Chapter 2**, I focus on the transcriptome in immune cells and comprehensively map the genetic control of gene expression and pre-mRNA splicing, by analyzing existing large-scale RNA-seq datasets in a uniform fashion. My integrative analysis on eQTL/sQTL and GWAS shows both the utility and limitations of eQTL data and generates several valuable insights. First, I highlight the pervasiveness of primary eQTL sharing among immune cell types and subtypes, contrary to previous believes that many eQTL are cell type-specific[29]. This implicates that collecting ever-more diversified immune cell populations in eQTL study will not lead to the discovery of new primary eQTL, as least in peripheral blood. Second, I demonstrate that sQTL explain 14% of GWAS loci that are not eQTL, highlighting the crucial role RNA splicing may play in disease progress. I show that significantly increase the sample size for eQTL study does not increase the number of colocalization with GWAS loci. In fact, increasing the number of samples by 100 folds only identifies 1.5-fold more colocalization events with GWAS. I also observe that GWAS loci not explained by eQTL tend to locate close to evolutionary constrained genes, and have more complex gene regulatory landscapes, consistent with recent theoretical analysis on the systematic differences between SNP found in GWAS and eQTL[34]. Finally, I use synovial fluid immune cells from RA patients to show that disease samples from the correct tissue capture genetic effects absent in healthy cells or *in vitro*-stimulated immune

cells. This highlights the importance to study disease-specific gene regulation.

In **Chapter 3**, I show that chromatin accessibility can greatly complement eQTL data in understanding disease GWAS. I compile one of the largest scATAC map in human PBMC to date and develop a workflow to demultiplex scATAC-seq libraries using low-pass WGS data. I also demonstrate that aggregated scATAC reads can be used to accurately impute genotype data. These strategies together makes population-scale scATAC study more cost-effective and increases the utility of scATAC data without genotype information for genetic studies, and can be readily applied to future experiment design. Using the harmonized scATAC data, I discover four times more caQTL compared to a previous study and show that caQTL colocalize with 50% more GWAS loci compared to eQTL alone[13]. Combined with eQTL/sQTL colocalization from **Chapter 2**, we can now explain nearly 75% of GWAS variants. While this is encouraging in terms of the number of colocalized loci, I show that it is often challenging to pinpoint the causal genes and contexts for caQTL-GWAS colocalization, especially when the GWAS loci do not colocalize with any eQTL. I propose that a regulatory element—for example, a peak in scATAC data—can be ***active*** in a given context, it is not necessarily ***functional*** in regulating gene expression. For a regulatory element to be functional, it needs additional mechanisms like certain transcription factor or co-factor binding and being physically linked to its target promoter. In line with this, scATAC data from *Drosophila* development revealed that pioneer TFs make chromatin accessible but other combinations of TFs are required for enhancer function[183]. I argue that studying disease-relevant contexts and integrating functional annotations on scATAC data is necessary to understand the biological process underlying caQTL.

The work I described in this dissertation is not without limitations. The study on molQTL here is restricted to primary QTL. This is partially due to the relatively small sample size that may not be sufficient for fine-mapping or conditional analysis. Statistical colocalization also assumes exactly one independent signal in each locus, which may

be violated in some cases. Indeed, a recent study reported that considering non-primary eQTL in adipose tissue increases colocalization by 46%[184]. Similarly, it is also difficult to conduct three-way colocalization in a caQTL-eQTL-GWAS tuple. I observe many cases in which both a caQTL and an eQTL colocalize with a GWAS locus, but the caQTL and eQTL do not colocalize. This is possibly because both eQTL and caQTL have limited power compared to GWAS in a colocalization test. Thus, it is challenging to directly link a GWAS locus to a regulatory element and to its target gene. Another limitation is the use of RA PBMC, rather than synovial fluid. This severely limits the ability to find disease-specific regulatory effects. While I define an RA-associated effector/memory $CD8^+$ T cells trajectory in PBMC population, it is far from the only RA-associated cell populations and it remains unclear whether this population has any role in disease progress in the synovial fluids. Last but not least, the eQTL and caQTL data analyzed in **Chapter 3** come from two independent studies. Different nomenclature used to identify cell subtypes are not completely consistent between the eQTL and caQTL data. This poses a challenge to identify cell type specific genetic effects in chromatin accessibility and gene expression. In the foreseeable future, scRNA and scATAC data from the same set of samples, or even multimodal measurements from the same cells can be used to address with issue.

4.2 Reflections on current paradigm and future directions

At the end of this dissertation, I would like to reflect on the current practices in the field and discuss possible future directions. The biological interpretation of GWAS results has been extensively studied in the past decade. This endeavor consists of at least two aspects: (1) knowing the causal gene of a given GWAS locus and (2) knowing the causal context in which it functions. To address these questions, the field has witnessed the formation of a research *paradigm*, namely, the collection of large-scale RNA-seq data and subsequent statistical integration of eQTL/sQTL with GWAS summary statistics. The recent develop-

ment in epigenetic QTL (histone modification, DNA methylation, and chromatin accessibility) mapping only reinforces this paradigm. The large number of scientific research articles that report newly collected data and tallying of colocalization with GWAS attest to the wide adoption of this research paradigm.

Conventional thinking has that if we can find a cell type-specific eQTL that is also a GWAS locus, we can simultaneously nominate the causal gene and cell type (or context) for this given GWAS locus. This line of thinking has been widely adopted in the field in the past few years and partially motivated the pursuit of cell type-specific eQTL and colocalization events. However, this interpretation is not necessarily true. Our work demonstrated that due to high level of QTL sharing, it is very difficult to pinpoint the correct causal context for GWAS variants, even more so than pinpointing the potential causal gene. Indeed, one study found that eQTL in whole blood (eQTLGen) colocalize with more brain GWAS loci than eQTL from brain tissues [185]. Similary, in **Chapter 1**, I found a large proportion of Alzheimer's disease (AD) GWAS colocalize with eQTL in peripheral blood monocytes. These signals likely arise from the shared eQTL between monocytes and microglias. Second, one GWAS locus often colocalizes with more than one eGene, possibly due to co-regulation. Third, important genes may be regulated by multiple redundant enhancers to maintain a stable expression level, thereby masking the effect of single enhancer and eQTL. Conversely, the same gene may be loosely regulated in an unimportant cell type, making the effect of a single eQTL detectable [134]. This implicates that the colocalized context is not guaranteed to be the causal context. Instead, it may precisely be the wrong context. Importantly, it is hard to disentangle these complex effects just based on eQTL and GWAS data, and it is impossible to estimate how much of the colocalization results are affected by the above reasons.

The recent advances in epigenetic, including chromatin accessibility, QTL (epiQTL) has expanded our understanding of disease GWAS, although I would like to argue that

current results raise more question than answers. On one hand, epiQTL have higher heritability than eQTL and they mediate more GWAS heritability. epiQTL also colocalize with more GWAS loci, ranging from ~25% to ~50% more depending on studies and traits analyzed. On the other hand, the majority of epiQTL has no effect on gene expression in the same context. It thus remains unclear how epigenetic features can mediate the genetic effects on phenotypes without affecting gene expression.

One hypothesis is that many enhancers are "primed", meaning that these enhancers are accessible but rapidly enhance the expression of their target gene in another context. Primed enhancers have been relatively well-documented in immune activation and development. In **Chapter 3**, I propose that similar ideas can be generalized to explain the lack of effect on expression for most caQTL across cell types. However, new data and methods are needed to fully understand this. In **Chapter 3**, I found that in $CD8^+$ T cells, enhancers underlying caQTL-GWAS colocalization without eQTL tend to be enriched for motifs of TFs related to long-term effector/memory and exhausted phenotypes. This suggests that adding TF binding information can help resolve the cell type and state functions of enhancers and help define their causal context. Indeed, many enhancers are highly accessible but not bound by TFs with specific functions, although they may have a strong caQTL and harbor motifs of multiple TFs. The functional consequence of these enhancers is only relevant with TF binding. One major challenge is that existing ChIP-seq data on TFs is scarce and current experimental methods can only profile one TF in each experiment. Unlike chromatin accessibility and histone modifications, there are at least hundreds of TFs in human genome, making it next to impossible to map all of them in a single study. Nevertheless, hypothesis-driven studies that focus on a few TFs at a time can still be immensely insightful on the selected TFs.

Alternatively, machine learning techniques can be used to carry out *in silico* experiments to understand how TF works. Specifically, deep neural networks have been trained

to predict scATAC signal at base-pair resolution[183,186]. Interpretation tools can subsequently be used to extract contribution scores for each of the four nucleotides at each base-pair position. These contribution scores can be clustered and aggregated to identify TF binding motifs, offering biological meanings to model outputs. Furthermore, manipulated sequences can be run through the trained model to decipher how sequence contexts affect chromatin accessibility. Using these techniques, a recent study reported that during *Drosophila* embryo development, pioneering TFs quickly drive chromatin accessibility throughout the genome, whereas more specific TFs later bind to these sites to induce tissue-specific gene expression patterns[183]. These discoveries are consistent with previous experimental approaches, but also offers novel insights on the sequence rules of TF lexicon and greatly improves the breath of questions that can be answered using scATAC-seq data.

It is well known that functional enhancers make contact with downstream promoters. Therefore, incorporating enhancer-to-promoter (E2P) links to caQTL/eQTL/GWAS data can help nominate causal genes. However, current E2P prediction methods often gives inconsistent results. It is unclear whether the inconsistency are due to technical differences or that they are capturing different aspect of gene regulation. Benchmarking these methods is also extremely challenging because of the scarcity of ground truth data. Therefore, future research could benefit from making better benchmark dataset with specialized experiments and developing E2P prediction methods that can better handle single-cell genomics data. Moreover, very few gene prioritization methods currently use E2P information. Developing methods that integrate E2P and other functional annotations with QTL/GWAS data in a more principled fashion may be useful.

Ultimately, knowing whether an epiQTL is also an eQTL or not would be tantamount to solving the "cis-regulatory code"[187], which is a mapping of DNA sequence to expression level of nearby genes that could predict when, where, and how much the gene should be

expressed. Unlike amino acid code that maps DNA sequence to protein sequence, "cis-regulatory code" is extremely poorly defined. This is challenging for several reasons. First, the sequence space of "cis-regulatory code" is huge, whereas the open chromatin sequences from a group of cell types is relatively small, and most accessibility status are shared across cell types. Second, "cis-regulatory code" dictates the binding and interaction of TFs, which can have a large number of potential combinations and can lead to competitive or cooperative relationships based on the exact interacting partners and the genome context. Third, many aspects jointly determine the effect of cis-regulatory elements on gene transcription, including but not restricted to genomic distance, chromatin folding, enhancer strength, promoter strength, and so on[187]. No single assay could generate the data modalities that are required to decipher the effect of these elements on transcription, and future studies not only need to generate more types of data, but they also need to come up with novel methods to integrate all the information from all the data.

To summarize, given all the advents in experimental and statistical methods in the past decade, understanding the biological mechanisms of disease GWAS variants remain very challenging. While the field has largely focused on expanding the number of samples and contexts in study design, I believe more attention could be paid to multimodal integration of data and to bridging the theories and methods between the fields of statistical genetics and biological experiments. These interdisciplinary approaches have the potential to reveal the complexities in gene regulation and offer a more complete picture of how genetics shape human complex phenotypes.

REFERENCES

[1] Axel Finckh, Benoît Gilbert, Bridget Hodkinson, Sang-Cheol Bae, Ranjeny Thomas, Kevin D Deane, Deshiré Alpizar-Rodriguez, and Kim Lauper. Global epidemiology of rheumatoid arthritis. *Nature reviews. Rheumatology*, September 2022.

[2] GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. *Lancet*, 396(10258):1204–1222, October 2020.

[3] N Risch and K Merikangas. The future of genetic studies of complex human diseases. *Science (New York, N.Y.)*, 273(5281):1516–1517, September 1996.

[4] Tony Burdett, Emma Hastings, Dani Welter, SPOT, EMBL-EBI, and NHGRI. GWAS catalog. `https://www.ebi.ac.uk/gwas/`. Accessed: 2023-10-22.

[5] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

[6] Joel Achenbach, Dan Keating, Laurie McGinley, Akilah Johnson, and Jahi Chikwendiu. Life expectancy in U.S. is falling amid surges in chronic illness. *The Washington Post*, October 2023.

[7] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[8] François Aguet, Kaur Alasoo, Yang I Li, Alexis Battle, Hae Kyung Im, Stephen B Montgomery, and Tuuli Lappalainen. Molecular quantitative trait loci. *Nature Reviews Methods Primers*, 3(1):1–22, January 2023.

[9] GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, September 2020.

[10] Yang I Li, Bryce van de Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, and Jonathan K Pritchard. RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, April 2016.

[11] Rachel E Gate, Christine S Cheng, Aviva P Aiden, Atsede Siba, Marcin Tabaka, Dmytro Lituiev, Ido Machol, M Grace Gordon, Meena Subramaniam, Muhammad Shamim, Kendrick L Hougen, Ivo Wortman, Su-Chen Huang, Neva C Durand, Ting

Feng, Philip L De Jager, Howard Y Chang, Erez Lieberman Aiden, Christophe Benoist, Michael A Beer, Chun J Ye, and Aviv Regev. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nature genetics*, 50(8):1140–1150, August 2018.

[12] Biao Zeng, Jaroslav Bendl, Chengyu Deng, Donghoon Lee, Ruth Misir, Sarah M Reach, Steven P Kleopoulos, Pavan Auluck, Stefano Marenco, David A Lewis, Vahram Haroutunian, Nadav Ahituv, John F Fullard, Gabriel E Hoffman, and Panos Roussos. Genetic regulation of cell-type specific chromatin accessibility shapes the etiology of brain diseases. *bioRxiv*, page 2023.03.02.530826, March 2023.

[13] Paola Benaglio, Jacklyn Newsome, Jee Yun Han, Joshua Chiou, Anthony Aylward, Sierra Corban, Michael Miller, Mei-Lin Okino, Jaspreet Kaur, Sebastian Preissl, David U Gorkin, and Kyle J Gaulton. Mapping genetic effects on cell type-specific chromatin accessibility and annotating complex immune trait variants using single nucleus ATAC-seq in peripheral blood. *PLoS genetics*, 19(6):e1010759, June 2023.

[14] Meritxell Oliva, Kathryn Demanelis, Yihao Lu, Meytal Chernoff, Farzana Jasmine, Habibul Ahsan, Muhammad G Kibriya, Lin S Chen, and Brandon L Pierce. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nature genetics*, pages 1–11, December 2022.

[15] Jingning Zhang, Diptavo Dutta, Anna Köttgen, Adrienne Tin, Pascal Schlosser, Morgan E Grams, Benjamin Harvey, Bing Yu, Eric Boerwinkle, Josef Coresh, and Nilanjan Chatterjee. Plasma proteome analyses in individuals of european and african ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nature genetics*, pages 1–10, May 2022.

[16] Benjamin B Sun, Joshua Chiou, Matthew Traylor, Christian Benner, Yi-Hsiang Hsu, Tom G Richardson, Praveen Surendran, Anubha Mahajan, Chloe Robins, Steven G Vasquez-Grinnell, Liping Hou, Erika M Kvikstad, Oliver S Burren, Jonathan Davitte, Kyle L Ferber, Christopher E Gillies, Åsa K Hedman, Sile Hu, Tinchi Lin, Rajesh Mikkilineni, Rion K Pendergrass, Corran Pickering, Bram Prins, Denis Baird, Chia-Yen Chen, Lucas D Ward, Aimee M Deaton, Samantha Welsh, Carissa M Willis, Nick Lehner, Matthias Arnold, Maria A Wörheide, Karsten Suhre, Gabi Kastenmüller, Anurag Sethi, Madeleine Cule, Anil Raj, Lucy Burkitt-Gray, Eugene Melamud, Mary Helen Black, Eric B Fauman, Joanna M M Howson, Hyun Min Kang, Mark I McCarthy, Paul Nioi, Slavé Petrovski, Robert A Scott, Erin N Smith, Sándor Szalma, Dawn M Waterworth, Lyndon J Mitnaul, Jose Szustakowski, Bradford W Gibson, Melissa R Miller, and Christopher D Whelan. Plasma proteomic associations with genetics and health in the UK biobank. *Nature*, pages 1–10, October 2023.

[17] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Seyhan Yazar,

Harm Brugge, Roy Oelen, Dylan H de Vries, Monique G P van der Wijst, Silva Kasela, Natalia Pervjakova, Isabel Alves, Marie-Julie Favé, Mawussé Agbessi, Mark W Christiansen, Rick Jansen, Ilkka Seppälä, Lin Tong, Alexander Teumer, Katharina Schramm, Gibran Hemani, Joost Verlouw, Hanieh Yaghootkar, Reyhan Sönmez Flitman, Andrew Brown, Viktorija Kukushkina, Anette Kalnapenkis, Sina Rüeger, Eleonora Porcu, Jaanika Kronberg, Johannes Kettunen, Bernett Lee, Futao Zhang, Ting Qi, Jose Alquicira Hernandez, Wibowo Arindrarto, Frank Beutner, BIOS Consortium, i2QTL Consortium, Julia Dmitrieva, Mahmoud Elansary, Benjamin P Fairfax, Michel Georges, Bastiaan T Heijmans, Alex W Hewitt, Mika Kähönen, Yungil Kim, Julian C Knight, Peter Kovacs, Knut Krohn, Shuang Li, Markus Loeffler, Urko M Marigorta, Hailang Mei, Yukihide Momozawa, Martina Müller-Nurasyid, Matthias Nauck, Michel G Nivard, Brenda W J H Penninx, Jonathan K Pritchard, Olli T Raitakari, Olaf Rotzschke, Eline P Slagboom, Coen D A Stehouwer, Michael Stumvoll, Patrick Sullivan, Peter A C 't Hoen, Joachim Thiery, Anke Tönjes, Jenny van Dongen, Maarten van Iterson, Jan H Veldink, Uwe Völker, Robert Warmerdam, Cisca Wijmenga, Morris Swertz, Anand Andiappan, Grant W Montgomery, Samuli Ripatti, Markus Perola, Zoltan Kutalik, Emmanouil Dermitzakis, Sven Bergmann, Timothy Frayling, Joyce van Meurs, Holger Prokisch, Habibul Ahsan, Brandon L Pierce, Terho Lehtimäki, Dorret I Boomsma, Bruce M Psaty, Sina A Gharib, Philip Awadalla, Lili Milani, Willem H Ouwehand, Kate Downes, Oliver Stegle, Alexis Battle, Peter M Visscher, Jian Yang, Markus Scholz, Joseph Powell, Greg Gibson, Tõnu Esko, and Lude Franke. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature genetics*, 53(9):1300–1310, September 2021.

[18] Lili Wang, Nikita Babushkin, Zhonghua Liu, and Xuanyao Liu. Trans-eQTL mapping in gene sets identifies network effects of genetic variants. *bioRxiv*, page 2022.11.11.516189, November 2022.

[19] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph, Peter J Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, September 2012.

[20] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics*, 6(4):e1000888, 2010.

[21] Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation be-

tween pairs of genetic association studies using summary statistics. *PLoS genetics*, 10(5):e1004383, May 2014.

[22] Eric R Gamazon, Ayellet V Segrè, Martijn van de Bunt, Xiaoquan Wen, Hualin S Xi, Farhad Hormozdiari, Halit Ongen, Anuar Konkashbaev, Eske M Derks, François Aguet, Jie Quan, GTEx Consortium, Dan L Nicolae, Eleazar Eskin, Manolis Kellis, Gad Getz, Mark I McCarthy, Emmanouil T Dermitzakis, Nancy J Cox, and Kristin G Ardlie. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature genetics*, 50(7):956–967, July 2018.

[23] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, Patrick F Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L Price, Päivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252, March 2016.

[24] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, and Jian Yang. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics*, 48(5):481–487, May 2016.

[25] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter A C 't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, September 2013.

[26] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, Rui Mei, and Others. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research*, 24(1):14–24, 2014.

[27] Sarah Kim-Hellmuth, François Aguet, Meritxell Oliva, Manuel Muñoz-Aguirre, Silva Kasela, Valentin Wucher, Stephane E Castel, Andrew R Hamel, Ana Viñuela,

Amy L Roberts, Serghei Mangul, Xiaoquan Wen, Gao Wang, Alvaro N Barbeira, Diego Garrido-Martín, Brian B Nadel, Yuxin Zou, Rodrigo Bonazzola, Jie Quan, Andrew Brown, Angel Martinez-Perez, José Manuel Soria, GTEx Consortium, Gad Getz, Emmanouil T Dermitzakis, Kerrin S Small, Matthew Stephens, Hualin S Xi, Hae Kyung Im, Roderic Guigó, Ayellet V Segrè, Barbara E Stranger, Kristin G Ardlie, and Tuuli Lappalainen. Cell type-specific genetic regulation of gene expression across human tissues. *Science*, 369(6509), September 2020.

[28] Lu Chen, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yan, Kousik Kundu, Simone Ecker, Avik Datta, David Richardson, Frances Burden, Daniel Mead, Alice L Mann, Jose Maria Fernandez, Sophia Rowlston, Steven P Wilder, Samantha Farrow, Xiaojian Shao, John J Lambourne, Adriana Redensek, Cornelis A Albers, Vyacheslav Amstislavskiy, Sofie Ashford, Kim Berentsen, Lorenzo Bomba, Guillaume Bourque, David Bujold, Stephan Busche, Maxime Caron, Shu-Huang Chen, Warren Cheung, Oliver Delaneau, Emmanouil T Dermitzakis, Heather Elding, Irina Colgiu, Frederik O Bagger, Paul Flicek, Ehsan Habibi, Valentina Iotchkova, Eva Janssen-Megens, Bowon Kim, Hans Lehrach, Ernesto Lowy, Amit Mandoli, Filomena Matarese, Matthew T Maurano, John A Morris, Vera Pancaldi, Farzin Pourfarzad, Karola Rehnstrom, Augusto Rendon, Thomas Risch, Nilofar Sharifi, Marie-Michelle Simon, Marc Sultan, Alfonso Valencia, Klaudia Walter, Shuang-Yin Wang, Mattia Frontini, Stylianos E Antonarakis, Laura Clarke, Marie-Laure Yaspo, Stephan Beck, Roderic Guigo, Daniel Rico, Joost H A Martens, Willem H Ouwehand, Taco W Kuijpers, Dirk S Paul, Hendrik G Stunnenberg, Oliver Stegle, Kate Downes, Tomi Pastinen, and Nicole Soranzo. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, 167(5):1398–1414.e24, November 2016.

[29] Benjamin J Schmiedel, Divya Singh, Ariel Madrigal, Alan G Valdovino-Gonzalez, Brandie M White, Jose Zapardiel-Gonzalo, Brendan Ha, Gokmen Altay, Jason A Greenbaum, Graham McVicker, Grégory Seumois, Anjana Rao, Mitchell Kronenberg, Bjoern Peters, and Pandurangan Vijayanand. Impact of genetic polymorphisms on human immune cell gene expression. *Cell*, 175(6):1701–1715.e16, November 2018.

[30] Mineto Ota, Yasuo Nagafuchi, Hiroaki Hatano, Kazuyoshi Ishigaki, Chikashi Terao, Yusuke Takeshima, Haruyuki Yanaoka, Satomi Kobayashi, Mai Okubo, Harumi Shirai, Yusuke Sugimori, Junko Maeda, Masahiro Nakano, Saeko Yamada, Ryochi Yoshida, Haruka Tsuchiya, Yumi Tsuchida, Shuji Akizuki, Hajime Yoshifuji, Koichiro Ohmura, Tsuneyo Mimori, Ken Yoshida, Daitaro Kurosaka, Masato Okada, Keigo Setoguchi, Hiroshi Kaneko, Nobuhiro Ban, Nami Yabuki, Kosuke Matsuki, Hironori Mutoh, Sohei Oyama, Makoto Okazaki, Hiroyuki Tsunoda, Yukiko Iwasaki, Shuji Sumitomo, Hirofumi Shoda, Yuta Kochi, Yukinori Okada, Kazuhiko Yamamoto, Tomohisa Okamura, and Keishi Fujio. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell*, 184(11):3006–3021.e17, May 2021.

[31] Sung Chun, Alexandra Casparino, Nikolaos A Patsopoulos, Damien C Croteau-Chonka, Benjamin A Raby, Philip L De Jager, Shamil R Sunyaev, and Chris Cotsapas. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nature genetics*, 49(4):600–605, April 2017.

[32] Douglas W Yao, Luke J O'Connor, Alkes L Price, and Alexander Gusev. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature genetics*, 52(6):626–633, June 2020.

[33] Zepeng Mu, Wei Wei, Benjamin Fair, Jinlin Miao, Ping Zhu, and Yang I Li. The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome biology*, 22(1):122, April 2021.

[34] Hakhamanesh Mostafavi, Jeffrey P Spence, Sahin Naqvi, and Jonathan K Pritchard. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nature genetics*, pages 1–10, October 2023.

[35] Alan P Boyle, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, January 2008.

[36] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218, December 2013.

[37] Xinchen Wang and David B Goldstein. Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease. *American journal of human genetics*, 106(2):215–233, February 2020.

[38] Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas R P Taylor, Linda Clarke, Katia Maccora, Christine Chen, Anthony L Cook, Chun Jimmie Ye, Kirsten A Fairfax, Alex W Hewitt, and Joseph E Powell. Single-cell eQTL mapping identifies cell typespecific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041, 2022.

[39] Aparna Nathan, Samira Asgari, Kazuyoshi Ishigaki, Cristian Valencia, Tiffany Amariuta, Yang Luo, Jessica I Beynor, Yuriy Baglaenko, Sara Suliman, Alkes L Price, Leonid Lecca, Megan B Murray, D Branch Moody, and Soumya Raychaudhuri. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature*, page 2021.07.29.454316, May 2022.

[40] Anna S E Cuomo, Daniel D Seaton, Davis J McCarthy, Iker Martinez, Marc Jan Bonder, Jose Garcia-Bernardo, Shradha Amatya, Pedro Madrigal, Abigail Isaacson, Florian Buettner, Andrew Knights, Kedar Nath Natarajan, HipSci Consortium, Ludovic Vallier, John C Marioni, Mariya Chhatriwala, and Oliver Stegle. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature communications*, 11(1):810, February 2020.

[41] Zijian Cheng, Thuy Do, Kulveer Mankia, Josephine Meade, Laura Hunt, Val Clerehugh, Alastair Speirs, Aradhna Tugnait, Paul Emery, and Deirdre Devine. Dysbiosis in the oral microbiomes of anti-CCP positive individuals at risk of developing rheumatoid arthritis. *Annals of the rheumatic diseases*, 80(2):162–168, February 2021.

[42] Stefano Alivernini, Gary S Firestein, and Iain B McInnes. The pathogenesis of rheumatoid arthritis. *Immunity*, 55(12):2255–2270, December 2022.

[43] A J MacGregor, H Snieder, A S Rigby, M Koskenvuo, J Kaprio, K Aho, and A J Silman. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis and rheumatism*, 43(1):30–37, January 2000.

[44] Yukinori Okada, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, Koichiro Ohmura, Akari Suzuki, Shinji Yoshida, Robert R Graham, Arun Manoharan, Ward Ortmann, Tushar Bhangale, Joshua C Denny, Robert J Carroll, Anne E Eyler, Jeffrey D Greenberg, Joel M Kremer, Dimitrios A Pappas, Lei Jiang, Jian Yin, Lingying Ye, Ding-Feng Su, Jian Yang, Gang Xie, Ed Keystone, Harm-Jan Westra, Tõnu Esko, Andres Metspalu, Xuezhong Zhou, Namrata Gupta, Daniel Mirel, Eli A Stahl, Dorothée Diogo, Jing Cui, Katherine Liao, Michael H Guo, Keiko Myouzen, Takahisa Kawaguchi, Marieke J H Coenen, Piet L C M van Riel, Mart A F J van de Laar, Henk-Jan Guchelaar, Tom W J Huizinga, Philippe Dieudé, Xavier Mariette, S Louis Bridges, Jr, Alexandra Zhernakova, Rene E M Toes, Paul P Tak, Corinne Miceli-Richard, So-Young Bang, Hye-Soon Lee, Javier Martin, Miguel A Gonzalez-Gay, Luis Rodriguez-Rodriguez, Solbritt Rantapää-Dahlqvist, Lisbeth Arlestig, Hyon K Choi, Yoichiro Kamatani, Pilar Galan, Mark Lathrop, RACI consortium, GARNET consortium, Steve Eyre, John Bowes, Anne Barton, Niek de Vries, Larry W Moreland, Lindsey A Criswell, Elizabeth W Karlson, Atsuo Taniguchi, Ryo Yamada, Michiaki Kubo, Jun S Liu, Sang-Cheol Bae, Jane Worthington, Leonid Padyukov, Lars Klareskog, Peter K Gregersen, Soumya Raychaudhuri, Barbara E Stranger, Philip L De Jager, Lude Franke, Peter M Visscher, Matthew A Brown, Hisashi Yamanaka, Tsuneyo Mimori, Atsushi Takahashi, Huji Xu, Timothy W Behrens, Katherine A Siminovitch, Shigeki Momohara, Fumihiko Matsuda, Kazuhiko Yamamoto, and Robert M Plenge. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, February 2014.

[45] Kazuyoshi Ishigaki, Saori Sakaue, Chikashi Terao, Yang Luo, Kyuto Sonehara, Kensuke Yamaguchi, Tiffany Amariuta, Chun Lai Too, Vincent A Laufer, Ian C Scott,

Sebastien Viatte, Meiko Takahashi, Koichiro Ohmura, Akira Murasawa, Motomu Hashimoto, Hiromu Ito, Mohammed Hammoudeh, Samar Al Emadi, Basel K Masri, Hussein Halabi, Humeira Badsha, Imad W Uthman, Xin Wu, Li Lin, Ting Li, Darren Plant, Anne Barton, Gisela Orozco, Suzanne M M Verstappen, John Bowes, Alexander J MacGregor, Suguru Honda, Masaru Koido, Kohei Tomizuka, Yoichiro Kamatani, Hiroaki Tanaka, Eiichi Tanaka, Akari Suzuki, Yuichi Maeda, Kenichi Yamamoto, Satoru Miyawaki, Gang Xie, Jinyi Zhang, Christopher I Amos, Edward Keystone, Gertjan Wolbink, Irene Van der Horst-Bruinsma, Jing Cui, Katherine P Liao, Robert J Carroll, Hye-Soon Lee, So-Young Bang, Katherine A Siminovitch, Niek de Vries, Lars Alfredsson, Solbritt Rantapää-Dahlqvist, Elizabeth W Karlson, Sang-Cheol Bae, Robert P Kimberly, Jeffrey C Edberg, Xavier Mariette, Tom Huizinga, Philippe Dieudé, Matthias Schneider, Martin Kerick, Joshua C Denny, BioBank Japan Project, Koichi Matsuda, Keitaro Matsuo, Tsuneyo Mimori, Fumihiko Matsuda, Keishi Fujio, Yoshiya Tanaka, Atsushi Kumanogoh, Matthew Traylor, Cathryn M Lewis, Stephen Eyre, Huji Xu, Richa Saxena, Thurayya Arayssi, Yuta Kochi, Katsunori Ikari, Masayoshi Harigai, Peter K Gregersen, Kazuhiko Yamamoto, S Louis Bridges, Jr, Leonid Padyukov, Javier Martin, Lars Klareskog, Yukinori Okada, and Soumya Raychaudhuri. Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nature Genetics*, pages 1–12, November 2022.

[46] Xinli Hu, Hyun Kim, Towfique Raj, Patrick J Brennan, Gosia Trynka, Nikola Teslovich, Kamil Slowikowski, Wei-Min Chen, Suna Onengut, Clare Baecher-Allan, and Others. Regulation of gene expression in autoimmune disease loci and the genetic basis of proliferation in CD4+ effector memory T cells. *PLoS genetics*, 10(6):e1004404, 2014.

[47] Gosia Trynka, Cynthia Sandor, Buhm Han, Han Xu, Barbara E Stranger, X Shirley Liu, and Soumya Raychaudhuri. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*, 45(2):124–130, February 2013.

[48] Farbod Ghobadinezhad, Nasim Ebrahimi, Fatemeh Mozaffari, Neda Moradi, Sheida Beiranvand, Mehran Pournazari, Fatemeh Rezaei-Tazangi, Roya Khorram, Maral Afshinpour, Rob A Robino, Amir Reza Aref, and Leonardo M R Ferreira. The emerging role of regulatory cell-based therapy in autoimmune disease. *Frontiers in immunology*, 13, 2022.

[49] Mélika Ben Ahmed, Nadia Belhadj Hmida, Nicolette Moes, Sophie Buyse, Maha Abdeladhim, Hechmi Louzir, and Nadine Cerf-Bensussan. IL-15 renders conventional lymphocytes resistant to suppressive functions of regulatory T cells through activation of the phosphatidylinositol 3-kinase pathway. *Journal of immunology*, 182(11):6763–6770, June 2009.

[50] Ilaria Peluso, Massimo Claudio Fantini, Daniele Fina, Roberta Caruso, Monica Boirivant, Thomas T MacDonald, Francesco Pallone, and Giovanni Monteleone. IL-

21 counteracts the regulatory T cell-mediated suppression of human CD4+ T lymphocytes. *Journal of immunology*, 178(2):732–739, January 2007.

[51] A Helena Jonsson, Fan Zhang, Garrett Dunlap, Emma Gomez-Rivas, Gerald F M Watts, Heather J Faust, Karishma Vijay Rupani, Joseph R Mears, Nida Meednu, Runci Wang, Gregory Keras, Jonathan S Coblyn, Elena M Massarotti, Derrick J Todd, Jennifer H Anolik, Andrew McDavid, Accelerating Medicines Partnership RA/SLE Network, Kevin Wei, Deepak A Rao, Soumya Raychaudhuri, and Michael B Brenner. Granzyme K^+ CD8 T cells form a core population in inflamed human tissue. *Science translational medicine*, 14(649):eabo0686, June 2022.

[52] Jae-Seung Moon, Shady Younis, Nitya S Ramadoss, Radhika Iyer, Khushboo Sheth, Orr Sharpe, Navin L Rao, Stephane Becart, Julie A Carman, Eddie A James, Jane H Buckner, Kevin D Deane, V Michael Holers, Susan M Goodman, Laura T Donlin, Mark M Davis, and William H Robinson. Cytotoxic CD8+ T cells target citrullinated antigens in rheumatoid arthritis. *Nature communications*, 14(1):319, January 2023.

[53] Kaur Alasoo, Julia Rodrigues, Subhankar Mukhopadhyay, Andrew J Knights, Alice L Mann, Kousik Kundu, HIPSCI Consortium, Christine Hale, Gordon Dougan, and Daniel J Gaffney. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature genetics*, 50(3):424–431, March 2018.

[54] Yohann Nédélec, Joaquín Sanz, Golshid Baharian, Zachary A Szpiech, Alain Pacis, Anne Dumaine, Jean-Christophe Grenier, Andrew Freiman, Aaron J Sams, Steven Hebert, Ariane Pagé Sabourin, Francesca Luca, Ran Blekhman, Ryan D Hernandez, Roger Pique-Regi, Jenny Tung, Vania Yotova, and Luis B Barreiro. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell*, 167(3):657–669.e21, October 2016.

[55] Mengjie Chen, Qi Zhan, Zepeng Mu, Lili Wang, Zhaohui Zheng, Jinlin Miao, Ping Zhu, and Yang I Li. Alignment of single-cell RNA-seq samples without overcorrection using kernel density matching. *Genome research*, 31(4):698–712, April 2021.

[56] Zepeng Mu, Wei Wei, Benjamin Fair, Jinlin Miao, Ping Zhu, and Yang I Li. The impact of cell-type and context-dependent regulatory variants on human immune traits, 2021.

[57] Sarah M Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature genetics*, 51(1):187–195, January 2019.

[58] Yang I Li, David A Knowles, Jack Humphrey, Alvaro N Barbeira, Scott P Dickinson, Hae Kyung Im, and Jonathan K Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nature genetics*, 50(1):151–158, January 2018.

[59] Katrina M de Lange, Loukas Moutsianas, James C Lee, Christopher A Lamb, Yang Luo, Nicholas A Kennedy, Luke Jostins, Daniel L Rice, Javier Gutierrez-Achury, Sun-Gou Ji, Graham Heap, Elaine R Nimmo, Cathryn Edwards, Paul Henderson, Craig Mowat, Jeremy Sanderson, Jack Satsangi, Alison Simmons, David C Wilson, Mark Tremelling, Ailsa Hart, Christopher G Mathew, William G Newman, Miles Parkes, Charlie W Lees, Holm Uhlig, Chris Hawkey, Natalie J Prescott, Tariq Ahmad, John C Mansfield, Carl A Anderson, and Jeffrey C Barrett. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics*, 49(2):256–261, February 2017.

[60] Shin Maeda, Li-Chung Hsu, Hongjun Liu, Laurie A Bankston, Mitsutoshi Iimura, Martin F Kagnoff, Lars Eckmann, and Michael Karin. Nod2 mutation in crohn's disease potentiates NF-κB activity and IL-1SS processing. *Science*, 307(5710):734–738, 2005.

[61] Subrata Ghosh, Eran Goldin, Fiona H Gordon, Helmut A Malchow, Jørgen Rask-Madsen, Paul Rutgeerts, Petr Vyhnálek, Zdena Zádorová, Tanya Palmer, and Stephen Donoghue. Natalizumab for active crohn's disease. *New England Journal of Medicine*, 348(1):24–32, 2003.

[62] Jimmy Z Liu, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, James C Lee, Luke Jostins, Tejas Shah, Shifteh Abedian, Jae Hee Cheon, Judy Cho, Naser E Dayani, Lude Franke, Yuta Fuyuno, Ailsa Hart, Ramesh C Juyal, Garima Juyal, Won Ho Kim, Andrew P Morris, Hossein Poustchi, William G Newman, Vandana Midha, Timothy R Orchard, Homayon Vahedi, Ajit Sood, Joseph Y Sung, Reza Malekzadeh, Harm-Jan Westra, Keiko Yamazaki, Suk-Kyun Yang, International Multiple Sclerosis Genetics Consortium, International IBD Genetics Consortium, Jeffrey C Barrett, Behrooz Z Alizadeh, Miles Parkes, Thelma Bk, Mark J Daly, Michiaki Kubo, Carl A Anderson, and Rinse K Weersma. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics*, 47(9):979–986, September 2015.

[63] Donald B Bloch, M Suzanne, Pavel Guigaouri, Andrew Filippov, and Kenneth D Bloch. Identification and characterization of a leukocyte-specific component of the nuclear body. *Journal of Biological Chemistry*, 271(46):29198–29204, 1996.

[64] Stuti Mehta, D Alexander Cronkite, Megha Basavappa, Tahnee L Saunders, Fatemeh Adiliaghdam, Hajera Amatullah, Sara A Morrison, Jose D Pagan, Robert M Anthony, Pierre Tonnerre, Georg M Lauer, James C Lee, Sreehaas Digumarthi, Lorena Pantano, Shannan J Ho Sui, Fei Ji, Ruslan Sadreyev, Chan Zhou, Alan C Mullen, Vinod Kumar, Yang Li, Cisca Wijmenga, Ramnik J Xavier, Terry K Means, and Kate L Jeffrey. Maintenance of macrophage transcriptional programs and intestinal homeostasis by epigenetic reader SP140. *Science immunology*, 2(9), March 2017.

[65] Mohamad Karaky, María Fedetz, Victor Potenciano, Eduardo Andrés-León, Anna Esteve Codina, Cristina Barrionuevo, Antonio Alcina, and Fuencisla Matesanz. SP140 regulates the expression of immune-related genes associated with multiple sclerosis and other autoimmune diseases by NF-κ B inhibition. *Human molecular genetics*, 27(23):4012–4023, 2018.

[66] Fuencisla Matesanz, Victor Potenciano, Maria Fedetz, Priscila Ramos-Mozo, María del Mar Abad-Grau, Mohamad Karaky, Cristina Barrionuevo, Guillermo Izquierdo, Juan Luis Ruiz-Peña, María Isabel García-Sánchez, and Others. A functional variant that affects exon-skipping and protein expression of SP140 as genetic mechanism predisposing to multiple sclerosis. *Human molecular genetics*, 24(19):5619–5627, 2015.

[67] Frank L Heppner, Richard M Ransohoff, and Burkhard Becher. Immune attack: the role of inflammation in alzheimer disease. *Nature Reviews Neuroscience*, 16(6):358–372, 2015.

[68] Xinli Hu, Hyun Kim, Eli Stahl, Robert Plenge, Mark Daly, and Soumya Raychaudhuri. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *The American Journal of Human Genetics*, 89(4):496–506, 2011.

[69] Briana E Mittleman, Sebastian Pott, Shane Warland, Tony Zeng, Zepeng Mu, Mayher Kaur, Yoav Gilad, and Yang Li. Alternative polyadenylation mediates genetic regulation of gene expression. *eLife*, 9, June 2020.

[70] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, and Others. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[71] Diego Calderon, Michelle L T Nguyen, Anja Mezger, Arwa Kathiria, Fabian Müller, Vinh Nguyen, Ninnia Lescano, Beijing Wu, John Trombetta, Jessica V Ribado, David A Knowles, Ziyue Gao, Franziska Blaeschke, Audrey V Parent, Trevor D Burt, Mark S Anderson, Lindsey A Criswell, William J Greenleaf, Alexander Marson, and Jonathan K Pritchard. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nature genetics*, 51(10):1494–1505, October 2019.

[72] Blagoje Soskic, Eddie Cano-Gamez, Deborah J Smyth, Wendy C Rowan, Nikolina Nakic, Jorge Esparza-Gordillo, Lara Bossini-Castillo, David F Tough, Christopher G C Larminie, Paola G Bronson, David Willé, and Gosia Trynka. Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nature genetics*, 51(10):1486–1493, October 2019.

[73] Philip Courtney and Michael Doherty. Joint aspiration and injection and synovial fluid analysis. *Best Practice & Research Clinical Rheumatology*, 27(2):137–169, 2013.

[74] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R Day, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R B Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J Daly, Nick Patterson, Benjamin M Neale, and Alkes L Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228–1235, November 2015.

[75] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357, 2012.

[76] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of ChIP-seq (MACS). *Genome biology*, 9(9):R137, September 2008.

[77] Ruteja A Barve, Marc D Zack, David Weiss, Ruo-Hua Song, David Beidler, and Richard D Head. Transcriptional profiling and pathway analysis of CSF-1 and IL-34 effects on human monocyte differentiation. *Cytokine*, 63(1):10–17, 2013.

[78] S Perrier, C Coussediere, J J Dubost, E Albuisson, and B Sauvezie. IL-1 receptor antagonist (IL-1RA) gene polymorphism in sjogren's syndrome and rheumatoid arthritis. *Clinical immunology and immunopathology*, 87(3):309–313, 1998.

[79] Silvia Lopa, Maarten J C Leijs, Matteo Moretti, Erik Lubberts, Gjvm van Osch, and Y M Bastiaansen-Jenniskens. Arthritic and non-arthritic synovial fluids modulate IL10 and IL1RA gene expression in differentially activated primary human monocytes. *Osteoarthritis and cartilage*, 23(11):1853–1857, 2015.

[80] Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv*, page 501114, June 2020.

[81] Christopher N Foley, James R Staley, Philip G Breen, Benjamin B Sun, Paul D W Kirk, Stephen Burgess, and Joanna M M Howson. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature communications*, 12(1):764, February 2021.

[82] Haruka Tsuchiya, Mineto Ota, Shuji Sumitomo, Kazuyoshi Ishigaki, Akari Suzuki, Toyonori Sakata, Yumi Tsuchida, Hiroshi Inui, Jun Hirose, Yuta Kochi, and Others. Synovial fibroblasts contribute to the genetic risk of rheumatoid arthritis through the synergistic action of cytokines. *bioRxiv*, page 861781, 2019.

[83] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, David Schlessinger,

Dwight Stambolian, Po-Ru Loh, William G Iacono, Anand Swaroop, Laura J Scott, Francesco Cucca, Florian Kronenberg, Michael Boehnke, Gonçalo R Abecasis, and Christian Fuchsberger. Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284–1287, October 2016.

[84] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W De Bakker, Mark J Daly, and Others. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[85] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904, 2006.

[86] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016.

[87] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, and Others. GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773, 2019.

[88] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[89] Bryce van de Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11):1061–1063, November 2015.

[90] Yang-Yang Feng, Avinash Ramu, Kelsy C Cotto, Zachary L Skidmore, Jason Kunisaki, Donald F Conrad, Yiing Lin, William Chapman, Ravindra Uppaulri, Ramaswamy Govindan, and Others. RegTools: Integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer. *bioRxiv*, page 436634, 2018.

[91] Olivier Delaneau, Halit Ongen, Andrew A Brown, Alexandre Fort, Nikolaos I Panousis, and Emmanouil T Dermitzakis. A complete tool set for molecular QTL discovery and analysis. *Nature communications*, 8:15452, May 2017.

[92] John D Storey. False discovery rate. *International encyclopedia of statistical science*, 1:504–508, 2011.

[93] Po-Ru Loh, Gleb Kichaev, Steven Gazal, Armin P Schoech, and Alkes L Price. Mixed-model association for biobank-scale datasets. *Nature genetics*, 50(7):906–908, July 2018.

[94] Manuel A Ferreira, Judith M Vonk, Hansjörg Baurecht, Ingo Marenholz, Chao Tian, Joshua D Hoffman, Quinta Helmer, Annika Tillander, Vilhelmina Ullemar, Jenny van Dongen, Yi Lu, Franz Rüschendorf, Jorge Esparza-Gordillo, Chris W Medway, Edward Mountjoy, Kimberley Burrows, Oliver Hummel, Sarah Grosche, Ben M Brumpton, John S Witte, Jouke-Jan Hottenga, Gonneke Willemsen, Jie Zheng, Elke Rodríguez, Melanie Hotze, Andre Franke, Joana A Revez, Jonathan Beesley, Melanie C Matheson, Shyamali C Dharmage, Lisa M Bain, Lars G Fritsche, Maiken E Gabrielsen, Brunilda Balliu, 23andMe Research Team, AAGC collaborators, BIOS consortium, LifeLines Cohort Study, Jonas B Nielsen, Wei Zhou, Kristian Hveem, Arnulf Langhammer, Oddgeir L Holmen, Mari Løset, Gonçalo R Abecasis, Cristen J Willer, Andreas Arnold, Georg Homuth, Carsten O Schmidt, Philip J Thompson, Nicholas G Martin, David L Duffy, Natalija Novak, Holger Schulz, Stefan Karrasch, Christian Gieger, Konstantin Strauch, Ronald B Melles, David A Hinds, Norbert Hübner, Stephan Weidinger, Patrik K E Magnusson, Rick Jansen, Eric Jorgenson, Young-Ae Lee, Dorret I Boomsma, Catarina Almqvist, Robert Karlsson, Gerard H Koppelman, and Lavinia Paternoster. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nature genetics*, 49(12):1752–1757, December 2017.

[95] Lavinia Paternoster, Marie Standl, Johannes Waage, Hansjörg Baurecht, Melanie Hotze, David P Strachan, John A Curtin, Klaus Bønnelykke, Chao Tian, Atsushi Takahashi, Jorge Esparza-Gordillo, Alexessander Couto Alves, Jacob P Thyssen, Herman T den Dekker, Manuel A Ferreira, Elisabeth Altmaier, Patrick Ma Sleiman, Feng Li Xiao, Juan R Gonzalez, Ingo Marenholz, Birgit Kalb, Maria Pino Yanes, Cheng-Jian Xu, Lisbeth Carstensen, Maria M Groen-Blokhuis, Cristina Venturini, Craig E Pennell, Sheila J Barton, Albert M Levin, Ivan Curjuric, Mariona Bustamante, Eskil Kreiner-Møller, Gabrielle A Lockett, Jonas Bacelis, Supinda Bunyavanich, Rachel A Myers, Anja Matanovic, Ashish Kumar, Joyce Y Tung, Tomomitsu Hirota, Michiaki Kubo, Wendy L McArdle, A J Henderson, John P Kemp, Jie Zheng, George Davey Smith, Franz Rüschendorf, Anja Bauerfeind, Min Ae Lee-Kirsch, Andreas Arnold, Georg Homuth, Carsten O Schmidt, Elisabeth Mangold, Sven Cichon, Thomas Keil, Elke Rodríguez, Annette Peters, Andre Franke, Wolfgang Lieb, Natalija Novak, Regina Fölster-Holst, Momoko Horikoshi, Juha Pekkanen, Sylvain Sebert, Lise L Husemoen, Niels Grarup, Johan C de Jongste, Fernando Rivadeneira, Albert Hofman, Vincent Wv Jaddoe, Suzanne Gma Pasmans, Niels J Elbert, André G Uitterlinden, Guy B Marks, Philip J Thompson, Melanie C Matheson, Colin F Robertson, Australian Asthma Genetics Consortium (AAGC), Janina S Ried, Jin Li, Xian Bo Zuo, Xiao Dong Zheng, Xian Yong Yin, Liang Dan Sun, Maeve A McAleer, Grainne M O'Regan, Caoimhe Mr Fahy, Linda E Campbell, Milan Macek, Michael Kurek, Donglei Hu, Celeste Eng, Dirkje S Postma, Bjarke Feenstra, Frank Geller, Jouke Jan Hottenga, Christel M Middeldorp, Pirro Hysi, Veronique Bataille, Tim Spector, Carla Mt Tiesler, Elisabeth Thiering, Badri Pahukasahasram, James J Yang, Medea Imboden, Scott Huntsman, Natàlia Vilor-Tejedor, Caroline L Relton, Ronny Myhre, Wenche Nystad, Adnan Custovic, Scott T Weiss, Deborah A Meyers, Cilla Söderhäll, Erik

Melén, Carole Ober, Benjamin A Raby, Angela Simpson, Bo Jacobsson, John W Holloway, Hans Bisgaard, Jordi Sunyer, Nicole M Probst Hensch, L Keoki Williams, Keith M Godfrey, Carol A Wang, Dorret I Boomsma, Mads Melbye, Gerard H Koppelman, Deborah Jarvis, Wh Irwin McLean, Alan D Irvine, Xue Jun Zhang, Hakon Hakonarson, Christian Gieger, Esteban G Burchard, Nicholas G Martin, Liesbeth Duijts, Allan Linneberg, Marjo-Riitta Jarvelin, Markus M Noethen, Susanne Lau, Norbert Hübner, Young-Ae Lee, Mayumi Tamari, David A Hinds, Daniel Glass, Sara J Brown, Joachim Heinrich, David M Evans, and Stephan Weidinger. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature genetics*, 47(12):1449–1456, December 2015.

[96] Manuel A R Ferreira, Riddhima Mathur, Judith M Vonk, Agnieszka Szwajda, Ben Brumpton, Raquel Granell, Bronwyn K Brew, Vilhelmina Ullemar, Yi Lu, Yunxuan Jiang, and Others. Genetic architectures of childhood-and adult-onset asthma are partly distinct. *The American Journal of Human Genetics*, 104(4):665–684, 2019.

[97] Florence Demenais, Patricia Margaritte-Jeannin, Kathleen C Barnes, William O C Cookson, Janine Altmüller, Wei Ang, R Graham Barr, Terri H Beaty, Allan B Becker, John Beilby, Hans Bisgaard, Unnur Steina Bjornsdottir, Eugene Bleecker, Klaus Bønnelykke, Dorret I Boomsma, Emmanuelle Bouzigon, Christopher E Brightling, Myriam Brossard, Guy G Brusselle, Esteban Burchard, Kristin M Burkart, Andrew Bush, Moira Chan-Yeung, Kian Fan Chung, Alexessander Couto Alves, John A Curtin, Adnan Custovic, Denise Daley, Johan C de Jongste, Blanca E Del-Rio-Navarro, Kathleen M Donohue, Liesbeth Duijts, Celeste Eng, Johan G Eriksson, Martin Farrall, Yuliya Fedorova, Bjarke Feenstra, Manuel A Ferreira, Australian Asthma Genetics Consortium (AAGC) collaborators, Maxim B Freidin, Zofia Gajdos, Jim Gauderman, Ulrike Gehring, Frank Geller, Jon Genuneit, Sina A Gharib, Frank Gilliland, Raquel Granell, Penelope E Graves, Daniel F Gudbjartsson, Tari Haahtela, Susan R Heckbert, Dick Heederik, Joachim Heinrich, Markku Heliövaara, John Henderson, Blanca E Himes, Hiroshi Hirose, Joel N Hirschhorn, Albert Hofman, Patrick Holt, Jouke Hottenga, Thomas J Hudson, Jennie Hui, Medea Imboden, Vladimir Ivanov, Vincent W V Jaddoe, Alan James, Christer Janson, Marjo-Riitta Jarvelin, Deborah Jarvis, Graham Jones, Ingileif Jonsdottir, Pekka Jousilahti, Michael Kabesch, Mika Kähönen, David B Kantor, Alexandra S Karunas, Elza Khusnutdinova, Gerard H Koppelman, Anita L Kozyrskyj, Eskil Kreiner, Michiaki Kubo, Rajesh Kumar, Ashish Kumar, Mikko Kuokkanen, Lies Lahousse, Tarja Laitinen, Catherine Laprise, Mark Lathrop, Susanne Lau, Young-Ae Lee, Terho Lehtimäki, Sébastien Letort, Albert M Levin, Guo Li, Liming Liang, Laura R Loehr, Stephanie J London, Daan W Loth, Ani Manichaikul, Ingo Marenholz, Fernando J Martinez, Melanie C Matheson, Rasika A Mathias, Kenji Matsumoto, Hamdi Mbarek, Wendy L McArdle, Mads Melbye, Erik Melén, Deborah Meyers, Sven Michel, Hamida Mohamdi, Arthur W Musk, Rachel A Myers, Maartje A E Nieuwenhuis, Emiko Noguchi, George T O'Connor, Ludmila M Ogorodova, Cameron D Palmer, Aarno Palotie, Julie E Park, Craig E Pennell, Göran Pershagen, Alexey Polonikov, Dirkje S Postma, Nicole Probst-Hensch, Valery P

Puzyrev, Benjamin A Raby, Olli T Raitakari, Adaikalavan Ramasamy, Stephen S Rich, Colin F Robertson, Isabelle Romieu, Muhammad T Salam, Veikko Salomaa, Vivi Schlünssen, Robert Scott, Polina A Selivanova, Torben Sigsgaard, Angela Simpson, Valérie Siroux, Lewis J Smith, Maria Solodilova, Marie Standl, Kari Stefansson, David P Strachan, Bruno H Stricker, Atsushi Takahashi, Philip J Thompson, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Carla M T Tiesler, Dara G Torgerson, Tatsuhiko Tsunoda, André G Uitterlinden, Ralf J P van der Valk, Amaury Vaysse, Sailaja Vedantam, Andrea von Berg, Erika von Mutius, Judith M Vonk, Johannes Waage, Nick J Wareham, Scott T Weiss, Wendy B White, Magnus Wickman, Elisabeth Widén, Gonneke Willemsen, L Keoki Williams, Inge M Wouters, James J Yang, Jing Hua Zhao, Miriam F Moffatt, Carole Ober, and Dan L Nicolae. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature genetics*, 50(1):42–53, January 2018.

[98] David L Morris, Yujun Sheng, Yan Zhang, Yong-Fei Wang, Zhengwei Zhu, Philip Tombleson, Lingyan Chen, Deborah S Cunninghame Graham, James Bentham, Amy L Roberts, Ruoyan Chen, Xianbo Zuo, Tingyou Wang, Leilei Wen, Chao Yang, Lu Liu, Lulu Yang, Feng Li, Yuanbo Huang, Xianyong Yin, Sen Yang, Lars Rönnblom, Barbara G Fürnrohr, Reinhard E Voll, Georg Schett, Nathalie Costedoat-Chalumeau, Patrick M Gaffney, Yu Lung Lau, Xuejun Zhang, Wanling Yang, Yong Cui, and Timothy J Vyse. Genome-wide association meta-analysis in chinese and european individuals identifies ten new loci associated with systemic lupus erythematosus. *Nature genetics*, 48(8):940–946, August 2016.

[99] Stephen Sawcer and Maria Ban. Multiple sclerosis genomic map implicates peripheral immune cells & microglia in susceptibility. 2019.

[100] William J Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A Kostadima, John J Lambourne, Suthesh Sivapalaratnam, Kate Downes, Kousik Kundu, Lorenzo Bomba, Kim Berentsen, John R Bradley, Louise C Daugherty, Olivier Delaneau, Kathleen Freson, Stephen F Garner, Luigi Grassi, Jose Guerrero, Matthias Haimel, Eva M Janssen-Megens, Anita Kaan, Mihir Kamat, Bowon Kim, Amit Mandoli, Jonathan Marchini, Joost H A Martens, Stuart Meacham, Karyn Megy, Jared O'Connell, Romina Petersen, Nilofar Sharifi, Simon M Sheard, James R Staley, Salih Tuna, Martijn van der Ent, Klaudia Walter, Shuang-Yin Wang, Eleanor Wheeler, Steven P Wilder, Valentina Iotchkova, Carmel Moore, Jennifer Sambrook, Hendrik G Stunnenberg, Emanuele Di Angelantonio, Stephen Kaptoge, Taco W Kuijpers, Enrique Carrillo-de Santa-Pau, David Juan, Daniel Rico, Alfonso Valencia, Lu Chen, Bing Ge, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yang, Roderic Guigo, Stephan Beck, Dirk S Paul, Tomi Pastinen, David Bujold, Guillaume Bourque, Mattia Frontini, John Danesh, David J Roberts, Willem H Ouwehand, Adam S Butterworth, and Nicole Soranzo. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429.e19, November 2016.

[101] Kyoko Watanabe, Sven Stringer, Oleksandr Frei, Maa Umievi Mirkov, Christiaan de Leeuw, Tinca J C Polderman, Sophie van der Sluis, Ole A Andreassen, Benjamin M Neale, and Danielle Posthuma. A global overview of pleiotropy and genetic architecture in complex traits. *Nature genetics*, 51(9):1339–1348, September 2019.

[102] Anubha Mahajan, Jennifer Wessel, Sara M Willems, Wei Zhao, Neil R Robertson, Audrey Y Chu, Wei Gan, Hidetoshi Kitajima, Daniel Taliun, N William Rayner, Xiuqing Guo, Yingchang Lu, Man Li, Richard A Jensen, Yao Hu, Shaofeng Huo, Kurt K Lohman, Weihua Zhang, James P Cook, Bram Peter Prins, Jason Flannick, Niels Grarup, Vassily Vladimirovich Trubetskoy, Jasmina Kravic, Young Jin Kim, Denis V Rybin, Hanieh Yaghootkar, Martina Müller-Nurasyid, Karina Meidtner, Ruifang Li-Gao, Tibor V Varga, Jonathan Marten, Jin Li, Albert Vernon Smith, Ping An, Symen Ligthart, Stefan Gustafsson, Giovanni Malerba, Ayse Demirkan, Juan Fernandez Tajes, Valgerdur Steinthorsdottir, Matthias Wuttke, Cécile Lecoeur, Michael Preuss, Lawrence F Bielak, Marielisa Graff, Heather M Highland, Anne E Justice, Dajiang J Liu, Eirini Marouli, Gina Marie Peloso, Helen R Warren, ExomeBP Consortium, MAGIC Consortium, GIANT Consortium, Saima Afaq, Shoaib Afzal, Emma Ahlqvist, Peter Almgren, Najaf Amin, Lia B Bang, Alain G Bertoni, Cristina Bombieri, Jette Bork-Jensen, Ivan Brandslund, Jennifer A Brody, Noël P Burtt, Mickaël Canouil, Yii-Der Ida Chen, Yoon Shin Cho, Cramer Christensen, Sophie V Eastwood, Kai-Uwe Eckardt, Krista Fischer, Giovanni Gambaro, Vilmantas Giedraitis, Megan L Grove, Hugoline G de Haan, Sophie Hackinger, Yang Hai, Sohee Han, Anne Tybjærg-Hansen, Marie-France Hivert, Bo Isomaa, Susanne Jäger, Marit E Jørgensen, Torben Jørgensen, Annemari Käräjämäki, Bong-Jo Kim, Sung Soo Kim, Heikki A Koistinen, Peter Kovacs, Jennifer Kriebel, Florian Kronenberg, Kristi Läll, Leslie A Lange, Jung-Jin Lee, Benjamin Lehne, Huaixing Li, Keng-Hung Lin, Allan Linneberg, Ching-Ti Liu, Jun Liu, Marie Loh, Reedik Mägi, Vasiliki Mamakou, Roberta McKean-Cowdin, Girish Nadkarni, Matt Neville, Sune F Nielsen, Ioanna Ntalla, Patricia A Peyser, Wolfgang Rathmann, Kenneth Rice, Stephen S Rich, Line Rode, Olov Rolandsson, Sebastian Schönherr, Elizabeth Selvin, Kerrin S Small, Alena Stanáková, Praveen Surendran, Kent D Taylor, Tanya M Teslovich, Barbara Thorand, Gudmar Thorleifsson, Adrienne Tin, Anke Tönjes, Anette Varbo, Daniel R Witte, Andrew R Wood, Pranav Yajnik, Jie Yao, Loïc Yengo, Robin Young, Philippe Amouyel, Heiner Boeing, Eric Boerwinkle, Erwin P Bottinger, Rajiv Chowdhury, Francis S Collins, George Dedoussis, Abbas Dehghan, Panos Deloukas, Marco M Ferrario, Jean Ferrières, Jose C Florez, Philippe Frossard, Vilmundur Gudnason, Tamara B Harris, Susan R Heckbert, Joanna M M Howson, Martin Ingelsson, Sekar Kathiresan, Frank Kee, Johanna Kuusisto, Claudia Langenberg, Lenore J Launer, Cecilia M Lindgren, Satu Männistö, Thomas Meitinger, Olle Melander, Karen L Mohlke, Marie Moitry, Andrew D Morris, Alison D Murray, Renée de Mutsert, Marju Orho-Melander, Katharine R Owen, Markus Perola, Annette Peters, Michael A Province, Asif Rasheed, Paul M Ridker, Fernando Rivadineira, Frits R Rosendaal, Anders H Rosengren, Veikko Salomaa, Wayne H-H Sheu, Rob Sladek, Blair H Smith, Konstantin Strauch, André G Uitterlinden, Rohit Varma, Cristen J Willer, Matthias Blüher, Adam S Butterworth,

John Campbell Chambers, Daniel I Chasman, John Danesh, Cornelia van Duijn, Josée Dupuis, Oscar H Franco, Paul W Franks, Philippe Froguel, Harald Grallert, Leif Groop, Bok-Ghee Han, Torben Hansen, Andrew T Hattersley, Caroline Hayward, Erik Ingelsson, Sharon L R Kardia, Fredrik Karpe, Jaspal Singh Kooner, Anna Köttgen, Kari Kuulasmaa, Markku Laakso, Xu Lin, Lars Lind, Yongmei Liu, Ruth J F Loos, Jonathan Marchini, Andres Metspalu, Dennis Mook-Kanamori, Børge G Nordestgaard, Colin N A Palmer, James S Pankow, Oluf Pedersen, Bruce M Psaty, Rainer Rauramaa, Naveed Sattar, Matthias B Schulze, Nicole Soranzo, Timothy D Spector, Kari Stefansson, Michael Stumvoll, Unnur Thorsteinsdottir, Tiinamaija Tuomi, Jaakko Tuomilehto, Nicholas J Wareham, James G Wilson, Eleftheria Zeggini, Robert A Scott, Inês Barroso, Timothy M Frayling, Mark O Goodarzi, James B Meigs, Michael Boehnke, Danish Saleheen, Andrew P Morris, Jerome I Rotter, and Mark I McCarthy. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nature genetics*, 50(4):559–571, April 2018.

[103] Iris E Jansen, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, Ida K Karlsson, Sara Hägg, Lavinia Athanasiu, and Others. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimers disease risk. *Nature genetics*, 51(3):404–413, 2019.

[104] Mike A Nalls, Cornelis Blauwendraat, Costanza L Vallerga, Karl Heilbron, Sara Bandres-Ciga, Diana Chang, Manuela Tan, Demis A Kia, Alastair J Noyce, Angli Xue, and Others. Identification of novel risk loci, causal insights, and heritable risk for parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology*, 18(12):1091–1102, 2019.

[105] Matthias Wuttke, Yong Li, Man Li, Karsten B Sieber, Mary F Feitosa, Mathias Gorski, Adrienne Tin, Lihua Wang, Audrey Y Chu, Anselm Hoppmann, and Others. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature genetics*, 51(6):957, 2019.

[106] Loic Yengo, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, Joel Hirschhorn, Jian Yang, Peter M Visscher, and Others. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. *Human molecular genetics*, 27(20):3641–3649, 2018.

[107] Qi Guo, Marjanka K Schmidt, Peter Kraft, Sander Canisius, Constance Chen, Sofia Khan, Jonathan Tyrer, Manjeet K Bolla, Qin Wang, Joe Dennis, and Others. Identification of novel genetic markers of breast cancer survival. *JNCI: Journal of the National Cancer Institute*, 107(5), 2015.

[108] Boxiang Liu, Michael J Gloudemans, Abhiram S Rao, Erik Ingelsson, and Stephen B Montgomery. Abundant associations with gene expression complicate GWAS follow-up. *Nature genetics*, 51(5):768–769, 2019.

[109] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12):1009–1015, 2010.

[110] Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research*, 44(W1):W160–W165, 2016.

[111] Fidel Ramírez, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature communications*, 9(1):1–15, 2018.

[112] Daniel Aletaha, Tuhina Neogi, Alan J Silman, Julia Funovits, David T Felson, Clifton O Bingham, 3rd, Neal S Birnbaum, Gerd R Burmester, Vivian P Bykerk, Marc D Cohen, Bernard Combe, Karen H Costenbader, Maxime Dougados, Paul Emery, Gianfranco Ferraccioli, Johanna M W Hazes, Kathryn Hobbs, Tom W J Huizinga, Arthur Kavanaugh, Jonathan Kay, Tore K Kvien, Timothy Laing, Philip Mease, Henri A Ménard, Larry W Moreland, Raymond L Naden, Theodore Pincus, Josef S Smolen, Ewa Stanislawska-Biernat, Deborah Symmons, Paul P Tak, Katherine S Upchurch, Jirí Vencovský, Frederick Wolfe, and Gillian Hawker. 2010 rheumatoid arthritis classification criteria: an american college of rheumatology/european league against rheumatism collaborative initiative. *Arthritis and rheumatism*, 62(9):2569–2581, September 2010.

[113] Hatice S Kaya-Okur, Steven J Wu, Christine A Codomo, Erica S Pledger, Terri D Bryson, Jorja G Henikoff, Kami Ahmad, and Steven Henikoff. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature communications*, 10(1):1–10, 2019.

[114] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, August 2009.

[115] Aaron R Quinlan. BEDTools: The swiss-army tool for genome feature analysis. *et al [Current protocols in bioinformatics]*, 47(1):11.12.1–34, September 2014.

[116] Yang Liao, Gordon K Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[117] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):R29, 2014.

[118] Ann B Begovich, Victoria E H Carlton, Lee A Honigberg, Steven J Schrodi, Anand P Chokkalingam, Heather C Alexander, Kristin G Ardlie, Qiqing Huang, Ashley M Smith, Jill M Spoerke, and Others. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *The American Journal of Human Genetics*, 75(2):330–337, 2004.

[119] Chieko Kyogoku, Ward A Ortmann, Annette Lee, Scott Selby, Victoria E H Carlton, Monica Chang, Paula Ramos, Emily C Baechler, Franak M Batliwalla, Jill Novitzke, and Others. Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. *The American Journal of Human Genetics*, 75(3):504–507, 2004.

[120] Nunzio Bottini, Lucia Musumeci, Andres Alonso, Souad Rahmouni, Konstantina Nika, Masoud Rostamkhani, James MacMurray, Gian Franco Meloni, Paola Lucarelli, Maurizio Pellecchia, and Others. A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nature genetics*, 36(4):337–338, 2004.

[121] Paul M Ridker, Nader Rifai, Lynda Rose, Julie E Buring, and Nancy R Cook. Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. *New England journal of medicine*, 347(20):1557–1565, 2002.

[122] John Danesh, Jeremy G Wheeler, Gideon M Hirschfield, Shinichi Eda, Gudny Eiriksdottir, Ann Rumley, Gordon D O Lowe, Mark B Pepys, and Vilmundur Gudnason. C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. *New England Journal of Medicine*, 350(14):1387–1397, 2004.

[123] Nasimudeen R Jabir, Chelapram K Firoz, Farid Ahmed, Mohammad A Kamal, Salwa Hindawi, Ghazi A Damanhouri, Hussein A Almehdar, and Shams Tabrez. Reduction in CD16/CD56 and CD16/CD3/CD56 natural killer cells in coronary artery disease. *Immunological investigations*, 46(5):526–535, 2017.

[124] Olivia Osborn and Jerrold M Olefsky. The cellular and signaling networks linking the immune system and metabolism in disease. *Nature medicine*, 18(3):363, 2012.

[125] Yang Shi and David M Holtzman. Interplay between innate immunity and alzheimer disease: APOE and TREM2 in the spotlight. *Nature Reviews Immunology*, 18(12):759–772, 2018.

[126] Yang I Li, Garrett Wong, Jack Humphrey, and Towfique Raj. Prioritizing parkinsons disease genes using population-scale transcriptomic data. *Nature communications*, 10(1):1–10, 2019.

[127] Hadas Keren-Shaul, Amit Spinrad, Assaf Weiner, Orit Matcovitch-Natan, Raz Dvir-Szternfeld, Tyler K Ulland, Eyal David, Kuti Baruch, David Lara-Astaiso, Beata Toth,

and Others. A unique microglia type associated with restricting development of alzheimers disease. *Cell*, 169(7):1276–1290, 2017.

[128] Susanne Krasemann, Charlotte Madore, Ron Cialic, Caroline Baufeld, Narghes Calcagno, Rachid El Fatimy, Lien Beckers, Elaine O'Loughlin, Yang Xu, Zain Fanek, and Others. The TREM2-APOE pathway drives the transcriptional phenotype of dysfunctional microglia in neurodegenerative diseases. *Immunity*, 47(3):566–581, 2017.

[129] Melvin T Korkor, Fan Bo Meng, Shen Yang Xing, Mu Chun Zhang, Jin Rui Guo, Xiao Xue Zhu, and Ping Yang. Microarray analysis of differential gene expression profile in peripheral blood cells of patients with human essential hypertension. *International journal of medical sciences*, 8(2):168, 2011.

[130] Zilun Wei, Yining Yang, Qiaoling Li, Yong Yin, Zhonghai Wei, Wenfeng Zhang, Dan Mu, Jie Ni, Xuan Sun, and Biao Xu. The transcriptome of circulating cells indicates potential biomarkers and therapeutic targets in the course of hypertension-related myocardial infarction. *Genes & Diseases*, 2020.

[131] Chris Wallace. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS genetics*, 16(4):e1008720, April 2020.

[132] Noah J Connally, Sumaiya Nazeen, Daniel Lee, Huwenbo Shi, John Stamatoyannopoulos, Sung Chun, Chris Cotsapas, Christopher A Cassa, and Shamil R Sunyaev. The missing link between genetic association and regulatory function. *eLife*, 11, December 2022.

[133] Katherine A Aracena, Yen-Lung Lin, Kaixuan Luo, Alain Pacis, Saideep Gona, Zepeng Mu, Vania Yotova, Renata Sindeaux, Albena Pramatarova, Marie-Michelle Simon, Xun Chen, Cristian Groza, David Lougheed, Romain Gregoire, David Brownlee, Yang Li, Xin He, David Bujold, Tomi Pastinen, Guillaume Bourque, and Luis B Barreiro. Epigenetic variation impacts ancestry-associated differences in the transcriptional response to influenza infection. *bioRxiv*, page 2022.05.10.491413, May 2022.

[134] Vivek Chandra, Sourya Bhattacharyya, Benjamin J Schmiedel, Ariel Madrigal, Cristian Gonzalez-Colin, Stephanie Fotsing, Austin Crinklaw, Gregory Seumois, Pejman Mohammadi, Mitchell Kronenberg, Bjoern Peters, Ferhat Ay, and Pandurangan Vijayanand. Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants. *Nature genetics*, 53(1):110–119, January 2021.

[135] Haley E Randolph, Jessica K Fiege, Beth K Thielen, Clayton K Mickelson, Mari Shiratori, João Barroso-Batista, Ryan A Langlois, and Luis B Barreiro. Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science*, 374(6571):1127–1133, 2021.

[136] Richard K Perez, M Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, Andrew Lu, Mike Thompson, Nadav Rappoport, Andrew Dahl, Cristina M Lanata, Mehrdad Matloubian, Lenka Maliskova, Serena S Kwek, Tony Li, Michal Slyper, Julia Waldman, Danielle Dionne, Orit Rozenblatt-Rosen, Lawrence Fong, Maria DallEra, Brunilda Balliu, Aviv Regev, Jinoos Yazdany, Lindsey A Criswell, Noah Zaitlen, and Chun Jimmie Ye. Single-cell RNA-seq reveals cell typespecific molecular and genetic associations to lupus. *Science*, 376(6589):eabf1970, 2022.

[137] Heini M Natri, Christina B Del Azodi, Lance Peter, Chase J Taylor, Sagrika Chugh, Robert Kendle, Mei-I Chung, David K Flaherty, Brittany K Matlock, Carla L Calvi, Timothy S Blackwell, Lorraine B Ware, Matthew Bacchetta, Rajat Walia, Ciara M Shaver, Jonathan A Kropski, Davis J Mccarthy, and Nicholas E Banovich. Cell type-specific and disease-associated eQTL in the human lung, March 2023.

[138] Adam W Turner, Shengen Shawn Hu, Jose Verdezoto Mosquera, Wei Feng Ma, Chani J Hodonsky, Doris Wong, Gaëlle Auguste, Yipei Song, Katia Sol-Church, Emily Farber, Soumya Kundu, Anshul Kundaje, Nicolas G Lopez, Lijiang Ma, Saikat Kumar B Ghosh, Suna Onengut-Gumuscu, Euan A Ashley, Thomas Quertermous, Aloke V Finn, Nicholas J Leeper, Jason C Kovacic, Johan L M Björkgren, Chongzhi Zang, and Clint L Miller. Single-nucleus chromatin accessibility profiling highlights regulatory mechanisms of coronary artery disease risk. *Nature genetics*, pages 1–13, May 2022.

[139] Maojun You, Liang Chen, Dawei Zhang, Peng Zhao, Zhu Chen, En-Qiang Qin, Yanan Gao, Mark M Davis, and Pengyuan Yang. Single-cell epigenomic landscape of peripheral immune cells reveals establishment of trained immunity in individuals convalescing from COVID-19. *Nature cell biology*, 23(6):620–630, June 2021.

[140] Malte D Luecken, M Büttner, K Chaichoompu, A Danese, M Interlandi, M F Mueller, D C Strobl, L Zappia, M Dugas, M Colomé-Tatché, and Fabian J Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, pages 1–10, December 2021.

[141] Simone Rubinacci, Diogo M Ribeiro, Robin J Hofmeister, and Olivier Delaneau. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature genetics*, 53(1):120–126, January 2021.

[142] Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427, June 2018.

[143] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, 3rd, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P Mimitou, Jaison Jain,

Avi Srivastava, Tim Stuart, Lamar M Fleming, Bertrand Yeung, Angela J Rogers, Juliana M McElrath, Catherine A Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021.

[144] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, June 2019.

[145] Yuanhua Huang, Davis J McCarthy, and Oliver Stegle. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome biology*, 20(1):273, December 2019.

[146] Peter Carbonetto, Kaixuan Luo, Abhishek Sarkar, Anthony Hung, Karl Tayeb, Sebastian Pott, and Matthew Stephens. Interpreting structure in sequence count data with differential expression analysis allowing for grades of membership. *bioRxiv*, page 2023.03.03.531029, March 2023.

[147] Peter Carbonetto, Abhishek Sarkar, Zihao Wang, and Matthew Stephens. Non-negative matrix factorization algorithms greatly improve topic model fits. *arXiv [stat.ML]*, May 2021.

[148] Jeffrey M Granja, M Ryan Corces, Sarah E Pierce, S Tansu Bagdatli, Hani Choudhry, Howard Y Chang, and William J Greenleaf. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature genetics*, 53(3):403–411, March 2021.

[149] Yao Chen, Ryan A Zander, Xiaopeng Wu, David M Schauder, Moujtaba Y Kasmani, Jian Shen, Shikan Zheng, Robert Burns, Elizabeth J Taparowsky, and Weiguo Cui. BATF regulates progenitor to cytolytic effector CD8+ T cell transition during chronic viral infection. *Nature immunology*, 22(8):996–1007, August 2021.

[150] Gang Xin, David M Schauder, Begoña Lainez, Jason S Weinstein, Zhengxi Dai, Yuhong Chen, Enric Esplugues, Renren Wen, Demin Wang, Ian A Parish, Allan J Zajac, Joe Craft, and Weiguo Cui. A critical role of IL-21-induced BATF in sustaining CD8-T-cell-mediated chronic viral control. *Cell reports*, 13(6):1118–1124, November 2015.

[151] Hsiao-Wei Tsao, James Kaminski, Makoto Kurachi, R Anthony Barnitz, Michael A DiIorio, Martin W LaFleur, Wataru Ise, Tomohiro Kurosaki, E John Wherry, W Nicholas Haining, and Nir Yosef. Batf-mediated epigenetic control of effector CD8+ T cell differentiation. *Science immunology*, 7(68):eabi4919, February 2022.

[152] Mateusz Legut, Zoran Gajic, Maria Guarino, Zharko Daniloski, Jahan A Rahman, Xinhe Xue, Congyi Lu, Lu Lu, Eleni P Mimitou, Stephanie Hao, Teresa Davoli,

Catherine Diefenbach, Peter Smibert, and Neville E Sanjana. A genome-scale screen for synthetic drivers of T cell proliferation. *Nature*, pages 1–8, March 2022.

[153] P E Lipsky. Systemic lupus erythematosus: an autoimmune disease of B cell hyperactivity. *Nature immunology*, 2(9):764–766, September 2001.

[154] Simone Caielli, Zurong Wan, and Virginia Pascual. Systemic lupus erythematosus pathogenesis: Interferon and beyond. *Annual review of immunology*, February 2023.

[155] Yakir A Reshef, Laurie Rumker, Joyce B Kang, Aparna Nathan, Ilya Korsunsky, Samira Asgari, Megan B Murray, D Branch Moody, and Soumya Raychaudhuri. Covarying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics. *Nature biotechnology*, October 2021.

[156] Emma Dann, Neil C Henderson, Sarah A Teichmann, Michael D Morgan, and John C Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature biotechnology*, September 2021.

[157] Stinne Ravn Greisen, Karen Kræmmer Schelde, Tue Kruse Rasmussen, Tue Wenzel Kragstrup, Kristian Stengaard-Pedersen, Merete Lund Hetland, Kim Hørslev-Petersen, Peter Junker, Mikkel Østergaard, Bent Deleuran, and Malene Hvid. CXCL13 predicts disease activity in early rheumatoid arthritis and could be an indicator of the therapeutic 'window of opportunity'. *Arthritis research & therapy*, 16(5):434, September 2014.

[158] Antoine W T van Lieshout, Jaap Fransen, Marcel Flendrie, Agnes M M Eijsbouts, Frank H J van den Hoogen, Piet L C M van Riel, and Timothy R D J Radstake. Circulating levels of the chemokine CCL18 but not CXCL16 are elevated and correlate with disease activity in rheumatoid arthritis. *Annals of the rheumatic diseases*, 66(10):1334–1338, October 2007.

[159] X Wang, D Liu, Y Ning, J Liu, X Wang, R Tu, H Shen, Q Chen, and Y Xiong. Siglec-9 is upregulated in rheumatoid arthritis and suppresses collagen-induced arthritis through reciprocal regulation of Th17-/treg-cell differentiation. *Scandinavian journal of immunology*, 85(6):433–440, June 2017.

[160] Haruyasu Ito, Kentaro Noda, Ken Yoshida, Kazuhiro Otani, Masayuki Yoshiga, Yohsuke Oto, Saburo Saito, and Daitaro Kurosaka. Prokineticin 2 antagonist, PKRA7 suppresses arthritis in mice with collagen-induced arthritis. *BMC musculoskeletal disorders*, 17(1):387, September 2016.

[161] Kentaro Noda, Bianca Dufner, Haruyasu Ito, Ken Yoshida, Gianfranco Balboni, and Rainer H Straub. Differential inflammation-mediated function of prokineticin 2 in the synovial fibroblasts of patients with rheumatoid arthritis compared with osteoarthritis. *Scientific reports*, 11(1):18399, September 2021.

[162] Natsuhiko Kumasaka, Andrew J Knights, and Daniel J Gaffney. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nature genetics*, 51(1):128–137, January 2019.

[163] Nicholas E Banovich, Yang I Li, Anil Raj, Michelle C Ward, Peyton Greenside, Diego Calderon, Po Yuan Tung, Jonathan E Burnett, Marsha Myrthil, Samantha M Thomas, Courtney K Burrows, Irene Gallego Romero, Bryan J Pavlovic, Anshul Kundaje, Jonathan K Pritchard, and Yoav Gilad. Impact of regulatory variation across human iPSCs and differentiated cells. *Genome research*, 28(1):122–131, January 2018.

[164] Joseph Nasser, Drew T Bergman, Charles P Fulco, Philine Guckelberger, Benjamin R Doughty, Tejal A Patwardhan, Thouis R Jones, Tung H Nguyen, Jacob C Ulirsch, Fritz Lekschas, Kristy Mualim, Heini M Natri, Elle M Weeks, Glen Munson, Michael Kane, Helen Y Kang, Ang Cui, John P Ray, Thomas M Eisenhaure, Ryan L Collins, Kushal Dey, Hanspeter Pfister, Alkes L Price, Charles B Epstein, Anshul Kundaje, Ramnik J Xavier, Mark J Daly, Hailiang Huang, Hilary K Finucane, Nir Hacohen, Eric S Lander, and Jesse M Engreitz. Genome-wide enhancer maps link risk variants to disease genes. *Nature*, 593(7858):238–243, May 2021.

[165] B J Strober, R Elorbany, K Rhodes, N Krishnan, K Tayeb, A Battle, and Y Gilad. Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, 364(6447):1287–1290, June 2019.

[166] Benjamin D Umans, Alexis Battle, and Yoav Gilad. Where are the disease-associated eQTLs? *Trends in genetics: TIG*, 37(2):109–124, February 2021.

[167] Yuwen Zhu, Sheng Yao, Bettina P Iliopoulou, Xue Han, Mathew M Augustine, Haiying Xu, Ryan T Phennicie, Sarah J Flies, Megan Broadwater, William Ruff, Janis M Taube, Linghua Zheng, Liqun Luo, Gefeng Zhu, Jianzhu Chen, and Lieping Chen. B7-H5 costimulates human T cells via CD28H. *Nature communications*, 4:2043, 2013.

[168] Katayoun Dolatkhah, Nazila Alizadeh, Hanieh Mohajjel-Shoja, Mahdi Abdoli Shadbad, Khalil Hajiasgharzadeh, Leili Aghebati-Maleki, Amir Baghbanzadeh, Negar Hosseinkhani, Noora Karim Ahangar, and Behzad Baradaran. B7 immune checkpoint family members as putative therapeutics in autoimmune disease: An updated overview. *International journal of rheumatic diseases*, 25(3):259–271, March 2022.

[169] Yan Yang, Yanfeng Wang, Qingwei Liang, Lutian Yao, Shizhong Gu, and Xizhuang Bai. MiR-338-5p promotes inflammatory response of fibroblast-like synoviocytes in rheumatoid arthritis via TargetingSPRY1. *Journal of cellular biochemistry*, 118(8):2295–2301, August 2017.

[170] Sylvan C Baca, Cassandra Singler, Soumya Zacharia, Ji-Heui Seo, Tunc Morova, Faraz Hach, Yi Ding, Tommer Schwarz, Chia-Chi Flora Huang, Jacob Anderson, André P Fay, Cynthia Kalita, Stefan Groha, Mark M Pomerantz, Victoria Wang, Simon

Linder, Christopher J Sweeney, Wilbert Zwart, Nathan A Lack, Bogdan Pasaniuc, David Y Takeda, Alexander Gusev, and Matthew L Freedman. Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation. *Nature genetics*, September 2022.

[171] Steven Gazal, Omer Weissbrod, Farhad Hormozdiari, Kushal K Dey, Joseph Nasser, Karthik A Jagadeesh, Daniel J Weiner, Huwenbo Shi, Charles P Fulco, Luke J O'Connor, Bogdan Pasaniuc, Jesse M Engreitz, and Alkes L Price. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nature genetics*, pages 1–10, June 2022.

[172] Alexandra Zhernakova, Eleanora M Festen, Lude Franke, Gosia Trynka, Cleo C van Diemen, Alienke J Monsuur, Marianna Bevova, Rian M Nijmeijer, Ruben van 't Slot, Roel Heijmans, H Marike Boezen, David A van Heel, Adriaan A van Bodegraven, Pieter C F Stokkers, Cisca Wijmenga, J Bart A Crusius, and Rinse K Weersma. Genetic analysis of innate immunity in crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP. *American journal of human genetics*, 82(5):1202–1210, May 2008.

[173] Ping Luo, Zhiwen Yang, Bin Chen, and Xiaoming Zhong. The multifaceted role of CARD9 in inflammatory bowel disease. *Journal of cellular and molecular medicine*, 24(1):34–39, January 2020.

[174] Manuel A Rivas, Mélissa Beaudoin, Agnes Gardet, Christine Stevens, Yashoda Sharma, Clarence K Zhang, Gabrielle Boucher, Stephan Ripke, David Ellinghaus, Noel Burtt, Tim Fennell, Andrew Kirby, Anna Latiano, Philippe Goyette, Todd Green, Jonas Halfvarson, Talin Haritunians, Joshua M Korn, Finny Kuruvilla, Caroline Lagacé, Benjamin Neale, Ken Sin Lo, Phil Schumm, Leif Törkvist, National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC), United Kingdom Inflammatory Bowel Disease Genetics Consortium, International Inflammatory Bowel Disease Genetics Consortium, Marla C Dubinsky, Steven R Brant, Mark S Silverberg, Richard H Duerr, David Altshuler, Stacey Gabriel, Guillaume Lettre, Andre Franke, Mauro D'Amato, Dermot P B McGovern, Judy H Cho, John D Rioux, Ramnik J Xavier, and Mark J Daly. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics*, 43(11):1066–1073, October 2011.

[175] Lena Krzyzak, Christine Seitz, Anne Urbat, Stefan Hutzler, Christian Ostalecki, Joachim Gläsner, Andreas Hiergeist, André Gessner, Thomas H Winkler, Alexander Steinkasserer, and Lars Nitschke. CD83 modulates B cell activation and germinal center responses. *The journal of immunology*, 196(9):3581–3594, May 2016.

[176] Hamish W King, Kristen L Wells, Zohar Shipony, Arwa S Kathiria, Lisa E Wagar, Caleb Lareau, Nara Orban, Robson Capasso, Mark M Davis, Lars M Steinmetz, Louisa K James, and William J Greenleaf. Integrated single-cell transcriptomics and

epigenomics reveals strong germinal center-associated etiology of autoimmune risk loci. *Science immunology*, 6(64):eabh3768, October 2021.

[177] Joyce B Kang, Amber Z Shen, Saori Sakaue, Yang Luo, Saisriram Gurajala, Aparna Nathan, Laurie Rumker, Vitor R C Aguiar, Cristian Valencia, Kaitlyn Lagattuta, Fan Zhang, Anna Helena Jonsson, Seyhan Yazar, Jose Alquicira-Hernandez, Hamed Khalili, Ashwin N Ananthakrishnan, Karthik Jagadeesh, Kushal Dey, Accelerating Medicines Partnership Program: RA/SLE Network, Mark J Daly, Ramnik J Xavier, Laura T Donlin, Jennifer H Anolik, Joseph E Powell, Deepak A Rao, Michael B Brenner, Maria Gutierrez-Arcelus, and Soumya Raychaudhuri. Mapping the dynamic genetic regulatory architecture of HLA genes at single-cell resolution. *medRxiv*, page 2023.03.14.23287257, March 2023.

[178] Kathryn Weinand, Saori Sakaue, Aparna Nathan, Anna Helena Jonsson, Fan Zhang, Gerald F M Watts, Zhu Zhu, Deepak A Rao, Jennifer H Anolik, Michael B Brenner, Laura T Donlin, Kevin Wei, Soumya Raychaudhuri, and Accelerating Medicines Partnership Program: RA and SLE Network. The chromatin landscape of pathogenic transcriptional cell states in rheumatoid arthritis. *bioRxiv*, page 2023.04.07.536026, April 2023.

[179] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, Richard Durbin, and Alkes L Price. Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443–1448, November 2016.

[180] Zuguang Gu. Complex heatmap visualization. *iMeta*, 1(3), September 2022.

[181] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[182] Buhm Han and Eleazar Eskin. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *American journal of human genetics*, 88(5):586–598, May 2011.

[183] Kaelan J Brennan, Melanie Weilert, Sabrina Krueger, Anusri Pampari, Hsiao-Yun Liu, Ally W H Yang, Jason A Morrison, Timothy R Hughes, Christine A Rushlow, Anshul Kundaje, and Julia Zeitlinger. Chromatin accessibility in the drosophila embryo is determined by transcription factor pioneering and enhancer activation. *Developmental cell*, August 2023.

[184] Sarah M Brotman, Julia S El-Sayed Moustafa, Li Guan, K Alaine Broadaway, Dongmeng Wang, Anne U Jackson, Ryan Welch, Kevin W Currin, Max Tomlinson, Swarooparani Vadlamudi, Heather M Stringham, Amy L Roberts, Timo A Lakka, Anniina Oravilahti, Lilian Fernandes Silva, Narisu Narisu, Michael R Erdos, Tingfen

Yan, Lori L Bonnycastle, Chelsea K Raulerson, Yasrab Raza, Xinyu Yan, Stephen C J Parker, Johanna Kuusisto, Paivi Pajukanta, Jaakko Tuomilehto, Francis S Collins, Michael Boehnke, Michael I Love, Heikki A Koistinen, Markku Laakso, Karen L Mohlke, Kerrin S Small, and Laura J Scott. Adipose tissue eQTL meta-analysis reveals the contribution of allelic heterogeneity to gene expression regulation and cardiometabolic traits. *bioRxiv*, page 2023.10.26.563798, October 2023.

[185] Ting Qi, Yang Wu, Jian Zeng, Futao Zhang, Angli Xue, Longda Jiang, Zhihong Zhu, Kathryn Kemper, Loic Yengo, Zhili Zheng, eQTLGen Consortium, Riccardo E Marioni, Grant W Montgomery, Ian J Deary, Naomi R Wray, Peter M Visscher, Allan F McRae, and Jian Yang. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nature communications*, 9(1):2282, June 2018.

[186] iga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature genetics*, 53(3):354–366, March 2021.

[187] Seungsoo Kim and Joanna Wysocka. Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular cell*, 83(3):373–392, February 2023.