THE UNIVERSITY OF CHICAGO

UNRAVELING THE GENETIC AND EPIGENETIC BASIS OF INTER-INDIVIDUAL VARIATION IN IMMUNE RESPONSE TO INFLUENZA INFECTION

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

DEPARTMENT OF HUMAN GENETICS

BY

KATHERINE ANNE ARACENA

CHICAGO, ILLINOIS

AUGUST 2023

To my mother, father, and husband

"A hustler's work is never through."

– The song Play Hard by David Guetta (feat. Ne-Yo and Akon)

Table of Contents

Supplementary tables are available as files online. The List of Tables gives the page number for each table's caption.

List of Figures

List of Tables

Acknowledgements

First and foremost, I must thank my advisor, Dr. Luis Barreiro for his unwavering support during my time in his lab. Thank you for trusting me, giving me both the attention and space I needed to grow, and being able to detect when I needed which. Thank you for allowing me to be myself and never making me feel ashamed for not knowing something, but instead taking the opportunity to teach me. I admire your creativity and ability to make any result compelling, yet always truthful. I am truly grateful to have completed my PhD in your lab.

I also must thank all current and past members of the Barreiro Lab for the supportive and fun work environment and allowing me to grow alongside you. In particular, I have to thank Paul Maurizio for teaching me how to log on to UChicago's cluster, a pivotal moment in my scientific career. I would also like to thank Yen-lung "Onta" Lin with whom I worked closely for many analyses. Lastly, a thank you to the "Only Child Club" rebranded to the "No Boys Club," including Dr. Haley Randolph, Mari Shiratori, Dr. Sarah Sun, Veronica Locher, Bridget Chak and Ellen Ketter. You all made the work environment more silly and enjoyable.

Thank you to my thesis committee members, Drs. Xin He, Carole Ober, and Yoav Gilad, for providing valuable discussions and feedback on my work. Thank you to all my collaborators without whom the work presented in this thesis would not be possible, especially Dr. Guillaume Bourque. I must also recognize my past research mentors, including the late Dr. Christopher (Casey) Brown and Dr. Benjamin Blackman whose lab environments and mentorship motivated me to continue pursuing a career in scientific research. Thank you to Peter Carbonetto for answering my email late at night when I thought I accidentally deleted all the raw ATACseq fastqs during my rotation and introducing me to snapshot on midway2 (I did not actually delete them,

Luis, by the way, I just renamed them). Similarly, thank you to Sue Levinson for always being available to ask for help and being an advocate for students.

To Dr. Anna Di Rienzo, the Human Genetics Program Director my first year—I will forever remember your empathy, compassion, and determination to ensure that I could care for my sick father and still preserve my place in the Human Genetics program. You arranged an independent schedule for me to be able to fulfill all the requirements and successfully complete the program. You assured me that I could take a leave of absence and return when I was ready, which was the best thing someone could have done for me. Thank you for how much you cared about me and all students.

To my friends and fellow graduate students: first, to the OG Genetics women, Kate Farris, Dr. Selene Clay and Dr. Sammy Keyport Kik, thank you for always being there to giggle, gossip, and complain in various proportions. We had many good times with Dr. Grace Hansen, Meike Lobb-Rabe, Édgar Correa, and so many more. To the Bachelors of Science, Shreya Ramachandran, Jennifer Blanc, JoJo Tang, and Maggie Steiner in addition to Kate and Selene, thank you for the critical and hilarious conversations inspired by the various reality tv shows we watched together. Lastly, thank you to my Advocacy Spouses, Elaine Kouame and Jake Higgins, for your constant uplifting support and raw friendship.

To my parents, Dr. Beth Roth and Ramón Aracena, a big part of the reason I came to the University of Chicago was to be like you. To my mom, thank you for instilling a love of learning in me and guiding me during every step of my education. Thank you for always being there to debrief and offer advice. To my father, I have never doubted how proud of me you would be on the day I officially received my PhD. I thought of you often during this journey, especially when

something small bothered me. I knew you would say something along the lines of “just forget about them.”

To my husband, Dr. Michael Zhang, thank you for being by my side on this journey start to finish. I still remember when we ate lunch together every day while working in our respective undergraduate labs in the summer. Thank you for your unwavering love, support, and knowledge of when modern humans moved out of Africa (approximately 60,000 years ago to populate the globe). With you the highs of life are sweeter and the lows easier to bear.

Abstract

Humans display remarkable inter-individual variation in immune response, in part due to population specific adaptations resulting from different pathogenic environments encountered during the out of Africa movement. To date, most of the variance in transcriptional differences in immune responses remains unaccounted for when considering genetics alone, suggesting other mechanisms at play. Epigenetics is known to play a central role in the regulation of immune response, yet few studies have investigated how genetic ancestry and genetic variation impact epigenetic signatures in response to pathogens. To investigate this, we carried out in-depth genetic, epigenetic, and transcriptional profiling on primary macrophages derived from a panel of individuals with varying proportions of European and African ancestry before and after infection with influenza A virus (IAV). We find that the immune response to IAV infection involves a coordinated response of the gene regulatory landscape, and that both transcriptional and epigenetic effects act to upregulate main inflammatory response pathways. Baseline epigenetic profiles are strongly predictive of the transcriptional response to IAV across individuals, and ancestry-associated differences in gene expression are tightly coupled with variation in enhancer activity. Quantitative trait locus (QTL) mapping reveals highly coordinated genetic effects on gene regulation with many cis-acting genetic variants impacting concomitantly gene expression and multiple epigenetic marks. We find epigenetic ancestry-associated differences are genetically controlled, even more so than gene expression, and that epigenetic QTL are also more frequently associated with immune-disease risk. Lastly, we explore inter-individual and intra-individual variability in DNA methylation levels and corresponding transcriptional response in DNA methylation-dependent enhancer regions, revealing that baseline DNA methylation predicts transcriptional response. Together, our results provide a comprehensive characterization of the

genetic and epigenetic basis of variation in immune response to IAV infection in individuals of European and African ancestry.

Chapter I: Introduction

The post-GWAS boom and rise of functional genomics

The field of human genetics has long been driven by the desire to understand the genetic basis of traits and disease. The release of the human reference genome in 2000 was followed by a drastic reduction in sequencing costs and the development of high-throughput genotyping arrays, leading to a significant increase in data generation for population-scale studies. A boom of genome-wide association studies (GWAS) using these sequencing data followed[1]. GWAS detects genetic variants associated with traits and disease by utilizing genotype data to test for differences in allele frequency across individuals. Since its first application in 2005, GWAS has identified hundreds of thousands of genetic variants associated with over 3,300 complex traits and diseases[2–5]. Though this brought new appreciation for the sheer number of genetic variants contributing towards complex traits, GWAS arguably fell short at being able to provide meaningful insights into disease mechanisms. The function of GWAS-identified variants and the mechanisms by which they act is largely unknown.

In the post-GWAS era, many questions on how phenotypic variation is influenced by genetic variation remain. The field of functional genomics has sought to address these by harnessing sequence-based multi-omics analyses to profile the function and regulation of gene

expression and other molecular traits on a genome scale. These maps of regulatory activity and interactions can aid in the uncovering of the mechanisms by which previously identified variants act to influence phenotype. Functional genomics focused studies can build off GWAS by filling in several key gaps, including understanding how environment, genetic ancestry and genetic variation impacting epigenetics converge to modify gene regulatory mechanisms.

A major limitation of GWAS is identifying the biological relevance of significant variants, most often single nucleotide polymorphisms (SNPs). The vast majority of GWAS significant SNPs fall within non-coding regions of the genome, leaving the functional effect of these variants unclear[6,7]. GWAS significant SNPs often act upon the closest gene, though this is not always the case, and chromatin remodeling in response to developmental cues or external stimuli may affect interactions[8–10]. Functional genomics studies can begin to uncover how a non-coding variant affects its target gene by measuring genomic interactions and response in disease-relevant environments. Genetic factors are rarely the sole contributor to variation in complex traits and diseases as the vast majority of phenotypes are also influenced by interactions with environmental factors. Context-specific gene regulatory mechanisms can be identified through experimental designs that closely simulate a disease state.

Furthermore, GWAS are not fully representative of all human populations and are still lacking for some disease traits. As of April 2023, over 95% of GWAS participants were of European ancestry, while only 0.17% were of African ancestry, despite individuals of African descent having the largest amount of genetic diversity of any human population resulting from the out of Africa migratory event[11]. This lack of data inevitably leads to many disease variants being missed and limits the transferability of disease associated variants to non-European populations. Moreover, despite immune response being one of the most divergent phenotypic traits between

populations, relatively few GWAS on susceptibility to infectious diseases have been performed[3]. Some diseases have been entirely overlooked, such as susceptibility to the parasite helminths, despite being one of the most common infections worldwide[12]. Functional genomics studies provide an opportunity to build our knowledge on how genetically diverse individuals have similar or different responses to varied environmental challenges, including infection.

Lastly, genetic variants identified by GWAS do not explain the total heritability of traits, which has been termed the "missing heritability problem". Heritability is defined as the total amount of phenotypic variation that can be attributed to genetic variation. Across traits, genetic components identified by GWAS do not adequately explain total heritability. One potential explanation is that we have missed the full spectrum of human genetic variation due to lack of sampling of sparsely distributed genetic variants, structural variation, and individuals of African descent[13,14]. On the other hand, interactions between loci, defined as epistasis, or with the environment, defined as gene by environment interactions, may lead to underestimating or missing associations as GWAS may not measure the effects of variants in the relevant disease context[15,16]. Additionally, there is increasing evidence that suggests trait heritability may be affected by epigenetic mechanisms[17]. With functional genomics studies that assay for a multitude of molecular traits, we can begin to understand how transcription and epigenetic mechanisms act together to coordinate gene regulation in specific environmental contexts. One such environmental context is in response to pathogens, one of the strongest sources of selective pressure in human evolutionary history[18].

Population differences in immune responses

Individuals from different populations vary considerably in their susceptibility to infectious diseases, chronic inflammatory disorders, and autoimmune disorders, in part due to population-specific adaptation to the different pathogenic environments encountered as humans moved out of Africa[18–21]. For instance, African American and European American individuals have up to a 3-fold difference in prevalence of tuberculosis, systemic lupus erythematosus, systemic sclerosis, psoriasis, and septicemia[22–24]. These differences make utilizing a diverse genetic ancestry cohort particularly important to capture the full spectrum of variation in immune response.

Environmental factors have also been found to impact immune response. Specifically, sex[25,26], age[27–29], gut microbiome diversity[30], and the social environment[31–33] have all been linked to variation in immune response. For example, young children and the elderly have increased susceptibility to infections, potentially due to changes in immune system cell composition over time[28]. Meanwhile, over 80% of autoimmune disease patients are female, possibly due to immune interactions with hormone regulation[26]. Since an individual's environment changes over time, it is difficult to pinpoint the exact root of these differences without carefully designed experimental models.

In addition to environmental factors, host genetics also contributes towards the heterogeneity in response to infection across populations. In human evolutionary history, selective pressures driven by pathogen exposure likely drove natural selection in genomic regions involved in the host immune response, causing allele frequencies to diverge at loci. Natural selection acting on gene regulatory mechanisms underlying susceptibility to infection likely contributes to the population level differences to infection we see today[18,19,21,34]. In fact, strong signatures of selection have been reported for genes differentially regulated between individuals of African and

European ancestry[21,35]. These complex interactions make immune response a particularly interesting study system to understand the relative contributions of genetic and environmental impact to disease. To determine if these differences are driven by genetic variation or merely environmental factors associated with ancestry, we can use QTL mapping.

Quantitative trait loci mapping

Quantitative trait loci (QTL) mapping identifies genetic variants that are associated with the level of a molecular trait. This technique is conceptually similar to GWAS, though GWAS are typically performed for non-molecular traits and scan the entire genome for associations[36]. QTL mapping can be performed with any molecular trait, including gene expression (eQTL), chromatin accessibility (caQTL), histone modifications (hQTL), and CpG methylation (meQTL). This allows us to test the effects of genetic variation not only on gene expression but on epigenetic traits as well. QTL mapping can be performed in both *cis,* usually defined as testing for associations within a 100kb to 1Mb window surrounding the gene/peak/CpG site and in *trans,* >5Mb away or on other chromosomes[36].

QTL may be tissue, cell type and environment-specific, or shared across many contexts[21,35,37,38]. This highlights the unique insights that performing QTL mapping in a specific context can bring. Immune response QTL mapping can identify genetic variants that underlie differential immune responses to infection[39]. In the context of immune stimuli, previous studies have shown infection-specific QTL[21,35,40]. For example 21.8% of QTL found in macrophages were only found after *Salmonella* or *Listeria* infection[21]. These genetic effects would have been missed in non-stimulated cells, showcasing the importance of performing QTL mapping studies in the

disease relevant contexts to capture context-specific effects that more closely mimic *in vivo* infection.

Previous immune QTL studies have primarily focused on eQTL mapping to identify how single-nucleotide polymorphisms (SNPs) functionally impact transcriptional differences in response to bacterial and viral stimuli[21,35,37,38]. Increasing evidence suggests that epigenetic modifications may play a role in regulatory changes affecting gene expression, yet little is known about the mechanism of these changes in response to immune stimulation[41–43]. Immune response QTL mapping for multiple traits in the same individual gives us the potential to interpret overlapping combinations of regulatory interactions and identify how genetic variation impacts these mechanisms in response to infection[44,45]. For example, if an eQTL acts by perturbing a particular regulatory mechanism, such as a histone modification, the eQTL SNPs will also be a histone mark QTL[43]. With a system-level approach to observe genomic changes in response to immune stimulation, we can harness QTL mapping to understand the combinatorial effect of regulatory changes across molecular traits. This makes QTL mapping a crucial tool in understanding the effects of genetic variation across epigenetic traits and how interactions modulate disease.

Epigenetic contributions on gene regulatory mechanisms

Epigenetics, meaning "above the genetics", is the study of mechanisms that affect gene expression without changing the underlying DNA sequence. Increasing evidence suggests that epigenetic modifications, particularly DNA methylation, are mitotically heritable and transmitted across generations, making them a potential contributor of heritability[15,17]. While these changes can be maintained for cell generations, epigenetic markings are also dynamic and adaptable to

varied environments. As with gene expression, epigenetic markings are unique to cell types, stages of development, and environmental exposures[46].

Types of epigenetic modifications include post-translational modification of histone tails and modulation of DNA methylation, which are both closely tied to chromatin accessibility. The modification of histones tails often occurs at enhancer and promoter regions and can take many different forms, including acetylation, methylation, ubiquitination, and phosphorylation. Specific patterns of histone modifications associated with active and repressed regions of the genome have been well characterized; for example, histone 3 lysine 27 acetylation (H3K27ac) is found at active promoters and enhancers, while H3K27me3 is associated with repressed regions[47,48]. Active promoters are marked by trimethylation of lysine 4 of histone H3 (H3K4), and active enhancers marked by monomethylation of H3K4[49]. DNA methylation, the addition of a methyl group on to the C-5 position of the cytosine ring of DNA, is a modification that acts directly on the DNA[50]. Typically, high levels of methylation are associated with repressed or inactive regions, while low levels of methylation are associated with active regions of the genome. Lastly, chromatin, the packaging structure of DNA, is a highly organized and dynamic system regulated by histones. Transcription occurs in regions of accessible and open chromatin and inaccessible regions will remain inactive. As described here, epigenetic modifications are highly inter-connected, though the causal relationships between epigenetic modifications in the gene regulatory cascade remain largely unknown.

At present, the relationship between the histone modifications and DNA methylation with chromatin accessibility remain largely correlative[51]. One hypothesis is that chromatin accessibility states are driven by pioneer transcription factor binding, promoting opening of chromatin and localized gene regulatory activity[51]. Disrupting transcription factor binding has been proposed as

a mechanism by which other epigenetic mechanisms act as well. In particular, it has been suggested that methylation may affect the binding dynamics of pioneer TFs, which play a crucial role in attracting secondary factors and downstream transcriptional changes[43]. In some cases, environmentally induced transcription factors may also be required to drive and regulate gene expression.

These mechanisms may help engage the rapid gene regulatory response required during infection. Previous studies have shown that certain stimuli or lifestyle factors are able to "educate" chromatin state of innate immune cells, notably monocytes and macrophages[52,53]. This epigenetic priming may affect transcriptional response during infection with a similar immune stimulus, increasing the speed or severity of the response. It has been hypothesized that this priming is due to the presence of H3K27ac, H3K4me1, H3K4me3, and chromatin remodelers promoting increased accessibility at enhancers and promoters of upregulated genes[51,54–56]. That is not to say that all or even most epigenetic changes are already in place in the baseline state. Though there is increased signal at pre-existing histone modifications associated with gene regulatory activity after infection, there is also evidence of de novo formation of enhancer elements absent in the baseline, non-infected state[51,57,58]. These findings recognize that epigenetic modifications are dynamic key contributors of immune response. Perhaps differences in these epigenetic responses contribute to population differences in immune response.

Despite meaning "above the genetics", epigenetic variation across individuals has been shown to be impacted by genetic variation[44,59–64]. Compared to the number of studies on gene expression, relatively few studies have explicitly tested the effects of genetic variation on epigenetic levels in response to immune stimuli, despite evidence of genetic variation impacting the levels of chromatin accessibility and DNA methylation in response to various pathogens[60,61,64].

Even fewer studies have investigated the impact of genetic variation in diverse ancestry cohorts. Investigating the interplay between epigenetic and genetic variation in different disease contexts may provide insights into how GWAS significant variants affect gene regulatory networks and reveal their biological relevance. A mere ~20% of GWAS loci colocalize with eQTL in the disease-relevant tissue or cell type[65,66], suggesting that alternative gene regulatory mechanisms may mediate genetic effects[67]. However, comprehensive datasets of population level variation in epigenetic levels in many disease contexts have yet to be created. Creating resources of epigenetic variation in specific disease contexts can help pinpoint mechanisms of disease and potentially become drug target candidates for disease treatment.

Dissertation Overview

In this thesis, I use a functional genomics approach to investigate how genetic variation impacts molecular traits in the immune response to infection. I introduce the first dataset to generate complete genetic and epigenetic profiles for the same individuals across two different populations known to have variation in immune response (European and African American). I perform in-depth analysis using these data to understand how population differences affect transcriptional and epigenetic changes in response to influenza A infection. I harness QTL mapping to connect genetic variation to variability across these molecular traits and understand the relative contributions of genetic variation on population level differences. I provide evidence that epigenetic traits can significantly contribute to our understanding of how genetic variants affect and explain GWAS-identified immune-related disease risk.

Secondly, I investigate the relationship between DNA methylation levels at baseline and transcriptional response to validate *in vitro* predictions of DNA methylation-dependent enhancer

activity. I show that pre-existing DNA methylation patterns can influence the response to subsequent environmental exposures. Together, this work emphasizes the value of generating multi-molecular trait datasets to study the impact of epigenetic variation on the gene regulatory network. This systems-based approach further untangles the mechanisms underlying the relationship between genotype and phenotype in humans and expands our knowledge of immune gene regulation to traditionally understudied populations.

Chapter II: Epigenetic variation impacts inter-individual differences in the transcriptional response to influenza infection

Note:

The following section (*Chapter II*) is reproduced verbatim, with the exception of figure numbering and reference labeling, form the manuscript titled "Epigenetic variation impacts ancestry-associated differences in the transcriptional response to influenza infection". A previous version of this manuscript was published on BioRxiv in 2022[68]. The following is a revised version pending publication.

Authors:

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, Carly Boye, Roger Pique-Regi, Yang Li, Xin He, David Bujold, Tomi Pastinen, Guillaume Bourque, Luis B Barreiro

Abstract

Humans display remarkable inter-individual variation in immune response when exposed to identical immune challenges[22,23,69]. Yet, our understanding of the genetic and epigenetic factors contributing to such variation remains limited. Here we carried out in-depth genetic, epigenetic, and transcriptional profiling on primary macrophages derived from a panel of European and African-ancestry individuals before and after infection with influenza A virus (IAV). We show that baseline epigenetic profiles are strongly predictive of the transcriptional response to IAV across individuals, and that ancestry-associated differences in gene expression are tightly coupled with variation in enhancer activity. Quantitative trait locus (QTL) mapping revealed highly coordinated genetic effects on gene regulation with many cis-acting genetic variants impacting concomitantly gene expression and multiple epigenetic marks. These data reveal that ancestry-associated differences in the epigenetic landscape are genetically controlled, even more so than variation in gene expression. Lastly, we show that among QTL variants that colocalized with immune-disease loci, only 7% were gene expression QTL, the remaining corresponding to genetic variants that impact one or more epigenetic marks, which stresses the importance of considering molecular phenotypes beyond gene expression in disease-focused studies.

Introduction

Inter-individual differences in the transcriptional response of innate immune cells to infectious agents are common and likely contribute to varying susceptibility to infectious diseases, inflammation, and autoimmune disorders[22,23,69]. Although a substantial fraction of transcriptional heterogeneity in the response to infection is likely attributable to environmental factors, several studies have shown that host genetics also plays an important role[21,27,29,35,40]. For example, it has been shown that ~30% of the transcriptional differences between European and African ancestry individuals in their immune responses to influenza A infection can be explained by expression quantitative trait loci (eQTL) that vary in allele frequency across populations[40]. Similar genetic contributions to ancestry-associated differences in the transcriptional response to intracellular bacterial pathogens and immune stimuli have been reported[21,35,70].

However, much of the variance in immune responses observed at the population level remains unexplained by genetics alone[27,29,71,72]. Other factors that have been linked to variation in immune responses include sex, age[27,29],gut microbiome diversity[30], and the social environment[31–33]. Although less studied, epigenetic variation is also likely to play an important role in explaining immune response variance. The most well studied epigenetic responses to immune stimuli involve the post-translational modification of histone tails at promoter and enhancer regions[73,74]. Histone acetylation is strongly associated with the activation of many pro-inflammatory genes[75,76], whereas histone deacetylation is often associated with gene repression in the context of inflammation[77]. Moreover, certain inflammatory signals (e.g., β-glucan or Bacillus Calmette–Guerin (BCG) vaccination) or even lifestyle factors (e.g., diet) are thought to be able to "educate" the chromatin state of innate immune cells, notably monocytes/macrophages, resulting in a stronger

transcriptional response during reinfection[52,53]. This suggests environmentally induced epigenetic changes may represent crucial determinants of an individual's ability to respond to pathogens.

Although the term epigenetics means "above the genetics", genetic variation has also been shown to play a substantial role in the degree of epigenetic variation across individuals[44,59–61,63,78,79]. In human lymphoblastoid cell lines, genetic variation has been shown to impact the levels of chromatin accessibility at thousands of enhancer and promoter elements throughout the genome[60]. Likewise, genetically controlled variation in chromatin accessibility has been shown to impact the magnitude of the response engaged by human macrophages in response to *Salmonella*[64]. Thus, it is likely that variation in epigenetic profiles across individuals and populations – whether genetically controlled or not – can ultimately represent a key contributor to population variation in innate immune responses and susceptibility to disease. However, despite intense efforts to generate comprehensive epigenomic atlases across many tissues and cell types[47,80–82], there are no comprehensive maps of population variation in epigenetic levels in primary innate immune cells before and after infection, preventing the formal evaluation of such hypotheses.

To address this gap, we carried out an in-depth genetic and epigenetic characterization of primary macrophages derived from 35 individuals with varying degrees of European and African ancestry at both baseline and after infection with influenza A. The data generated herein helps fill a critical gap in biomedical research: the lack of non-European ancestry individuals among cohorts designed to study immune variance in the general population and in genomic studies more generally. All data generated in this study are freely accessible via a custom web-based browser that enables easy querying and visualization of all the data generated (https://computationalgenomics.ca/tools/epivar).

Results

Transcriptional and epigenetic response to influenza infection

We infected monocyte-derived macrophages (MDMs) derived from a diverse panel of 35 healthy males with influenza A virus (IAV), commonly known as flu. We focused on macrophages as they are the primary source of type I interferon (IFN) and pro-inflammatory cytokines during flu infection, and therefore play a central role in viral clearance and the regulation of the pathology during infection[83–85]. Following 24-hours of flu infection, we collected from matched non-infected (NI) and infected samples data on *(i)* gene expression (RNA sequencing), *(ii)* chromatin accessibility (assay for Transposase-Accessible Chromatin using sequencing; ATAC-seq), *(iii)* levels of histone marks associated with promoters (H3K4 trimethylation, or H3K4me3), enhancers (H3K4 monomethylation, or H3K4me1), and their activation levels (H3K27 acetylation, or H3K27ac), as well as a general repressive mark (H3K27 trimethylation, or H3K27me3), and *(iv)* methylation levels (as measured by whole genome bisulfite sequencing; WGBS) (Fig. 2-1A). The 24 hour time point was chosen due to previous work showing that chromatin accessibility and DNA methylation changes in response to infection are mostly detectable at late time points post-infection[41]. In addition, we identified genetic variants for each individual using high-coverage (30X) whole genome sequencing. In total, we obtained over 211 billion reads across the different assays, generating the most extensive dataset to date of the combined transcriptional and epigenetic response to flu across individuals of different genetic ancestries (Table S2-1). All assay-specific quality control metrics, including percentage of mapped reads, number of CpG sites covered per sample, or the fraction of all mapped reads that fall into the called peak regions (i.e., FRIP scores) indicate that the data is of high quality (Table S2-1).

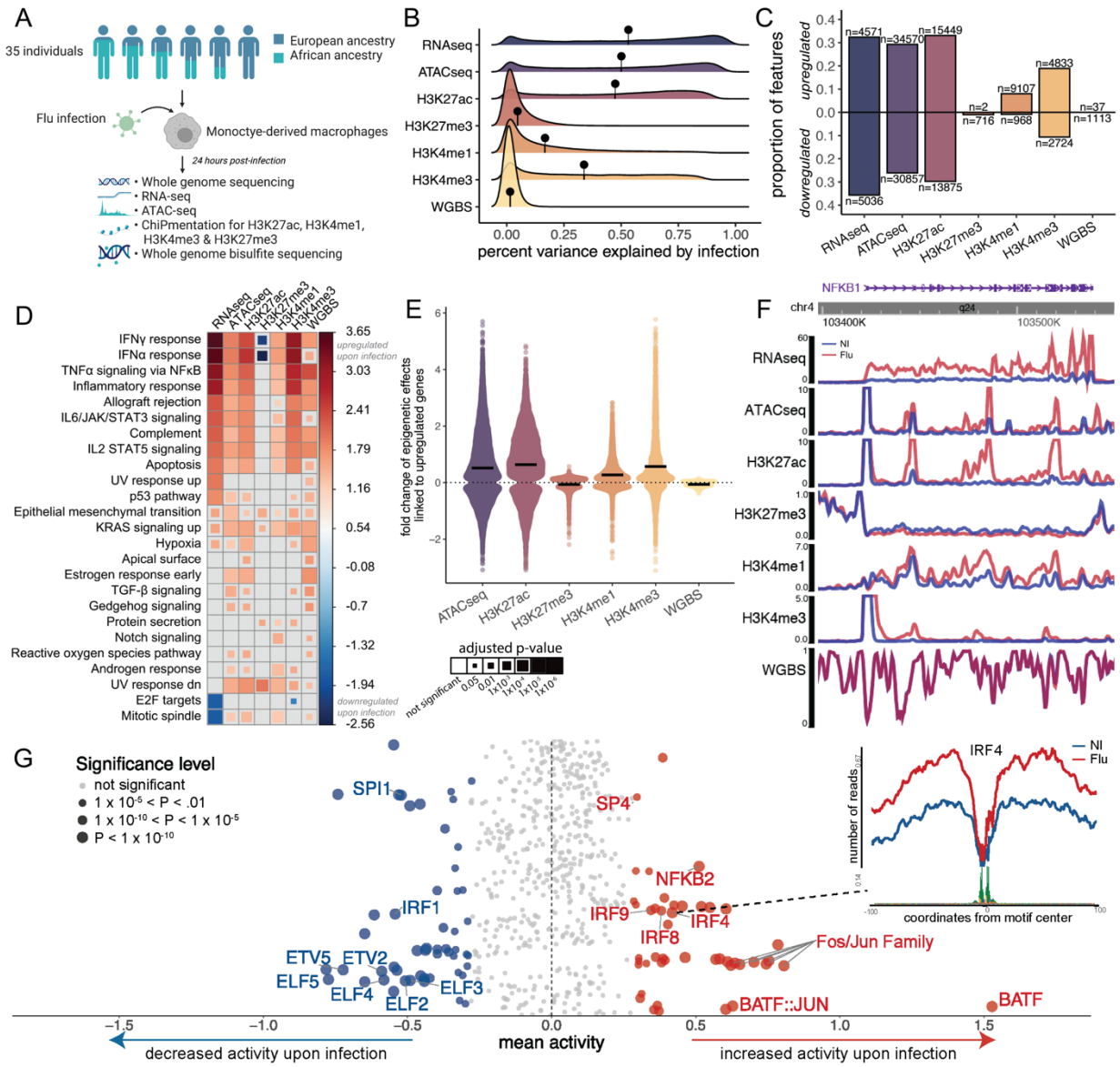


Figure 2-1. Flu infection remodels the epigenetic landscape of human macrophages. (A) Study design schematic. Monocyte-derived macrophages from 35 individuals were exposed to influenza A virus or media, for 24 hours. DNA collection and libraries for 7 types of regulatory marks were prepared and sequenced. Figure was created using BioRender.com (B) The distribution of the percent variance explained by infection for each feature in each data type. The mean is represented by the black lollipop. For comparison, the average PVE when comparing mock and NI samples ranged only from 1.6-2.7% (Fig. S2-1B). (C) Proportion and number of features significantly upregulated and downregulated in response to flu infection (FDR<.10, beta = ± 0.1 for WGBS and ± 0.5 for all other data types). (D) Hallmark pathways enriched among genes nearby epigenetic features (or the actual gene in case of gene expression) that respond to flu infection. Pathways shown have Benjamini-Hochberg adjusted $P < 0.001$ in at least 1 data type and a |normalized enrichment score| > 1.5 in at least 2 data types. Red marks upregulation for that data type and blue downregulation. (E) Distribution depicting the relationship between gene expression changes and epigenetic changes in response to flu infection. Mean is represented by the

Figure 2-1, continued. black line. Upregulated genes are defined as genes with beta > 0.5 and FDR<.01. Epigenetic changes are those with FDR<.01, with the exception of methylation changes for which we use a less stringent threshold (FDR<.20) due to the relatively smaller number of changes. A similar plot for downregulated genes can be found in Fig. S2-1B. (F) The region surrounding *NFKB1*, an example of a region where gene expression and epigenetic changes occur in a coordinated fashion. (G) Transcription factor activity changes after flu infection sorted based on alphabetical order of TF names (y-axis). Upper right plot shows an example of a footprint centered on theIRF4 motif. The footprint is stronger in the flu-infected condition indicating higher levels of IRF4 activity after flu infection. Mean activity across the samples (x-axis) is plotted in the main plot. The size of the dots reflects significance levels.

We first investigated the impact of flu infection across the different data types. For DNA methylation, because most CpG sites are fully methylated and static[57], we focused uniquely on CpG sites overlapping putative regulatory elements as identified by the chromatin segmentation program ChromHMM[86] (~7.3 million CpG sites out of a total of 19.5 million surveyed across the genome were used in all downstream analyses). Principal component analysis (PCA) on the matrices of gene expression and peak intensities revealed a strong infection effect, with NI and flu samples consistently separating on either PC1 or PC2 for most datasets (Fig. S2-1A). Such separation was not observed for DNA methylation or H3K27me3. To quantify the impact of flu infection on each of the molecular traits, we calculated the percent of variance explained (PVE) by infection for each feature in each data type. PVE by flu infection was highest for gene expression, chromatin accessibility, and H3K27ac histone modifications (average PVE ranging from 53-47%), followed by changes in H3K4me3 (34%) and H3K4me1 levels (17%). The least dynamic response to flu was observed for repressive marks, H3K27me3 and DNA methylation, with average PVE values across all features tested of only 5% and 2%, respectively (Fig. 2-1B), not far from the estimated PVE due to technical variation alone (Fig. S2-1B). Consistent with the PVE analyses, 68% of all genes tested (n=9,607) were found to be differently expressed in response to flu infection (FDR<0.10 with fold change >|0.5|), Fig. 2-1C, Table S2-2). Despite

differences in cell type (monocyte vs macrophages), technologies (scRNA-seq vs bulk), and time points (6 hours vs 24 hours post infection*)* the influenza-infection effects on gene expression we identified are strongly correlated with those previously reported[40] (R=.63, $P < 2.2e-16$, Fig. S2-1C). At the epigenetic level, 55% (n=65,427) and 63% (n=29,324) of regions tested changed chromatin accessibility and H3K27ac levels, respectively, in contrast to less than 0.02% for methylation and 1% of H3K27me3 levels (FDR<0.10 with fold change >|0.5|) for all data types or >|0.1| for WGBS). We also see a bias in the direction of the infection effects across the data types: repressive marks (H3K27me3 and DNA methylation) tend to be downregulated in response to flu infection (>97% of all significant changes are associated with H3K27me3 and DNA methylation losses) whereas marks associated with active enhancer and promoter regions (H3K4me1 and H3K4me3) are primarily upregulated (64-90%).

Consistent with previous work[87,88], we found that genes upregulated in response to flu infection are strongly enriched for gene sets involved in interferon α and γ responses as well as the activation of inflammatory responses (normalized enrichment score (NES)>3; FDR<$1x10^{-6}$; Fig. 2-1D). Our data shows that epigenetic changes in response to flu converge to the same pathways, indicating that transcriptional and epigenetic changes act in a coordinated manner to establish the immune regulatory networks required for the host response to flu infection. To further investigate the relationship between gene expression changes and epigenetic changes in response to flu infection, we asked how epigenetic features nearby genes that are up- or downregulated in response to infection respond. Regulatory elements nearby upregulated genes (Fig. 2-1E) show, on average, increased opening of chromatin, increased activation marks in enhancer and promoters, and a reduction of repressive marks (P<$7.147x10^{-6}$ for H3K27me3; P<$2.2x10^{-16}$ for all other data types; Fig. 2-1F for an example at the *NFKB1* locus). In contrast, regulatory elements near downregulated

genes tend to be associated with closing of chromatin and the loss of activation marks (Fig. S2-1D).

To investigate the role played by transcription factors (TF) to the epigenetic changes identified in response to IAV infection we used TF footprinting to compute TF activity scores (Fig. 2-1G, Table S2-3). TF footprinting characterizes regions where TFs are likely bound based on chromatin accessibility patterns at known TF motifs. We were particularly interest in TFs which activity levels change between NI and flu-infected samples (Fig. 2-1G inlet). We find that many immune-related TFs, such as those in the Fos/Jun family and Interferon Regulatory Factors (e.g., IRF4, IRF8 and IRF9), significantly increase activity after infection ($P<1x10^{-5}$). Of note, we find that several ETS family members are downregulated in response to flu infection, which is concordant with ETV7 acting as a negative regulator of the type I IFN response[89,90]. Unexpectedly, BATF, which is not a classical macrophage-response TF but can act as a pioneer TF in other cell types (often requiring dimerization with Jun (BATF:JUN)[91]), showed the greatest increase in activity upon infection. Our results thus suggest that BATF likely plays a previously unappreciated role in the macrophage response to flu infection, paralleling its already established role in the induction of effector programs and epigenetic landscape of $CD8^+$ T cells and innate lymphoid cells infected with flu[92–94]. Collectively, our results show that transcriptional and epigenetic changes in response to flu infection are highly coordinated and likely driven by the activation of infection-induced TFs involved in the regulation of antiviral responses.

Ancestry-associated differences across transcriptional and epigenetic responses to influenza infection

We next investigated ancestry-associated differences across the data types. To do so, we used the genotype data to estimate genome-wide levels of European and African ancestry in each sample using ADMIXTURE (v1.3.0)[95]. Consistent with previous reports[96], we found that self-identified African American (AF) individuals have a high proportion of European ancestry (mean = 28%, range 13%–57%; Fig. S2-2A). In contrast, self-identified European Americans (EU) showed virtually undetectable levels of African admixture (mean = 0.05%, range 0.001%–0.69%; Fig. S2-2A). We confirmed that K=2 is the most appropriate K for our study group, as it shows the lowest cross validation error (Fig. S2-2B). Performing ADMIXTURE analysis with larger K values further breaks the estimated African ancestry components into different sub-components but does not affect the overall proportion of the genomes that are estimated as being *"African-associated"* (Fig. S2-2C). In all downstream analyses, African ancestry level was used as a continuous variable unless otherwise noted.

We first identified genes/regions where gene expression, accessibility, histone changes, or methylation are correlated with quantitative genetic ancestry estimates at baseline, after flu infection, or both. We termed these genes/regions as "population differentially expressed" (popDE). Combining both FDR and a multivariate adaptive shrinkage method (mash)[97,98], we identified both shared and condition-specific popDE features for each data type (conservatively defined as genes/regions significant at a FDR<10% & local false sign rate (lfsr) < 10%) (Fig. 2-2A, Table S2-4). Mash leverages the correlation structure across conditions increasing statistical power and enabling the detection of shared popDE effects. We found that gene expression and H3K4me1 levels show the largest proportion of significant differences between ancestry groups

were the most divergent between ancestry groups – 23% of genes and 21% of H3K4me1 peaks tested were classified as popDE across infected and non-infected macrophages. In contrast, only 1% of promoter-associated H3K4me3 peaks were classified as popDE at the same significance thresholds, suggesting that ancestry-associated differences in gene expression are primarily driven by variation in enhancer activity as opposed to variation at the level of core promoters. We also performed popDE analysis with local ancestry estimates calculated using RFMix. We find that global versus local popDE effect sizes are strongly correlated ($R \geq 0.92$, Fig. S2-3A) across marks, which is probably due to the fact that across most regions of the genome local and global ancestry estimates are very strongly correlated given that our European-ancestry group shows negligible levels of African admixture (i.e., for these individuals local or global ancestry estimates of African ancestry will always be zero). As with flu infection effects, we found that popDE effects at the transcriptional and epigenetic level were highly coordinated: genes more highly expressed in individuals with a greater proportion of African ancestry were linked to epigenetic changes indicative of increased transcriptional activity in African- as compared to European-ancestry individuals, including increased chromatin accessibility, histone acetylation levels (H3K27ac), as well as mono- and tri-methylation of H3K4 (Fig. 2-2B, Fig. S2-3B).

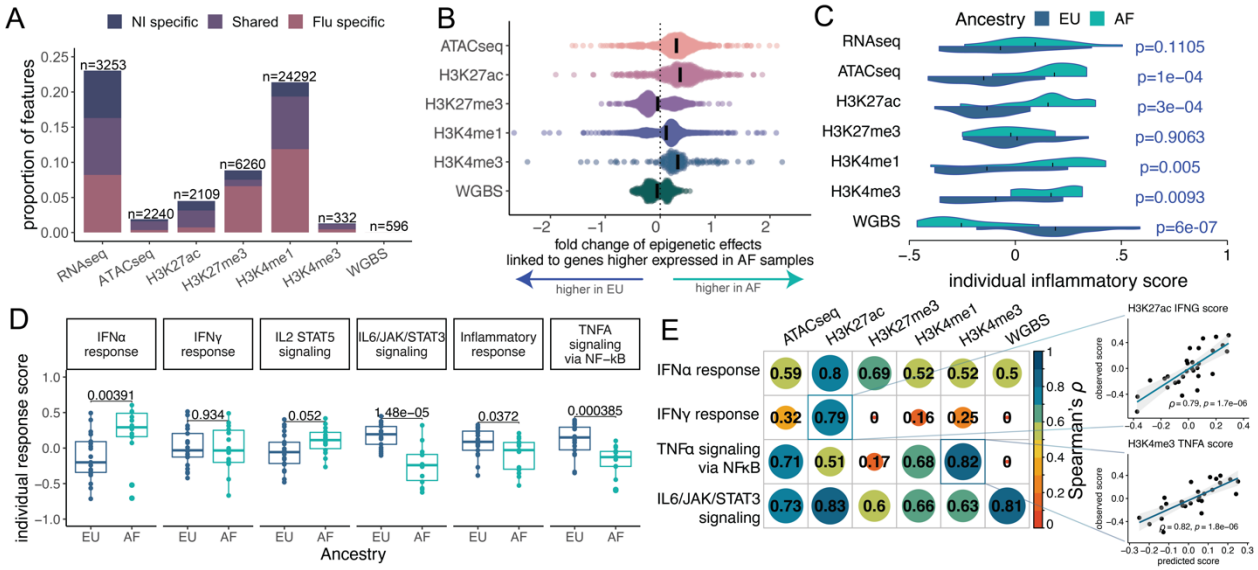


Figure 2-2. Ancestry-associated differences in the gene regulatory response to flu infection. (A) Proportion and number of popDE features that are either condition-specific (FDR<.10 and lfsr<.10 in only one condition) or shared (FDR<.10 and lfsr<.10 in both conditions). (B) Distribution depicting the relationship between popDE genes and popDE epigenetic changes across both conditions. Genes more highly expressed in individuals with high proportions of African ancestry (fold change> 0.5, FDR< 0.10) are nearby popDE epigenetic regions showing increased levels of chromatin accessibility, H3K27ac, H3K4me1 and H3K4me3 in individuals with increased African ancestry. Black lines represent means. (C) Distributions of individual mean score for the Hallmark "inflammatory pathway" in the non-infected condition. A higher score value indicates a stronger expression of genes or epigenetic marks nearby genes within this inflammatory response pathway. *P* values were calculated using a Wilcoxon rank sum test. (D) Boxplots of individual transcriptional response scores for 6 immune response pathways. Pathway response levels were measured as the difference in the per individual pathways' score between the flu-infected and non-infected conditions. (E) Spearman's correlation of observed and predicted transcriptional response scores.

Inflammation levels have consistently been shown to vary between individuals of European and African ancestry, with an overall tendency for higher inflammation in individuals with increased African ancestry[21,23,35]. To evaluate if differences in inflammation could result from baseline epigenetic differences between ancestry groups, we computed a per-sample score of inflammatory activity, the "inflammation score", which provides an estimate of the average expression or peak height of all features (genes or peaks nearby genes) in the Hallmark

inflammatory response pathway[99]. Consistent with previous reports, we found a clear trend towards higher "inflammation score" at the gene expression level in African ancestry individuals relative to Europeans in the non-infected condition (1.5-fold, albeit non-significant, P=0.11, Fig. 2-2C). More strikingly, we also found an epigenetic signature of higher inflammation in individuals of African-ancestry, relative to European-ancestry individuals. Specifically, we found that increased levels of African-ancestry were strongly associated with increased levels of chromatin accessibility (P=1x10^{-4}), H3K27ac (P=3x10^{-4}), H3K4me1 (P=5x10^{-3}), H3K4me3 (P=9x10^{-3}) as well as lower levels of CpG methylation (P=6x10^{-7}) nearby genes involved in the regulation of inflammatory responses (Fig. 2-2C (NI), see Fig. S2-3C for similar effects in the flu-infected condition).

Our dataset provides a unique opportunity to evaluate if baseline differences in epigenetic landscape contribute to ancestry-associated differences in transcriptional response to flu. To test such hypothesis, we started by characterizing genes for which the gene expression response to infection (i.e., individual-based fold-change) significantly correlated with genetic ancestry (hereafter referred to as population differently responsive genes; or popDR). We found 2,149 popDR genes (FDR<0.20; Table S2-4), reinforcing the notion that genetic ancestry has a marked impact on the transcriptional response to flu[21,35,40]. Focusing on this set of popDR genes and on a curated set of immune pathways known to be involved in anti-viral responses[99], we found that (at 24 hours post-infection) individuals with higher proportions of African ancestry show a significantly stronger IFN-α response (P=0.004) and weaker IL6/JAK/STAT3 (P=1.5x10^{-5}), TNFα (P=3.9x10-4) and inflammatory (P=0.0372) responses relative to EU individuals (Fig. 2-2D). These findings are robust to stricter FDR thresholds (Fig. S2-3C, S2-3D, S2-3E).

To evaluate if the observed differences in gene expression responses between European and African-ancestry individuals could stem from baseline differences in epigenetic profiles, we then used elastic net regression to assess the predictive power of baseline (non-infected) epigenetic levels to the transcriptional responses of the pathways described above. We found that the response to all pathways tested could be predicted with high accuracy (Spearman's $\rho \geq 0.79$, $P \leq 2\text{x}10^{-6}$) by the baseline levels of at least one epigenetic mark. Across the different marks, baseline levels of H3K27ac showed the most consistent predictive value across all the pathways (ρ range: 0.51 to 0.83, $P \leq 5\text{x}10^{-3}$, Fig. 2-2E, Fig. S2-3F). These results support the idea that an individual's gene expression changes in response to flu infection are, at least in part, driven by the epigenetic landscape of the genome surrounding the gene prior to infection.

Single nucleotide polymorphisms and short tandem repeats independently drive differences in regulatory marks

To evaluate the contribution of genetic variation to ancestry-associated differences, we mapped genetic variants that are associated with variation in gene expression or epigenetic marks across individuals (i.e., quantitative trait loci (QTL); hereafter we will refer to the mapping of the different molecular traits as the following: gene expression (eQTL), chromatin accessibility (caQTL), H3K4me1 (K4me1QTL), H3K4me3 (K4me3QTL), H3K27ac (K27acQTL), H3K27me3 (K27me3QTL), and methylation (meQTL)). To map QTL, we used a linear regression model that accounts for population structure and principal components of the expression data, thus limiting the effect of unknown confounding factors (see methods for details). Given that our sample size is too small to robustly detect trans-acting QTL, we focused our analyses on local associations that, for simplicity, we refer to as cis-QTL, defined as variants located within a gene body/peak or in

the 100 kb flanking the gene/peak of interest. For methylation levels, the window was limited to ±5kb from the CpG site being tested[100,101]. Power calculations indicate that we have reasonable power (>80% for SNPs with MAF>10%) to detect *cis*-QTLs of effect sizes equal or larger to 0.3 (equivalent to ~20% change in the molecular trait being measured per alternative allele) but that for smaller effect sizes our power is limited (Fig. S2-4A).

We leveraged our deep whole-genome sequencing data to obtain genetic information not only on single nucleotide polymorphisms (SNPs) but also on short tandem repeats (STRs), which constitute one of the most polymorphic and abundant types of repetitive elements in the human genome[102,103]. Across individuals, we identified approximately 7.38 million SNPs and 440,000 STRs with a minor allele frequency above 5% for SNPs and 10% for STRs, which were used for QTL mapping. Despite our limited power, we identified at least one cis-eQTL (FDR<10%) for 3,880 genes (eGenes) across one or more conditions (28% of all genes tested, Fig. 2-3A, Fig. S2-5A, S2-5B, S2-5C, Table S2-5). Among epigenetic marks, H3K4me1 was associated with the largest proportion of QTL (18.5% of all peaks tested, FDR <10%), followed by chromatin accessibility (20%) and H3K4me3 (11.6%) (Fig. 2-3A, Fig. S2-5C). For methylation levels, we found over 43,182 CpG sites associated with at least one meQTL, but, given the large number of CpG sites tested (over 7 million), the relative number of associations was the smallest of all epigenetic marks. Importantly, the estimated effect sizes of QTL identified in our data set are very strongly correlated ($0.49 < R > 0.92$, all $P< 2.2\text{x}10^{-16}$) with those derived from these previously published datasets in non-infected cells, attesting for the robustness of our QTL results (Fig. S2-4B)[21,40,61,64]. In addition, we found that across all molecular traits tested, features with *cis*-QTL were significantly enriched for features with allele-specific expression (ASE), compared to the

background of all genes/features tested (Fig. S2-4C, $P<2.2x10^{-16}$ for all conditions), further validating our genetic associations.

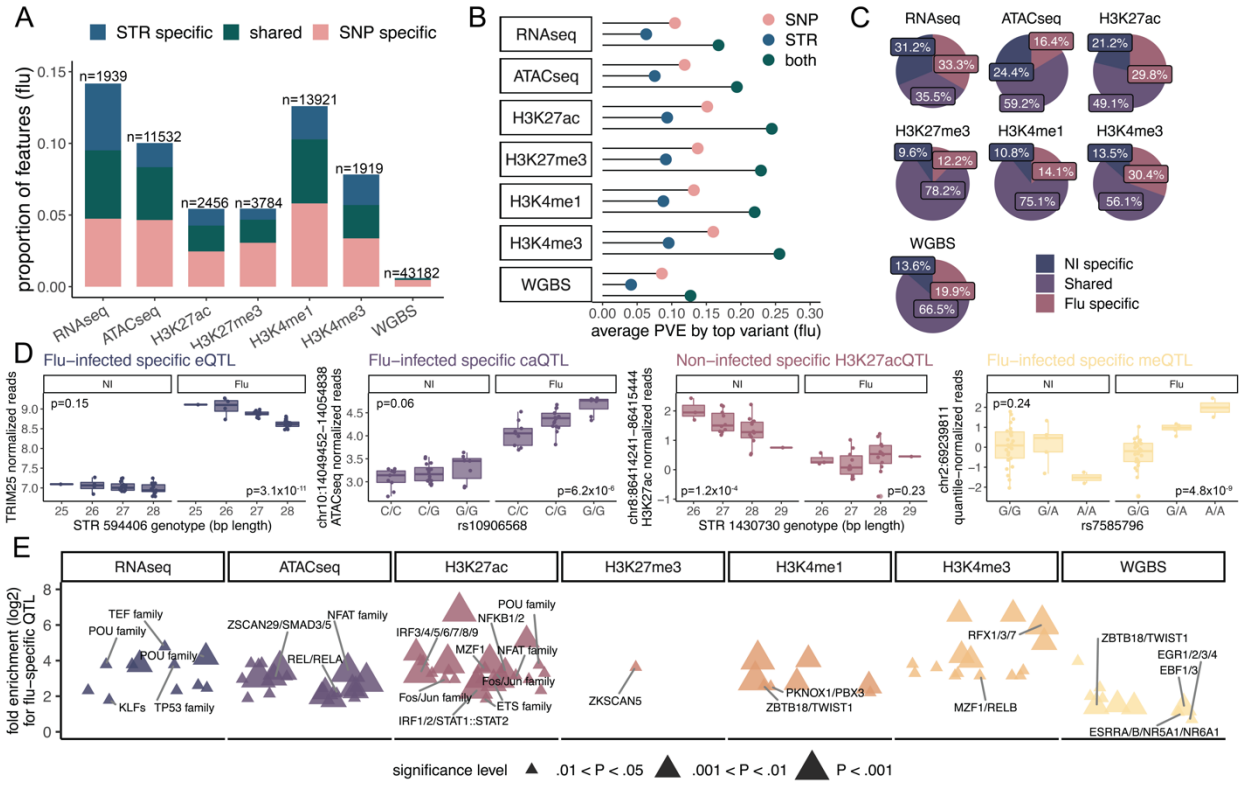


Figure 2-3. Cis-regulatory variation drives ancestry-associated differences in the transcriptional and epigenetic response to flu infection. (A) Proportion and number of genes/features associated with at least one SNP or STR QTL (in flu-infected samples, see Fig. S2-5C for the non-infected samples). Shared QTL were defined as those genes/features associated with a QTL at an FDR<.10 when performing the QTL mapping against SNPs and STRs separately. SNP- or STR-specific are those only identified as significant (FDR<.10) against either SNPs or STRs (B) The mean percent variance explained by the top SNP and STR across all features in the flu-infected condition. Both is the sum of the PVE of the top SNP and top STR (C) Pie charts showing the percentage of condition specific (FDR<.10 and lfsr <.10 in only one condition with either SNP or STR) and shared QTL (FDR<.10 and lfsr <.10 in both conditions with either SNP or STR) across the data types. (D) *Far left* - An example of a flu-infected specific STR-eQTL. *Middle left* - An example of a flu-infected specific SNP-caQTL. *Middle right* - An example of a non-infected specific STR- K27acQTL. *Far right* - An example of a flu-infected specific meQTL. (E) The enrichment of TF binding sites across flu-infected specific SNP-QTL. Immune-related TF cluster names are highlighted.

Strikingly, across all molecular traits tested, 5-33% of features with a cis-QTL were only identified through their association with STRs and not SNPs (Fig. 2-3A, Fig. S2-5C). We next used variance partitioning to disentangle the relative contribution of STRs and SNPs to variation in gene expression and epigenetic marks (Fig. 2-3B, Fig. S2-5D). Across molecular traits, STRs contribute, on average, to 4-10% of the *cis* heritability among genes/peaks associated with at least one QTL in infected cells, not far from the amount of variance independently explained by SNPs (9-16%). Similar results were found for non-infected cells (Fig. S2-5D). Thus, our findings highlight the unique contribution of STRs to the genetic architecture of human gene regulation.

We found that a large fraction of eGenes (33.3%) were only detectable in the infected condition (Fig. 2-3C, Fig. 2-3D), further reinforcing the pervasive role of gene by environment (GxE) interactions on human gene expression[21,35,70,104,105]. Our epigenetic data expands on previous work on gene expression by showing that that GxE interactions are ubiquitous across the entire gene regulatory landscape, and not only transcription: across epigenetic marks, 12.2-30.4% of peaks (or CpG sites) showed an infection-specific QTL. We hypothesized that one potential mechanism accounting for infection-specific QTLs is that the causal SNPs/STRs disrupt the binding site of TFs that become more active in response to flu infection. To test this hypothesis, we investigated if infection-specific QTL were significantly enriched for TF footprints (non-infected specific results can be found in Fig. S2-5E). We found that cis-eQTL and cis-epigenetic QTL detected only in infected cells were markedly enriched for a diverse array of immune-activating TF footprints (e.g., IRF and NfK-B family members; Fig. 2-3E, Table S2-3), suggesting that many infection-specific QTL are likely to be driven by the differential binding of infection-induced TFs.

QTL are shared across regulatory marks

To further investigate the connection between genetically regulated variation in epigenetic marks and gene expression levels, we tested if SNPs that are QTL in one data type are also QTL for the other data types. Briefly, for each condition, we took all significant SNPs (FDR <.10, Fig. S2-5A) for a feature and collected their corresponding p-value for all additional features. We used a cutoff of FDR<.10 to define if the SNP is also a significant QTL in other data types. We find striking patterns of sharing across the data types. For example, in non-infected macrophages, on average, 60% of QTL identified in one data type are shared with at least one other data type (ranging from minimum 45% for meQTL and at maximum 76% for K27acQTL (Fig. 2-4A), compared to a null expectation of only 1.8% (Fig. S2-6A) when permuting the data (see methods for details). We find a similar pattern for the QTL identified in the flu-infected condition (Fig. S2-6A, Fig. S2-5A). We note our approach assumes a null hypothesis of non-sharing and estimates of sharing across the data types are likely conservative. Relaxing the FDR threshold for sharing increases the number of overlaps, but also the null expectation.

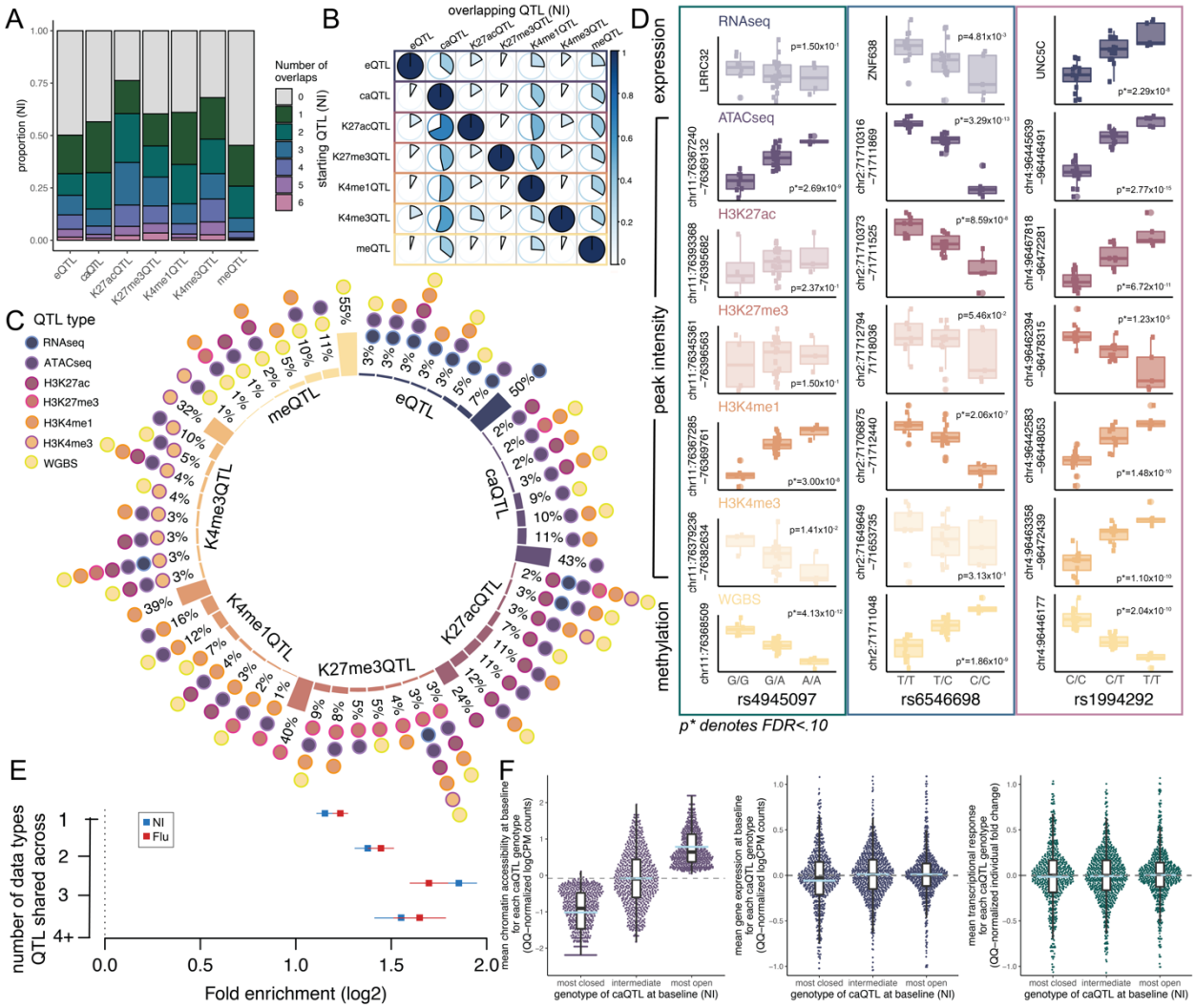


Figure 2-4. Overlap of regulatory QTLs along the cascade of gene regulatory elements. (A) The number of overlaps for each QTL type in the non-infected condition. In this figure, one (dark green) means that the QTL is only a QTL for that datatype alone. More than one overlap means that the QTL is shared with at least one other datatype, with 6 referring to cases where the QTL is shared across all datatypes. (B) The percentage of QTL in one data type that are also QTL for another data type at baseline (NI condition). The starting QTL (rows) are the QTL that are tested for sharing while the overlapping QTL (columns) are the percentage of each starting QTL that are shared with that datatype. For example, 36% of eQTL are also caQTL (row 1), while only 8% of caQTL are also eQTL (row 2). The color of each circle corresponds to the percentage of sharing. (C) The top patterns for QTL integration for each data type at baseline (NI condition). The size of the bar represents the percentage of significant QTL (FDR<.10) that share the pattern reported by the dots. (D) Examples of SNPs that are shared QTL. Grayed out plots are molecular QTL that are not significant at FDR<.10. Left: rs4945097 is a QTL for chromatin accessibility, H3K4me1 and methylation. Center: rs6546698 is a QTL for chromatin accessibility, H3K27ac, H3K4me1 and methylation. Right: rs1994292 is a QTL for all 7 data types. Notably, the T allele is associated with higher expression of *UNC5C* and epigenetic marks indicative of activated regions of the genome. (E) QTL enrichments (x axis) in actively regulated TF binding sites annotated by ATAC-

Figure 2-4, continued. seq footprinting. Error bars show 95% confidence intervals. QTLs that are shared across multiple data types are more likely to be enriched among TF footprints. (F) Association between genetically encoded baseline differences in chromatin accessibility and the magnitude of transcriptional response to IAV infection. *Left*- Meta caQTL plot at baseline condition across all caQTLs identified. For each caQTL locus, individuals were binned based on their genotype: homozygous for the genotype associated with more closed chromatin (most closed), heterozygous (intermediate), or homozygous for the allele associated with increased chromatin accessibility (most open). The light blue line marks the mean for each genotype and the gray dotted line is the median across all genotypes. We focused specifically on caQTLs nearby upregulated genes (n= 681 caQTLs associated with 506 genes) and that did not impact baseline expression levels (as shown in the middle plot). *Right*- Genotypes for chromatin accessibility levels at baseline have no impact on the transcriptional response of nearby genes.

We found consistent sharing patterns across each of the data types (Fig. 2-4B (non-infected), Fig. S2-6B (flu), Table S2-6). For example, in the non-infected condition, approximately 36% of eQTL are also caQTL. In fact, caQTL are the most commonly shared QTL type for all other data types, such that genetic variants impacting gene expression, a histone mark, or methylation will ~50% of the time (range 36%-69% in the non-infected condition) also be associated with changes in chromatin accessibility (Fig. 2-4B, Fig. S2-6C). When considering genetic variants that are shared across three or more molecular traits, the most common pattern is sharing between caQTL, K4me1QTL, and meQTL (Fig. 2-4C, Fig. S2-6D), suggesting a high-level of co-regulation of chromatin accessibility, H3K4me1, and methylation levels at enhancers elements. An example of this most common pattern is rs4945097 which is a QTL for chromatin accessibility, H3K4me1 and methylation (Fig. 2-4D Left). The genetic variant rs1994292 is a QTL for all 7 molecular traits (Fig. 2-4D Right). Interestingly, QTL impacting multiple regulatory marks are also more likely to overlap TF footprints (Fig. 2-4E), supporting the idea that TFs are the primary mediators of genetically-driven variation in gene regulatory programs[44,79].

We hypothesized that some of the shared QTLs effects might be associated differential activation of enhancer RNAs (eRNAs). Of the putative eRNAs identified (n=17,142), 88% are

differentially expressed in response to flu infection (FDR 10%, beta ± 0.5), the majority (70%) being up-regulated (Table S2-2). As with the other molecular traits, we mapped SNPs associated with the eRNAs, finding 11% (n= 1,889) associated with at least one eRNA-QTL (Table S2-5). Interestingly, eRNA-QTLs were found to be highly enriched for other regulatory QTL (range 6-130 fold, all $P<2.2x10^{-16}$, Fig. S2-6E, Table S2-7). The 130-fold enrichment of gene expression eQTL among variants classified as eRNA-QTLs in flu-infected samples raises the possibility that some of the eQTL are driven by variation in the activity of eRNA engaged upon infection.

It is commonly believed that increased chromatin accessibility primes immune cells to respond faster and stronger to an immune challenge or infection[52,106], but the data supporting such a model remains circumstantial. We sought to use our genetic and epigenetic data to test this model. To do so, we focused on caQTL found in non-infected cells and their associated genes (i.e., the closest coding gene to the caQTL) and asked whether, across the genome, individuals homozygous for the genotype associated with more open chromatin showed a stronger transcriptional response as compared to individuals heterozygous or homozygous for the alleles associated with reduced opening. We limited the analyses to genetically regulated accessibility peaks associated with genes that are upregulated in response to flu infection, and that are not concomitantly eQTLs to avoid the confounding effect of baseline differences in gene expression to variation in transcriptional responses. We found that genetically driven variation in chromatin accessibility levels had no impact on the magnitude of transcriptional responses upon IAV infection (Fig. 2-4F, see Fig. S2-7A for similar results when conditioning for genetically driven variation for other epigenetic marks). This is surprising given that increased levels of open chromatin were also associated with baseline increased levels of H3K27ac and H3K4me1 (Fig. S2-7B), all of which have been postulated as “priming marks” for a stronger transcriptional response to infection. Overall, these

data suggest that the relationship between baseline chromatin accessibility levels and transcriptional response to infectious agents is more complex than generally believed.

Cis-regulatory variation explains ancestry associated differences to varying extents across the regulatory marks

We next sought to examine the connection between regulatory QTL and ancestry-associated differences in gene expression and epigenetic profiles. Consistent with an important role for genetics in ancestry-associated differences in the gene regulatory landscape, we found that genes/peaks associated with regulatory QTLs were more likely to be classified as popDE than expected by chance (Fig. 2-5A). Interestingly, the strongest enrichments were observed for epigenetic marks, especially DNA methylation, for which we observed that CpG sites with a meQTL were >28-fold more likely to be classified as popDE than those without ($P<2x10^{-16}$; Fig. 2-5A). In contrast, the enrichment of popDE genes among genes with an eQTL, albeit significant ($P<3.0x10^{-4}$), was only 1.2 fold (in NI, 1.5 fold in flu), suggesting a much greater contribution of genetics to epigenetic variation across populations compared to gene expression.

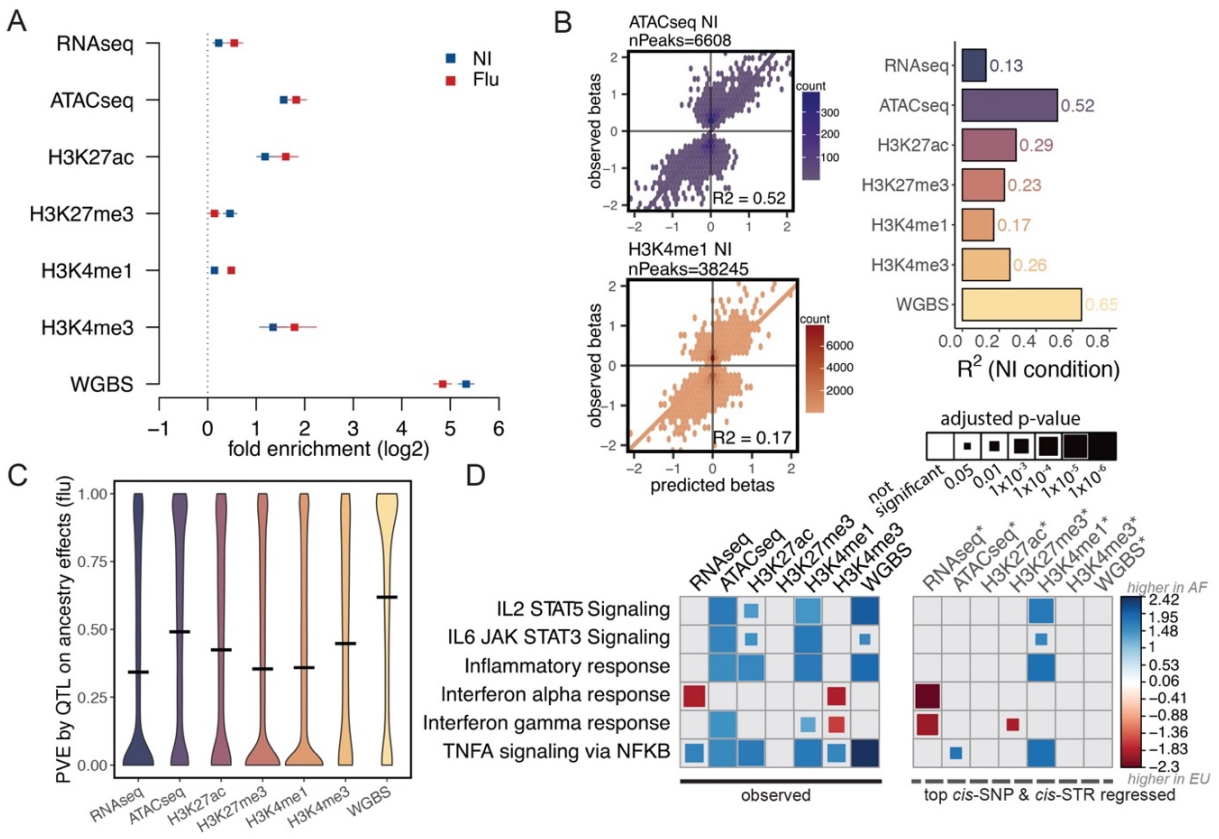


Figure 2-5. Cis-regulatory variation contributes to ancestry-associated differences. (A) The enrichment of QTL in popDE features across the data types. Log2 fold enrichments and a 95% confidence interval are plotted. (B) (Left) Examples of the correlation between the observed and predicted betas for popDE features (FDR<.10, Pearson's correlation coefficient reported). (Right) Bar plot summarizing the correlation between observed and predicted betas for popDE features (FDR<.10) across all marks in the non-infected condition. (C) Violin plot of the percent variance explained by the top SNP- and STR-QTL on ancestry effects for each feature in each data type in the flu-infected condition. Median PVE indicated by the black line. (D) Gene set enrichment using the popDE results originally and after regressing out the top cis-SNP and cis-STR for each feature. Immune-related pathways from the Hallmark gene sets are shown. Blue indicates that the genes or features associated with genes in the pathway are more highly expressed in individuals with high levels of African ancestry. Red indicates increased expression in individuals of primarily European ancestry.

These enrichments suggest that ancestry-associated differences in gene expression are likely to be explained, at least in part, by population differences in allele frequencies at QTLs. To test this hypothesis, we calculated, for each of the molecular traits, the correlation between estimated and predicted genetic ancestry effects. The estimated values were obtained from our

popDE analysis whereas the predicted effects were based on the effect size of the top SNP/STR for each feature and the dosage genotype for those variants across individuals (restricted to features with popDE effects). Differences in the genotype distribution between ancestry groups for the best SNP and STR explain up to 65% of the variance in genetic ancestry effect sizes across molecular traits (Fig. 2-5B, Fig. S2-8A). The strongest genetic contributions were observed for CpG site methylation and chromatin accessibility, for which the genotype of the best SNP and STR explain 65% and 52% of the variance in ancestry-associated differences in methylation and chromatin accessibility, respectively. Conversely, only 13% of the variance in gene expression differences is explained by the top SNP and STR, suggesting an important contribution of additional cis-regulatory variants, trans-regulatory variants, or environmental factors. We also calculated the change in the percent variance explained by genetic ancestry before and after regressing out the top SNP and STR. We found an analogous pattern: the lead STR and SNP plays a more significant role in explaining population differences for epigenetic marks than for gene expression (average of 62% and 49% for methylation and ATAC-seq, respectively, versus only 34% for RNA-seq, Fig. 2-5C, Fig. S2-8B (NI)).

To determine if the ancestry-associated differences in immune-related pathway activity we observed (Fig. 2-2C) remain significant after removing the effect of the best SNP and STR, we performed gene set enrichment analysis and compared the enrichments both before and after removing the top genetic effects (Fig. 2-5D). For the epigenetic effects (with the exception of H3K4me1), any baseline significant enrichment is reduced or eliminated, indicating that the top SNP and STR are important contributors to the differences in pathway activity detected between ancestry groups. For example, the observed enrichments of open chromatin and H3K27 acetylation levels near genes involved in inflammatory responses among individuals with increased African

ancestry (FDR<$1x10^{-5}$) completely disappeared (FDR>0.5) when the QTL effects were regressed out. Accounting for cis-acting genetic effects is also enough to eliminate the transcriptional differences in inflammatory response to IAV infection identified between individuals of European and African ancestry ($P_{original}$ = 0.03 (Fig. 2-2D) ; $P_{cis\text{-}regressed}$ = 0.434 (Fig. S2-8C)). In sharp contrast, the ancestry-associated differences in type-I interferon response remain unaltered when regressing out the effects of *cis* eQTL ($P_{original}$ = 0.004 (Fig. 2-2D); $P_{cis\text{-}regressed}$ = 0.006 (Fig. S2-8C)), suggesting that the ancestry-associated differences in interferon signaling are likely to be driven by environmental differences that correlate with genetic ancestry rather than by *cis* genetic variation.

Epigenetic variants provide insight into immune-related disease risk

To evaluate the impact of regulatory QTL on susceptibility to immune-related disorders, we first assessed the colocalization[107,108] between regulatory QTL hits and 14 publicly available genome-wide association study (GWAS) hits for 11 immune-related diseases. For each trait, we identified the lead GWAS SNPs with p-values below $1x10^{-5}$ and defined a "locus" as a 100kb (5 kb window for methylation QTL) centered around the lead GWAS SNP, removing the HLA region from the analysis. We find that many epigenetic variants colocalize with variants implicated in immune-related traits (Fig. 2-6A, Fig. S2-9A, Table S2-8), most of which would have been missed when considering eQTL alone. Indeed, across all colocalized variants, only 7% were eQTL, the remaining corresponding to genetic variants that impact one or more epigenetic marks but not gene expression levels (e.g., Fig. 2-6B).

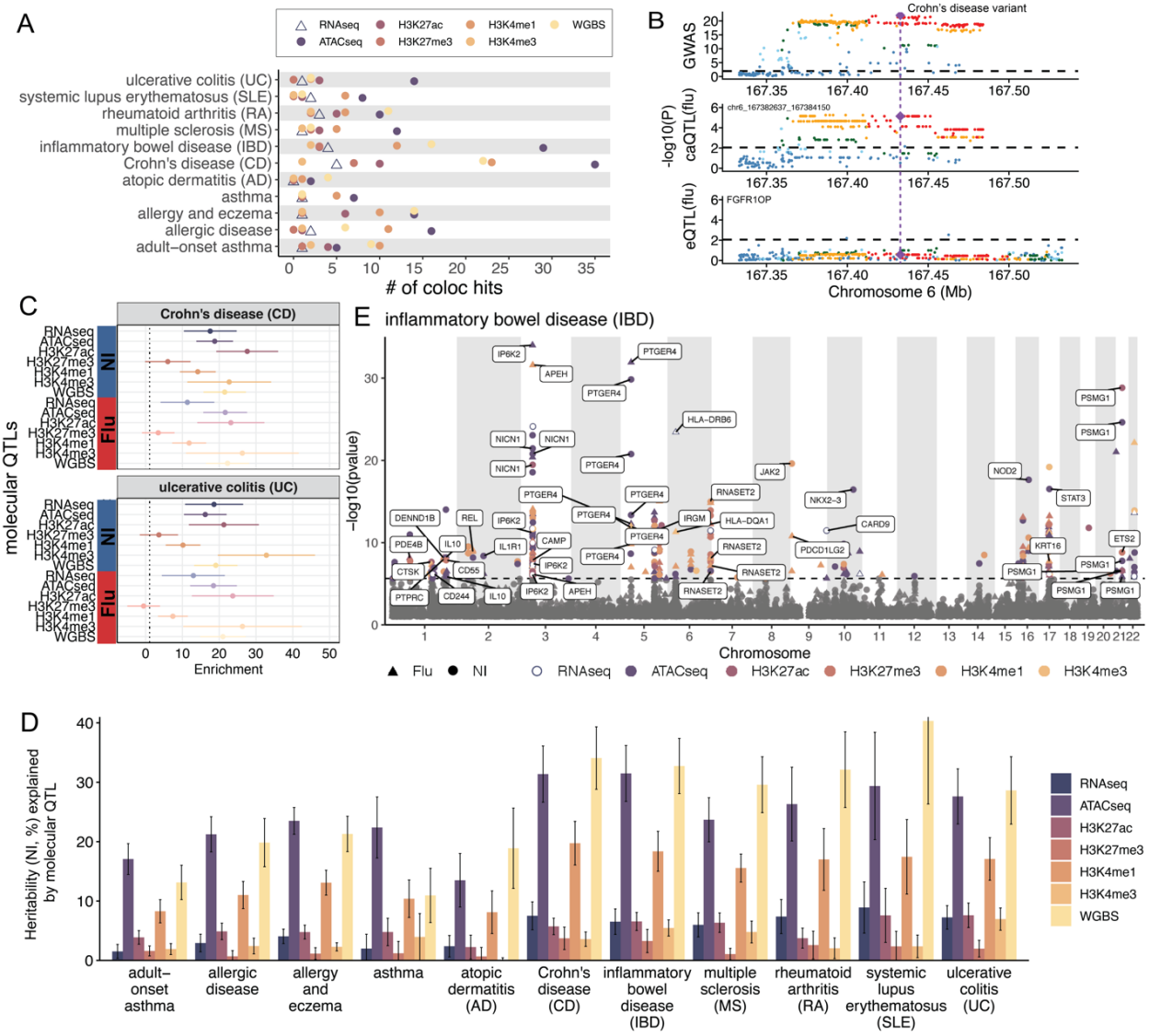


Figure 2-6. Variants controlling epigenetic marks affect immune-related disease traits. (A) Summary of colocalization results for immune related diseases. Points represent the number of significant hits defined as PP3+PP4 > 0.5 and PP4/(PP3+PP4) > 0.8 in either condition. (B) An example of a region in the flu-infected condition where a caQTL colocalizes with the GWAS variant for Crohn's disease (the purple diamond), but the eQTL does not. *P*=0.01 is represented by the dotted line. The color of the points represents the r2, the measure of linkage disequilibrium between the SNPs. (C) Examples of enrichment of heritability across the different regulatory QTLs identified. A 95% confidence interval is displayed. Enrichment results for the full 14 GWAS studies are shown in Fig. S2-10A. (D) Bar plots, with standard error, representing the percent of heritability explained by each of the molecular QTL in the non-infected condition. (E) Manhattan plot showing an example of the PrediXcan results for inflammatory bowel disease susceptibility loci. Each point represents a gene or peak that is significantly associated with the disease trait. Peaks are assigned to the closest gene and labels denote genes present in the Gene Ontology immune response set[109]. The dotted line marks the *P*=0.05 Bonferroni corrected P value cutoff.

We used Stratified LD score regression (S-LDSC)[110–112] to partition the heritability of complex traits and estimate heritability enrichment for each type of molecular QTL. S-LDSC is a tool for assessing how the heritability of a complex trait is partitioned among functional features, while controlling for LD, allele frequency and other baseline features. We first investigated the enrichment of heritability for each molecular QTL type, estimating heritability enrichment as a ratio of the proportion of heritability explained by a particular class of regulatory QTLs divided by the proportion of SNPs that belonged to that class. We found a significant enrichment of heritability across most diseases and QTL-types tested, with the strongest enrichments observed for K27acQTL and K4me3QTL and susceptibility to Crohn's disease and ulcerative colitis (up to 32-fold, Fig. 2-6C, Fig. S2-10A, Table S2-9), suggesting that genetically driven epigenetic variation in macrophages plays an important role in susceptibility to gut inflammatory disorders. We next estimated how much heritability can be explained by each type of molecular QTL, finding that chromatin accessibility and methylation QTL explain the largest percentage of heritability relative to the other data types (Fig. 2-6D, Fig. S2-10B, Table S2-9). Importantly, the observed enrichments are robust to changes in the baseline model used for the s-LDSC analyses (Fig. S2-10C, Table S2-9, see methods for details).

Lastly, we applied S-PrediXcan to identify genes for which the component of gene expression or epigenetic values determined by an individual's genetic profile (i.e., the regulatory QTLs identified herein) differed between cases and controls for the immune-related diseases described above[113]. Again, we found that genetically driven differences in epigenetic marks were more frequently associated with disease status across various immune-related diseases as compared to genetically encoded variation in gene expression levels (Fig. 2-6E, Fig. S2-9B, Table S2-10 provides the full results of S-PrediXcan analyses across all molecular traits and 11 immune-related

diseases). For IBD, for example, we found 23 genes putatively associated with disease susceptibility via changes in gene expression versus 178 genes (~8-fold more) when focusing on genetically encoded epigenetic differences. In sum, our results consistently highlight the link between genetically encoded epigenetic variation and susceptibility to immune-related disorders.

Discussion

Together, our results provide an extensive characterization of the gene regulatory landscape associated with variation in the immune response to flu infection between individuals of European and African ancestry. Our findings expand on previous work measuring genetic ancestry effects on the gene expression response to pathogens or immune stimuli[21,35,40] by showing that many of the ancestry-associated differences in transcriptional responses to pathogens are accompanied by epigenetic differences between ancestry groups. Consistent with previous findings[21–23], we found that increased levels of African ancestry are associated with a gene expression signature of increased inflammation both before and after flu infection[21–23]. Remarkably, we found that this signature of increased inflammatory potential among African ancestry individuals is even more accentuated when looking at the epigenetic landscape surrounding inflammation-associated genes. We also found that genetic ancestry was strongly associated with variation in the regulation of type-I interferon responses, although these effects appear to be temporally regulated. We show that at 24 hours post influenza infection macrophages from individuals with increased African ancestry have stronger type-I interferon response, which is the opposite pattern to what was described in peripheral blood mononuclear (PBMCs) cells soon after influenza infection[40]. Interferon responses are regulated by complex cellular-interactions and, therefore, it is plausible that different cell types (PBMCs vs macrophages) will result in different

outcomes in what concerns the relationship between ancestry and the response. Another possible explanation is that the main difference across ancestry groups is in the dynamics of the type-I interferon response rather than the overall magnitude of such response. If European-ancestry individuals engage a faster IFN response after flu infection[40] (relative to African-ancestry individuals) they might also start down-regulating such response sooner leading to what would appear as an apparent disconnect between the direction of ancestry differences at early versus late time-points. Future time-course experiments in different cell types will be critical to fully characterize the relationship between genetic ancestry and the regulation of interferon responses.

Since samples were derived from individuals with unknown life histories and environmental exposures, the ancestry-related differences we observed could be derived from a combination of environmental and genetic factors. The integration of ancestry-effects on gene regulation with our QTL results allowed us to demonstrate that cis-genetic variants account, in large part, for the identified ancestry-associated differences in inflammatory response. In stark contrast, ancestry-associated differences in type-I interferon response – one of the pathways most commonly identified as divergent between European and African ancestry individual[35,40] – do not appear to be explained by differences in allele frequency of *cis* genetic variants. These data suggest, therefore, that variation in interferon responses is likely environmentally driven or explained by *trans* genetic variants that we are unpowered to identify in this study. More generally, we show that genetics contributes more to epigenetic variation at the population level than it does towards variation in gene expression, corroborating previous findings only focused on population variation in DNA methylation levels[61,62]. We speculate that this finding reflects a more direct causal role of variation in TF binding to epigenetic variation versus gene expression levels that often require the combined action of several transcriptional regulators and regulatory elements. In

general, our data points to a driving role for differential TF binding in many of the molecular QTL identified (especially the epigenetic QTL), suggesting that additional effort should be invested to developing large scale datasets of TF-binding QTL, which as of now remain scarce and limited to very few TFs[79,114,115].

Our data raises questions about the commonly accepted notion that increased chromatin accessibility at baseline allows for a stronger transcriptional response to infection[52,53]. Although we cannot exclude the possibility that this is true at a limited number of loci[64], we show that this is not a generalizable feature across the genome. We show that an increase in chromatin accessibility prior to infection –coupled with higher levels of other activation marks, such as H3K4me1 and H3K27ac – is not in itself sufficient to "prime" cells to respond differently to a pathogenic attack. It is therefore likely that enhancer priming requires, in addition to epigenetic modification, active changes in the baseline activity of particular TFs as well as changes to the metabolic state of macrophages[116]. Our conclusion, however, must be considered within the limitations of our experiment; notably the fact that we have only measured transcriptional responses at a single time point (24 hours post infection) and that we are limited in our ability to link specific enhancers to the genes that they regulate. Using chromatin conformation capture data to more confidently link enhancers to the genes they regulate will be important to further evaluate the priming model.

Finally, our results indicate that epigenetic QTL are a powerful means to identify the mechanisms of disease-associated genetic variation. About 90% of GWAS variants map to non-coding regions of the genome, suggesting that they likely affect traits through gene regulation[117]. Despite immense efforts to characterize eQTL across thousands of individuals, tissue types and experimental conditions[118,119], only ~40% of GWAS variants colocalize with eQTLs[120]. The

modest overlap between GWAS loci and eQTLs is often attributed to the fact that many of the GWAS variants may only have an impact on gene expression during development, in specific cell types, or under environmental/experimental conditions not yet profiled. Our data indicates that epigenetic QTLs help fill this gap, by providing a means to markedly increase the number that colocalize (by about 10-fold) between GWAS variants and regulatory variants beyond those identified using eQTLs alone. We caution interpreting epigenetic variation as the causal mechanism behind variation in disease traits. We speculate, instead, that these epigenetic QTLs allow for the identification of sites associated with variation in TF binding. Therefore, they may serve as a proxy for genetic variation that, under particular environmental conditions, will have an impact on gene expression levels (Fig. S2-9C for a schematic model). Collectively, our data indicates that our understanding of disease etiology, genetic heritability, and disease risk can be greatly increased by considering molecular traits beyond gene expression.

Acknowledgments

We thank Silvia Vidal from McGill University for a gift of the Influenza strain. We thank all members from the Barreiro and Bourque labs for their comments on the paper. This work was supported by National Institute of Health Research grants R01-GM134376 and P30-DK042086 to L.B.B. It is also supported by a Canada Institute of Health Research (CIHR) program grant (CEE-151618) for the McGill Epigenomics Mapping Center, which is part of the Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) Network, to G.B., L.B.B. and T.M.P. K.A.A. is supported by a grant to University of Chicago from the Howard Hughes Medical Institute through the James H. Gilliam Fellowships for Advanced Study program. G.B. is supported by a Canada Research Chair Tier 1 award, a FRQ-S, Distinguished Research Scholar award and by the World Premier International Research Center Initiative (WPI), NEXT, Japan. The Canadian Center for Computational Genomics (C3G) is supported by a Genome Canada Genome Technology Platform grant. Computational resources were provided by the University of Chicago Research Computing Center (Barreiro team) and Calcul Québec and Compute Canada (Bourque team). Figures 2-1A and S2-9C were created with BioRender.com.

Author Contributions

L.B.B, G.B and T.M.P conceived the project. L.B.B. directed the study. V.Y., R.S., A.P (Albena), and M.S. performed experimental work. K.A.A. led the computational analyses, with contributions from Y.L, A.P (Alain), S.G., Z.M., K.L, C.G., X.C., X.H.,Y.L., C.B. and R.P. A.P. (Alain) and D.L. developed and implemented the EpiVar browser with help from R.G., D.B., and D.B. K.A.A. and L.B.B. wrote the manuscript, with input from all authors.

Declaration of interests

The authors declare no competing interests.

Lead contact

Reagent and resource requests should be addressed and will be fulfilled by the Lead Contacts, Luis Barreiro (lbarreiro@uchicago.edu) and Guillaume Bourque (guil.bourque@mcgill.ca).

Materials availability

This study did not generate new unique reagents.

Data availability

Sequence data has been deposited at the European Genome-phenome Archive (EGA), under accession numbers EGAD00001008422 (RNA-seq, ATAC-seq and ChIPmentation) and EGAD00001008359 (WGS and WGBS). We also constructed a versatile QTL browser (https://computationalgenomics.ca/tools/epivar), which allows users to explore and visualize mapped QTLs for gene expression, chromatin accessibility, histone modifications and DNA methylation.

Code availability

All original code is available at https://github.com/katiearacena/EU_AF_ancestry_flu_code. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Materials and Methods

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sample collection

Buffy coats from 39 healthy donors were obtained from the Indiana Blood Center (Indianapolis, IN, USA). A signed written consent was obtained from each participant and the project was approved by the ethics committee at the CHU Sainte-Justine (protocol #4022). All individuals recruited in this study were males, self-identified as African-American (AF) (n = 19) or European-American (EU) (n = 20) between the age of 18 and 54 years old. We focused on males to avoid potential effects of sex-specific differences in expression, which would reduce the power of our study. Because genetic effects on gene expression are largely indistinguishable between sexes[25], it is reasonable to assume that the molecular QTLs reported herein are likely generalizable to both males and females. The average age across AF and EU samples was similar (38.7 years for AF versus 38.6 years for EU). We only collected male samples to avoid the potentially confounding effects of sex-specific differences in immune responses to infection. Only individuals self-reported as currently healthy and not under medication were included in the study. In addition, each donor's blood was tested for Hepatitis B, Hepatitis C, Human Immunodeficiency Virus (HIV), and West Nile Virus, and only samples negative for all of the tested pathogens were used. All data sample collection and data generation were completed before the COVID-19 pandemic.

Monocytes isolation and macrophages generation

Blood mononuclear cells were isolated by Ficoll-Paque centrifugation. Monocytes were purified from peripheral blood mononuclear cells (PBMC) by positive selection with magnetic CD14

MicroBeads (Miltenyi Biotech) using the autoMACS Pro Separator. The purity of the isolated monocytes was verified using an antibody against CD14 (BD Biosciences) and only samples showing > 90% purity were used to differentiate into macrophages. To generate the monocytes-derived macrophages (MDM), the cells were cultured for 6 days in RPMI-1640 (Fisher) supplemented with 10% heat-inactivated FBS (FBS premium, US origin, Wisent), L-glutamine (Fisher), gentamicin (10ug/mL LifeTechologies) and M-CSF (20ng/mL; R&D systems) and incubated at 37°C and 5% CO_2. Cell cultures were fed every 2 days with complete medium. The macrophages obtained using this differentiation protocol have the expected phenotype of pure, non-activated primary macrophages. They stain positive for all the classical macrophage surface markers (CD14, CD16, CD11b, CD206), and negative/low for the activation markers HLA-DR.

Infection of macrophages

On day 6, the macrophages were harvested with CellStripper (Corning), counted, replated with the fresh media (previously mentioned) without antibiotic and incubated overnight. The next day, the cells were infected at a multiplicity of infection (MOI) of 0.1 for Influenza A virus strain *PR8WT* (Flu). A control group of non-infected macrophages (NI) was treated the same way but with only medium without virus. For some samples (n=4-6 for each assay), we added Mock at the same volume as for the Flu and NI conditions. The mock condition refers to macrophages that were treated with the supernatant from growing the Madin-Darby canine kidney (MDCK) cells used to propagate the virus but without any virus present. 24 hr post-infection, the cells were collected for downstream experiments.

gDNA extraction

Genomic DNA extraction was performed on 0.6 to 7 million (from NI or Flu macrophages) using the DNeasy Blood & Tissue kit (Qiagen). The genomic DNA was quantified using Quant-iT PicoGreen ds DNA Assay Kit (ThermoFisher Scientific).

Whole genome sequencing (WGS)

Libraries were generated from 400 ng of genomic DNA fragmented to 300–400 bp peak sizes using the Covaris focused-ultrasonicator E210. Library preparation was done using NxSeq AmpFREE Low DNA Library Kit (Lucigen) according to the manufacturer's instructions. The libraries were size selected using Ampure XP Beads (Beckman Coulter) and quantified using the KAPA Library Quantification kit - Universal (KAPA Biosystems). Sequencing of the WGS libraries was performed on the Illumina HiSeqX system using 150-bp paired-end sequencing.

Whole genome bisulfite sequencing (WGBS)

Libraries were generated from 1500 ng of genomic DNA spiked with 0.1% (w/w) unmethylated λ DNA (Roche Diagnostics) fragmented to 300–400 bp peak sizes using the Covaris focused-ultrasonicator E210. Library preparation was done using NxSeq AmpFREE Low DNA Library Kit (Lucigen) according to manufacturer's instructions, followed by bisulfite conversion with the EZ-DNA Methylation Gold Kit (Zymo Research) according to the manufacturer's protocol. Libraries were amplified by 6 cycles of PCR using the Kapa Hifi Uracil + DNA polymerase (KAPA Biosystems) according to the manufacturer's protocol. The amplified libraries were size selected using Ampure XP Beads (Beckman Coulter) and quantified using the KAPA Library

Quantification kit - Universal (KAPA Biosystems). Sequencing of the WGBS libraries was performed on the Illumina HiSeqX system using 150-bp paired-end sequencing.

RNA extraction

Macrophages were directly lysed from the culture plate with 1mL of Qiazol from 0.5 to 2.5 million cells (NI, Flu and Mock) and extracted using the miRNeasy kit (QIAGEN) following the manufacturer instruction. RNA integrity was assessed with the Agilent 2100 Bioanalyzer System (Agilent Technologies).

RNA sequencing (RNA-Seq)

RNA library preparations were carried out on 100-500 ng of RNA with RIN 1.2 to 9.8 using the Illumina TruSeq Stranded Total RNA Sample preparation kit, according to manufacturer's protocol. The libraries were size-selected using Ampure XP Beads (Beckman Coulter) and quantified using the KAPA Library Quantification kit – Universal (KAPA Biosystems). Sequencing of the RNA-Seq libraries was performed on the Illumina NovaSeq 6000 system using 100-bp paired-end sequencing.

ATAC-seq

ATAC-seq library preparation was performed using the Omni-ATAC protocol[121]. 50,000 macrophages (from NI, Flu and Mock conditions) were resuspended in 1 ml of cold ATAC-seq resuspension buffer (RSB; 10 mM Tris-HCl pH 7.4, 10 mM NaCl, and 3 mM $MgCl_2$ in water). Cells were centrifuged at 500 g for 5 min in a pre-chilled (4 °C) fixed-angle centrifuge. After centrifugation, supernatant was aspirated and cell pellets were then resuspended in 50 µl of ATAC-

seq RSB containing 0.1% IGEPAL, 0.1% Tween-20, and 0.01% digitonin by pipetting up and down three times. This cell lysis reaction was incubated on ice for 3 min. After lysis, 1 ml of ATAC-seq RSB containing 0.1% Tween-20 (without IGEPAL and digitonin) was added, and the tubes were inverted to mix. Nuclei were then centrifuged for 10 min at 500 rcf in a pre-chilled (4 °C) fixed-angle centrifuge. Supernatant was removed and nuclei were resuspended in 50 uL transposition mix (2x TD Buffer, 100 nM final transposase, 16.5 uL PBS, 0.5 uL 1% digitonin, 0.5 uL 10% Tween-20, 5 uL H2O). Transposition reactions were incubated at 37 °C for 30 min in a thermomixer with shaking at 1000 rpm. Reactions were cleaned up with Zymo DNA Clean and Concentrator 5 columns. Primers (i5 and i7) were added by amplification (12 cycles) using NEBNext 2x MasterMix. Sequencing of the ATAC-seq libraries was performed on the Illumina NovaSeq 6000 system using 100-bp paired-end sequencing.

ChIPmentation

Crosslink step

For ChIPmentation, 1 to 5 million macrophages (from NI, Flu and Mock conditions) were washed in cold PBS prior proceed the cross-linking of DNA with formaldehyde (0.75%) by shaking the tube for 10 min at RT and adding Glycine (125nM) for additional 5 min. Cells were washed with cold PBS and centrifuged for 5 minutes at 2500 xg at 4°C. The supernatant was discarded and the cell pellet immediately frozen at -80°C.

After cell lysis, sonication of nuclei was performed on a BioRuptor UCD-300 targeting 150-500 bp size. Immunoprecipitation of the histone marks H3K27ac, H3K4me1, H3K27me3 and H3K4me3 was performed following the Auto-ChIPmentation protocol for Histones (Diagenode inc, Denville, USA) according to the manufacturer's instructions. The libraries were size selected

using Ampure XP Beads (Beckman Coulter) and quantified using the KAPA Library Quantification kit – Universal (KAPA Biosystems). Sequencing of the ChIPmentation libraries was performed on the Illumina NovaSeq 6000 system using 100-bp paired-end sequencing.

QUANTIFICATION AND STATISTICAL ANALYSIS

WGS processing and genotyping

Raw reads were trimmed using Skewer[122] and the resulting reads were aligned to the hg19 human reference genome using BWA-MEM[123]. Insertion/deletion realignment and base quality score recalibration were performed using GATK[124] and duplicates were marked using Picard (http://broadinstitute.github.io/picard/). We used GATK's *HaplotypeCaller* to perform SNV and INDEL calling. We filtered the joint genotyped file to exclude non-autosomal and non-biallelic variants. Additionally, we removed SNPs that had a call rate of <90% across all samples, that deviated from Hardy–Weinberg equilibrium at $p < 10^{-5}$, and with minor allele frequency less than 5%. We used the resulting 7,383,243 SNPs in QTL mapping and other downstream analyses. We annotated the SNPs using dbSNP (human_9606_b151_GRCH37p13)[125].

Estimation of genome-wide admixture levels

We used the clustering algorithm ADMIXTURE (v1.3.0) to calculate the percentage of African and European ancestry in each individual[95]. Notably, we only obtained genotyping data for 17/19 self-identified African individuals and 18/20 European individuals, thus, we only calculated admixture estimates for samples we had data (n=35). We included Yoruba (YRI) and European (CEU) individuals from the 1000 Genomes reference panel and estimated ancestry proportions using K=2 ancestral clusters. We validated our choice of K=2 by varying K from 2 to 6 in the

ADMIXTURE analysis and plotting the cross-validation errors for each. K=2 had the lowest cross-validation error. We applied Genotype Harmonizer[126] to align and combine the 1000 genomes reference data. We used 362,075 unlinked SNPs (r^2 between all pairs < 0.1) to estimate genetic ancestry. ADMIXTURE analyses showed that 3 AF individuals were likely mislabeled by the blood center as they presented 99.9% of European ancestry. Ancestry labels were adjusted accordingly resulting in 14 African American and 21 European American individuals. Estimated ancestry proportions for each individual were used to calculate population differences unless specified otherwise.

Local ancestry

Genotypes were phased using Shapeit4 v4.2.2 with 1000 genomes phase 3 samples as the reference. We used bcftools (v2.29.1) fixref function to match allele calls. We started by validating our global genetic ancestry estimates using RFMix[127] and found a near perfect correlation $R \approx 1$ between ADMIXTURE and RFMix estimates of global genetic ancestry (Fig. S2-11A). Global estimates calculated using RFMix were obtained by collapsing all the chromosomes and calculating the total length of CEU and YRI ancestry tracks for each individual. To estimate local ancestry around each gene/feature being tested we averaged local ancestry estimated for 100kb +/- surrounding each feature, weighted based on proportion overlap.

Estimation of technical variation

Given the high cost associated with generation of complete epigenetic profiles across large number of individuals, we do not have technical replicates for all the samples. However, for a subset of individuals (n=4-6 for each assay), we obtained gene expression and epigenetic profiles for both

mock and non-infected conditions. We used these data to get an estimate of technical variation across datasets given that we expect mock and non-infected samples to show similar molecular profiles. Accordingly, and across all data types, we found a very high correlation between the gene expression/epigenetic estimates obtained in mock (i.e., samples treated with the vector in the absence of flu) and non-infected samples (R>= 0.96, Fig. S2-1A, S2-1B, S2-11B), suggesting little technical variation in our different data sets.

RNA-seq data processing

Adaptor sequences and low-quality score bases (Phred score < 30) were first trimmed using Trimmomatic[128]. The resulting reads were aligned to the hg19 human reference genome assembly, using STAR[129]. Read counts are obtained using HTSeq[130] with parameters -m intersection-nonempty -stranded=yes.

ChIPmentation and ATAC-seq data processing

ChIPmentation and ATAC-seq reads were first trimmed for adapter sequences and low-quality score bases using Trimmomatic[128]. The resulting reads were mapped to the human reference genome (hg19) using BWA-MEM[123] in paired-end mode at default parameters. Only reads that had a unique alignment (mapping quality > 20) were retained and PCR duplicates were marked using Picard tools (https://broadinstitute.github.io/picard/). Peaks were called using MACS2 software suite[131].

Enhancer RNA calling

Read counts for enhancer RNAs (eRNAs) were obtained using featureCounts[132] over regions of the genome annotated as putative enhancers using our chromHMM-annotations. Enhancers located within a gene body were excluded from the analysis since for those regions it is difficult to distinguish between eRNAs and coding transcripts. This resulted in the quantification of 94,933 putative eRNA regions. After excluding lowly expressed eRNAs, defined as those that did not have at least 10 reads in ≥ 50% of Flu *or* non-infected samples, we were able to study a total of 17,142 putative eRNAs.

Filtering phenotype data

In our RNAseq dataset, we excluded any genes that did not have an average RPKM > 2 in Flu *or* non-infected samples. For the ChIPmentation and ATACseq datasets, we calculated median peak size and required 50% of median value overlap for peaks to be called as the same peak between samples using bedtools merge[133]. We then filtered to exclude peaks that were not present in ≥ 50% of Flu *or* non-infected samples, and those that fall within blacklisted regions[134]. As described above, for the eRNA dataset, we excluded any regions that did not have at least 10 reads in ≥ 50% of Flu *or* non-infected samples. The number of features remaining after these thresholds are present in Table 2-1 below.

Data	**Median peak size**	**50% of median for peak merging**	**Total # of samples included in analysis**	**≥50% NI condition threshold**	**≥50% Flu condition threshold**	**Features *before* filtering**	**Features *after* filtering**
RNAseq	N/A	N/A	70	N/A	N/A	57905	**14122**
ATACseq	608	304	70	18/35	18/35	974189	**118201**
H3K27ac	349	174.5	58	15/29	15/29	608338	**46657**

Table 2-1. Filtering criteria for phenotype data.

H3K27me3	592	296	56	14/28	14/28	694827	**70675**
H3K4me1	498	249	60	15/30	15/30	749325	**113584**
H3K4me3	843	421.5	54	14/27	14/27	144765	**25568**
eRNA	N/A	N/A	78	20/39	20/39	94933	**17142**

Table 2-1, continued. Filtering criteria for phenotype data.

We used featureCounts[132] to calculate the number of reads for each genomic feature for each sample. We used the resulting counts matrices for all downstream analyses.

Partitioning the genome using ChromHMM

We generated genome-wide, gene regulatory annotation maps for noninfected and flu infected MDMs using the ChromHMM chromatin segmentation program[135]. We used samples for which there was data for all 4 histone marks (n= 27 samples, 10 AF, 17 EU) and 7 emission states. We used ChromHMM profiles from the Roadmap Epigenetics project to annotate our results[47].

Whole genome bisulfite sequencing data processing

Adaptor sequences and low-quality score bases were first trimmed using Trimmomatic[128]. The resulting reads were mapped to the human reference genome (hg19) and lambda phage genome using Bismark[136], which uses a bisulfite converted reference genome for read mapping. Only reads that had a unique alignment were retained and PCR duplicates were marked using Picard tools (https://broadinstitute.github.io/picard/). Methylation levels for each CpG site were estimated by counting the number of sequenced C ('methylated' reads) divided by the total number of reported C and T ('unmethylated' reads) at the same position of the reference genome using Bismark's methylation extractor tool. We performed a strand-independent analysis of CpG methylation where counts from the two Cs in a CpG and its reverse complement (position on the plus strand

and position i+1 on the minus strand) were combined and assigned to the position of the C in the plus strand. To assess MethylC-seq bisulfite conversion rate, the frequency of unconverted cytosines (C basecalls) at lambda phage CpG reference positions was calculated from reads uniquely mapped to the lambda phage reference genome.

We obtained methylation counts for 19,492,906 loci. Due to the high coverage of the data, we opted to not perform smoothing. To reduce the total number of statistical tests performed, we limited our analyses to CpG sites in open chromatin regions using ChromHMM data (states E3-E7). We also excluded C nucleotides that overlapped with SNPs identified in the whole genome sequencing data for our samples. After these filtering steps we analyzed methylation levels across 7,463,164 CpG sites.

Infection effects: Infection-Related Differential Effects

We used all samples we had collected data for, not just those with genotyping data to calculate infection effects. Note that this only increased the sample size for RNAseq data (from n=35 to n=39 individuals). The sample size for all other data types remained the same. For RNAseq, ATACseq and ChIPmentation datasets, we calculated normalization factors to scale the raw library sizes using calcNormFactors in edgeR (v 3.28.1)[137]. We used the voom function in limma (v 3.42.2)[138] to apply these factors, estimate the mean-variance relationship and convert raw read counts to logCPM values. Because samples were sequenced on different flowcells at different times (i.e., hereafter defined as "Batch") we regressed out these putative batch effects by fitting a linear model that estimates the technical effect of sequencing batch on the different datasets. We kept the residuals from this model (i.e, batch-corrected "expression" estimates) using the residuals.MArraLM function.

To calculate global infection effects for RNAseq, ATACseq and ChIPmentation datasets, batch-corrected read counts of samples corresponding to the same individual were compared in a paired design by introducing individuals as additional covariates. The following model was run using limma for each data type independently:

$$M_1 : E(i,j) \sim \begin{cases} \beta_0(i,j) + \varepsilon^{NI}(i,j) \text{ if Condition} = NI \\ \beta_0(i,j) + \beta_{flu}(i) + \varepsilon^{flu}(i,j) \text{ if Condition} = flu \end{cases}$$

Here, *E(i,j)* represents the batch-corrected estimate of each feature *i* for individual *j* and $\beta_0(i,j)$ represents the intercept corresponding to feature *i* and individual *j* (i.e., the expectation of gene or peaks *i*'s expression level in the non-infected sample for individual *j*). $\beta_{flu}(i)$ is the effect of flu infection on feature *i*. We performed 1000 permutations obtained by randomly reshuffling the condition labels in each condition in order to estimate FDR using the qvalue R package (v 2.18.0)[97].

Identification of differentially methylated loci

We identified differentially methylated loci (DML) in response to flu infection using the R package DSS and a fixed effects model[139]. We used the DMLfit.multiFactor function in DSS, using the same model described above (M1). We performed 10 permutations and FDR correction using the same approach detailed above.

Percent Variance Explained by Infection

The R package relaimpo (v 2.2-3)[140] was used in order to calculate the relative contribution of each predictor in the infection effects linear models to the R^2. The same batch corrected counts matrices and weights were used as before with the exception of the methylation loci, which were additionally filtered to remove sites that did not have coverage ≥4 sequence reads in at least half

of the non-infected or Flu-infected samples and those with 0 variation across all NI or Flu samples. DSS accounts for both low coverage and variation which is why these sites were previously included in the model. The same model (M1) was run for all datatypes. We followed the same workflow to compare the mock and non-infected samples.

GSEA of Infection effects and popDE effects

The R package fgsea[141] was used to perform gene set enrichment analysis (GSEA) to determine which biological pathways were enriched or depleted among DE genes/regions and popDE genes/regions. We connected CpG loci, ChIPmentation peaks, and ATACseq peaks to the nearest gene using the R package ChIPseeker[142] (using the default parameters). For GSEA each gene can only be included once. Thus, in situations where more than one peak was mapped to the same gene, we kept the peak with the highest t-statistic when modeling flu-infection effects or popDE effects. For the WGBS infection effects we used the difference between CpG methylation in the flu-infected and non-infected conditions to perform GSEA since the Wald test statistic from the model does not indicate the direction of the effect. For the WGBS popDE GSEA the Wald test statistic was used. GSEA were performed against the Hallmark gene set[99].

Relationship between expression and epigenetic changes in response to infection

To evaluate the relationship between gene expression changes and epigenetic changes in response to flu infection we connected peaks and CpG loci to the nearest gene using the R package ChIPseeker using the same parameters as detailed previously[142].

We first subset on upregulated genes defined as those genes with beta> 0.5 & FDR < .01. Additionally, we subset to include only epigenetic marks that change in response to flu infection

using FDR < .20 for CpG loci and FDR<.01 for all other marks. We then evaluate in which direction the epigenetic features associated with genes upregulated in response to flu infection change. The same analysis is done using downregulated genes (beta < -0.5 and FDR < .01). We used a Wilcoxon test to determine significance levels using peaks for all genes (not just those upregulated and downregulated) as the null.

Transcription Factor activity scores

Footprints were called using HINT-ATAC from the Regulatory Genomics Toolbox[143] on the subset of peak regions called using MACS2[131]. Footprint calling was performed by first merging aligned ATACseq reads within each condition using *samtools merge,sort,index*. A meta-footprint set was created for each pair by merging the respective footprint calls with *bedtools merge*[133]. Using this meta-footprint set, transcription factor motif matching was performed on the subset of regions falling within meta-footprints.

Motif matching was done using the JASPAR CORE Vertebrates set of curated position frequency matrices[144]. Because of similarity across TF motifs, we chose to group TFs into clusters based on similarity. To do this, we first computed pairwise TOMTOM[145] *E* value metrics to assess motif similarity. The log10 *E* values were then used as distance metrics for hierarchical clustering (base *R*; hclust(method="ward.D2")). A cutoff height of 10 (base *R*; cutree(h=10)) was used to define TF motif clusters, resulting in a total of 200 clusters which were used for the TF enrichment analysis described later.

Using the set of motif match regions for each TF, motif count enrichment was performed using the *rgt-motifanalysis enrichment* function. Background regions were defined as all meta-footprints. Foreground regions are the footprints overlapping regions of the genome of interest**.** A

two-sided Fisher's exact test was computed from the output frequencies of motif occurrences within the foreground and background regions (base *R*; fisher.test(alternative="two.sided")). P values were corrected using the Benjamini Hochberg method[146].

Using the set of motif match regions for each TF, activity analysis was performed with the RGT *differential* function. The activity score metric is described further in Li et al. 2019[143]. Parameters for footprinting, motif matching, and differential activity analysis were set as default. Activity statistics were calculated per sample between conditions using the ATACseq profiles of each sample independently. Combined activity scores were computed as the mean across samples, and meta p-values were calculated by Fisher's combined probability test (python scipy.stats.combine_pvalues) to summarize across all samples.

Correcting for technical effects in popDE, popDR and QTL mapping analysis

All popDE, popDR and QTL mapping analyses were performed on count matrices corrected for age and potential sequencing batch effects. Age and batch correction were done separately for NI and Flu-infected samples. We started by calculating normalization factors to scale the raw library sizes using calcNormFactors in edgeR (v 3.28.1)[137]. Then, we used the voom function in limma (v 3.42.2)[138] to apply these factors, estimate the mean-variance relationship and convert raw read counts to logCPM values. Batch effects, which are categorical variables, were regressed out using ComBat from the sva Bioconductor, fitting a model that also includes age (mean centered) and admixture. We subsequently regressed out age effects using limma.

Detection of population differentially expressed (popDE) features

Using the age and batch corrected matrices described above, we used limma to detect the effect of African admixture for RNAseq, ATACseq and ChIPmentation datasets using the following nested model for each data type independently:

$$M_2: E(i,j) \sim \begin{cases} \beta_0(i) + \beta_{AF}^{NI}(i) \cdot AF(j) + \varepsilon^{NI}(i,j) \text{ if Condition} = NI \\ \beta_0(i) + \beta_{flu}(i) + \beta_{AF}^{flu}(i) \cdot AF(j) + \varepsilon^{flu}(i,j) \text{ if Condition} = flu \end{cases}$$

Here, *E(i,j)* represents the age and batch corrected estimate of feature *i* for individual *j*, $\beta_0(i)$ is the global intercept accounting for the expected expression of feature *i* in a 100% European-ancestry non-infected individual, $\beta_{AF}^{NI}(i)$ and $\beta_{AF}^{flu}(i)$ indicate the effects of African admixture on feature *i* within each condition. The model was fit using limma and the estimates of $\beta_{AF}^{NI}(i)$ and $\beta_{AF}^{flu}(i)$ of the genetic ancestry effects were extracted across all features. We used 1000 permutations obtained by randomly reshuffling admixture estimates in order to estimate FDR, as described in detail above.

Because of the different nature of the methylation data (i.e., percentage methylation per CpG site instead of counts) we used DSS (Dispersion Shrinkage for Sequencing data) instead of limma to model population differentially methylated loci. DSS is specifically designed for the analyses of bisulfite sequencing (BS-seq) differential methylation. The core of DSS is a procedure based on Bayesian hierarchical model to estimate and shrink CpG site-specific dispersions, then conduct Wald tests for detecting differential methylation. We used the same model as M2 described above but including age and batch as covariables. We permuted the 10 times by shuffling the admixture estimates to obtain null p value distributions for FDR calculations.

To increase our power to detect condition and shared effects we applied Multivariate Adaptive Shrinkage in R (mashr v0.2.28)[98] to the outputs of the popDE effects for each data type

independently. For each condition, effect sizes were obtained from limma and standard error of the effect size was calculated by multiplying the square root of the posterior variance (s2.post) of each feature by the unscaled standard deviation for the effect size of interest for that feature (stdev.unscaled). NI and Flu effect sizes and standard error of effect sizes were formatted into n x m matrices, where: n = number of features for each data type, m = 2 conditions (NI and Flu).

We estimated the null correlation of the data using the "estimate_null_correlation_function" in mashr. We included canonical covariance and data-driven covariance matrices. The data-driven covariance matrix is the top 5 PCs from a PCA performed on the significant (local false sign rate (lfsr) < .05) signals identified in the condition-by-condition model learned from our data in the mash model fit. We then fit the mash model using the mash function. For the methylation data we made some modifications to the mash procedure. First, we removed any NAs from the DSS results, resulting from insufficient coverage at a particular CpG site. As described above, DSS uses a Wald statistical test to test each gene/CpG site for differential methylation, so we used the Wald test statistic as the effects, setting all standard errors to 1. Instead of using all the tests to estimate the null correlation structure like the other datasets, we obtained a random subset of 200,000 tests and applied the estimate_null_correlation_simple function. As with the other data types, we included canonical covariance and data-driven covariance matrices and fit the mash model on all the tests performed.

After running mash, we conservatively used both qvalue and local false sign rate (lfsr) to determine if popDE effects were condition-specific (i.e., only showing an effect in the non-infected or flu-infected conditions) or shared (i.e., showing an effect in both conditions). Specifically, we require popDE features to have both FDR <.10 and lfsr < .10 in only one condition to be considered condition specific. popDE features are shared if FDR < .10 and lfsr <.10 in both conditions.

We also performed popDE analysis using local ancestry estimates. The average local ancestry estimate for each feature was obtained for 100kb +/- surrounding each feature. A weighted average, based on proportion overlap, was calculated and used as the local ancestry estimate for each feature. As in our cis-regulatory popDE analysis we used 5 permutations due to the computational limitations of looping through each feature and adding its local ancestry estimate. We were also unable to perform this analysis for the methylation data because of this constraint as there are over 7 million CpG sites that would have needed to be estimated.

Detection of population differentially responsive features

We used the age and batch corrected count matrices and weights to model the effects of African admixture on the intensity of the response to flu infection (popDR effects). We build individual-wise fold-change (FC) matrices by subtracting non-infected counts from flu-infected counts for each individual (Flu-NI) using weights calculated using the same method as in Harrison et al. 2019[147]. Specifically, given the fold-change entry of: $FC = E^{Flu} - E^{NI}$, we calculate expected variance of the FC: $\sigma^2(FC) = \sigma^2(E^{flu}) + \sigma^2(E^{NI})$. Within condition weights are: $\omega_{NI} = 1/\sigma^2(E^{NI})$ and $\omega_{flu} = 1/\sigma^2(E^{flu})$, thus the fold change weight:

$$\omega_{FC} = \frac{1}{\sigma^2(FC)} = \frac{1}{\frac{1}{\omega_{NI}} + \frac{1}{\omega_{flu}}}$$

We subset the fold-change matrices to only those features with FDR<.10 for infection effects, since if a feature is not significantly differentially expressed, it cannot be differentially responsive. We then used limma with weights and modeled the effect of admixture on fold changes:

$$M_3: E(i,j) \sim \{\beta_{AF}(i) \cdot AF(j) + \varepsilon(i,j)$$

Here, $E(i,j)$ represents the fold change for feature *i* for individual *j* and $\beta_{AF(i)}$ signifies the effects of African admixture on feature *i*. For the WGBS, we constructed individual-wise fold-change matrices and subset on CpG sites with FDR < .20 for infection effects. The model used was the same as M3 but including age and batch which for the methylation data are not corrected for *a priori*.

For all data types we performed 1000 permutations obtained by randomly reshuffling admixture estimates in order to estimate FDR, as described previously.

Calculation of pathway activity scores across individuals

We used the R packages Gene set variation analysis (GSVA), which estimates variation of pathway activity, to calculate individual mean scores for several Hallmark pathways and combinations of gene sets[148]. The input for GSVA is a matrix of counts and database of gene sets. To apply GSVA to the popDE results, we first obtained all features that are popDE (FDR <.10) in at least one of the conditions. We took the mean of features (peaks or CpG sites) which shared the same closest gene such that there was only 1 value for each gene listed in the Hallmark pathway set. We split the batch and age corrected counts matrix by condition. We then applied the gsva function to each matrix, calculating an individual mean score for each gene set. We used an analogous workflow to calculate individual mean response scores for gene sets using the popDR (FDR<.20) results. The only modification is that we used the fold-change matrices instead of the batch and age corrected counts.

To evaluate the effect of the top SNP and STR on these pathway scores we used an analogous workflow as described above but with the following modifications. We first subset on features associated with genes that were included in any of the immune response pathways tested.

We also filtered out the few features that did not have both a SNP and STR associated with the feature. We then obtained the residuals after removing the top SNP and STR effects and followed the same steps as detailed above for the popDR features to calculate ancestry scores. without the effects of the top SNP and STR.

Elastic net regression to predict transcriptional response based on baseline epigenetic data

We used the glmnet R package[149] to build an elastic net model to determine if epigenetic features at baseline can predict transcriptional response to flu. Because the number of features is much larger than the number of samples, *glmnet* uses an elastic net penalty to shrink predictor coefficients toward 0. Optimal alpha parameters were identified by grid searching across a range of alphas from 0 (equivalent to ridge regression) to 1 (equivalent to Lasso) by increments of 0.1. We defined the optional alpha as the value that maximized Spearman's ρ between predicted and true transcriptional response values across samples. We set the regularization parameter lambda to the value that minimized mean-squared error during n-fold internal cross-validation.

To generate predicted transcriptional responses for a given sample, we used a leave-one-out cross-validation approach. Specifically, we separate the samples into training (n-1 individuals) and test (1) samples, where n is the sample size. Training samples were scaled independently of the test sample in each leave-one-out model to avoid bleed-through of information from the test data into the training data. To do so, for each of the datasets, we first quantile normalized the counts data for each feature (or methylation ratios in the case of methylation data) within each sample to a standard normal distribution. Training samples were then separated from the test sample and the normalized counts for each feature (e.g., peak intensity for ATAC-seq data) in the training set were quantile normalized across samples to a standard normal distribution. To predict

the transcriptional response in the test sample, we compared read counts for each feature in the test sample to the empirical cumulative distribution function for the training samples (at the same feature) to estimate the quantile in which the training sample ratio fell. The training sample was then assigned the same quantile value from the standard normal distribution using the function *qnorm* in R. A few specific settings were required for the methylation data. First, raw methylation counts were filtered to remove sites that did not have coverage ≥4 sequence reads in at least half of the non-infected or Flu-infected samples and those with 0 variation across all NI samples. Moreover, due to restrictions of the cv.glmnet function, we also removed any CpG sites that had any missing data for any individual, resulting in an input set of 5,528,187 CpG sites.

Relationship between ancestry-associated gene expression and epigenetic differences

Similar to the analysis described in "Relationship between expression and epigenetic changes in response to infection" we wanted to evaluate if there is a relationship between ancestry-associated gene expression differences and ancestry-associated epigenetic differences. We subset on popDE genes that are higher expressed in individuals with African Ancestry (beta > 0.5, FDR< .10) and evaluated how popDE (FDR<.10) epigenetic differences corresponding to these genes behave. The same is done for popDE features that are higher expressed in primarily European Ancestry individuals (beta < -0.5, FDR < .10). A Wilcoxon test was used to determine significance.

Power calculations for QTL mapping analysis

We performed power calculations using the R package powerEQTL[150], using the following for parameters: n=35, $\sigma_{y.}$ = standard deviation of the molecular trait, slope varying from 0.1 to 0.3. For each of the molecular traits, the standard deviation used in the power calculation was derived from

the empirical data. Specifically, for each of the data types, we calculate for each of the features (e.g., each gene in the gene expression data) the standard deviation across all samples. Then, we calculated the median standard deviation across features and used those estimates in our power calculations. Power was calculated assuming a family-wise type I error rate of 0.05, while correcting for the number of features tested for each of the molecular traits.

SNP genotype-phenotype association analysis

We used the R package Matrix eQTL[151] to examine the associations between SNP genotypes and multiple phenotypes of interest (gene expression, chromatin accessibility, DNA methylation levels, and histone marks) in each condition separately. To increase the power to detect cis-QTL, we accounted for unmeasured-surrogate confounders by performing principal component analysis (PCA) on the age and batch corrected expression/peak/methylation matrices. The number of PCs chosen for each data type empirically led to the identification of the largest QTL in each condition and are reported in Table 2-2 below.

Analysis	**Condition**	**Regressed PCs**
eQTL	Non-infected	1 to 4
	Flu-infected	1 to 4
caQTL	Non-infected	1 to 3
	Flu-infected	1 to 3
H3K27acQTL	Non-infected	1
	Flu-infected	1
H3K27me3QTL	Non-infected	1
	Flu-infected	1 to 2
H3K4me1QTL	Non-infected	1 to 2
	Flu-infected	1 to 4
H3K4me3QTL	Non-infected	1 to 2
	Flu-infected	1 to 2

Table 2-2. Principal components regressed for SNP-QTL mapping.

meQTL	Non-infected	None
	Flu-infected	None
eRNAQTL	Non-infected	1 to 3
	Flu-infected	1 to 4

Table 2-2, continued. Principal components regressed for SNP-QTL mapping.

Mapping was performed combining individuals in order to increase power, thus, we included the first eigenvector obtained from a PCA on the SNP genotype data as a covariate in our linear model to correct for population structure (Fig. S2-11C, S2-11D, S2-11E). For gene expression, chromatin accessibility and histone QTL mapping we used the following model:

$$M_4: E(i,j) \sim \beta^{NI}_{genotype} \cdot genotype(j) + EV1(j) + \varepsilon^{NI}(i,j) \text{ if Condition} = NI$$

$$E(i,j) \sim \beta^{flu}_{genotype} \cdot genotype(j) + EV1(j) + \varepsilon^{flu}(i,j) \text{ if Condition} = flu$$

Here, $E(i,j)$ represents the batch and age corrected expression estimates with PCs regressed for feature *i* and individual *j*. EV1 is the first eigenvector derived from the PCA on the SNP genotype data. Local associations (i.e., putative cis QTL) were tested against all SNPs located within the peak or 100Kb upstream and downstream of each peak.

Some modifications were made when performing QTL mapping using methylation proportions due to the nature of the data. First, we quantile normalized across each CpG site using the qqnorm function in R. Since we do not previously account for age and batch as for the other data types, we included mean-centered age and batch as covariates in our model in addition to the first eigenvector obtained from the PCA on the SNP genotype data for the meQTL analysis. Finally, we used a window size of 5kb up and downstream of each CpG site. We did so both to limit the number of tests and because previous studies show that SNPs associated with variation in methylation tend to be located very close to the CpG that they associate with[100,101,152].

For all data sets, we recorded the strongest association (minimum p-value) for each gene/region/CpG site, which we used as statistical evidence for the presence of at least one QTL for each of the loci tested. We permuted the genotypes ten times, re-performed the linear regressions, and recorded the minimum p-value for each gene/region/CpG site for each permutation. We used the R package qvalue[97] to estimate FDR using the permuted p-values as our null expectation. In all cases, we assume that alleles affect phenotype in an additive manner.

In addition to qvalue, we also applied Multivariate Adaptive Shrinkage in R (mashr v0.2.28)[98] to the outputs of the QTL mapping results for each data type independently. For each condition, full Matrix eQTL outputs are loaded (every SNP-feature pair tested). We obtained the effect sizes and the standard error of the effect size, calculated by dividing the beta by the t statistic. For each feature, we chose a single, top *cis*-SNP, defined as the SNP with the lowest pvalue across the two conditions. We recorded the corresponding effect sizes and standard errors of these betas for these top *cis*-SNPs and defined these as our set of "strong" tests. Additionally, we randomly sampled 200,000 rows from all the SNP-feature pairs (including both null and non-null tests). We estimated the null correlation structure using the set of random tests using the zero_Bhat_Shat_reset=2.22044604925031e-16 flag. The data driven covariance matrices were learned using the set of strong tests. We then fit the mash model to the random subset using canonical and data-driven covariance matrices. Lastly, we calculated the posterior summaries for the strong test subset using the fit from the random subset.

Evaluating the impact of potential mapping biases in our QTL results

To check if potential mapping biases over polymorphic sites could have in impact on our QTL results, we re-mapped our data using the variant aware aligner HISAT2. Adaptor sequences and

low quality score bases (Phred score < 30) were first trimmed using Trimmomatic[128] and PCR duplicates were marked using Picard tools. Trimmed reads were aligned using HISAT2[153] and its predefined variant aware index GRCh37_snp files at default parameters. Read counts were obtained using HTSeq[130] with parameters *-m intersection-nonempty -stranded=y.*

We found that the estimated effect sizes of the identified QTL remain virtually unchanged whether we use a variant aware aligner or not (minimum Pearson's r 0.96, $P<2.2x10^{-16}$, range 0.96-0.99, Fig. S2-11F). Moreover, the number of FDR<.10 significant QTL do not change across mapping methods (Fig. S2-11G). These data indicate that mapping biases, if existent, are not large enough to have a measurable impact in our QTL mapping results.

ASE Analysis

Out of the 7,383,243 SNPs for QTL mapping we excluded 588,500 that are not found in common SNPs from dbSNP used in HISAT2 mapping, resulting in 6,794,743 SNPs used as potential inputs for ASE analysis. The reads from HiSAT2 using a variant aware reference minimizes biases in the allele quantification. From this, we extracted only heterozygous genotypes for each individual. We then calculated the number of reads mapping to each allele using samtools mpileup with the following options (-d 1000000 --min-MQ 30 --min-BQ 30) for each experimental library corresponding to that individual. For all ASE analyses we focused on heterozygous SNPs with a read coverage greater than 20 and located on autosomes. Testing of ASE was performed for each library separately using QuASAR[154,155] beta-binomial model using 0.5 as the null hypothesis allelic imbalance and 0.001 as base calling error rate.

To connect the QTL to ASE events we first overlapped snps tested for ASE with each QTL-tested feature (gene or peak) using bedtools intersect (v2.29.1). For each ASE SNP we obtained

the results for each heterozygous individual and the QTL results for the overlapping feature. To evaluate a potential enrichment of QTL for ASE events, we determined if there was a significant QTL (FDR <.10) and ASE event (QuASAR adjusted p-value <.10). The ASE event was classified as significant if there was evidence of a significant ASE event for at least 1 individual. We then performed a logistic regression in to determine if there is a relationship between a feature being a QTL and overlapping an ASE event.

Validation of QTL using previously published data sets

We compared the effect sizes of our QTL by re-capitulating the results previously described in Randolph et al, Nedelec et al, Alasoo et al and Husquin et al[21,40,61,64]. If necessary, coordinates were lifted over from hg38 to hg19 using CrossMap (v0.6.5). We also removed all A/T and G/C SNPs to remove ambiguous SNPs since we did not have strand information. Using matched genotypes, counts matrices and the same model as previously reported, we re-ran QTL mapping for each of these studies in Matrix EQTL with a 100KB window size. For peak coordinates for Alasoo et al., we used CrossMap to transfer the coordinates from hg38 to hg19, removing any regions that mapped to multiple regions. We then overlapped the peaks with those from our study (~118k) and built a new read counts matrix by summing the counts of any overlapping peaks. This resulted in 100,871 peaks used for caQTL mapping. As in our study, the number of PCs chosen for each data type empirically led to the identification of the largest QTL in each condition are reported in Table 2-3.

Study data set	Condition	Phenotype PCs regressed
Randolph eQTL	Non-infected	1 to 5
Randolph eQTL	Flu-infected	1 to 3
Nedelec eQTL	Non-infected	1 to 8
Alasoo caQTL	Non-infected	1 to 4
Husquin meQTL	Non-infected	None

Table 2-3. Principal components regressed for validation with previously published data sets.

STR calling and filtering

To robustly genotype the highly repetitive STR variants in our samples, we employed the HipSTR algorithm (v0.6.2)[156] which accounts for potential sequencing errors of STR introduced through PCR due to the highly repetitive nature of these sequences. Briefly, HipSTR models the PCR stutter noise of the repetitive sequence at each STR locus and determines the most likely STR allele using population-scale data and phased SNP scaffolds. We genotyped a set of 1,504,432 GRCh37 autosomal STRs smaller than 100 using GRCh37.hipstr_reference.bed.gz. We filtered the calls using HipSTR's supplied script with the recommended thresholds (min-call-qual=0.9, max-call-flank-indel=0.15, max-call-stutter=0.15, --min-call-allele-bias=−2, min-call-strand-bias=−2) to remove unreliable calls. This resulted in 1,465,954 robustly genotyped STRs.

STR genotype – phenotype analysis

We use the additive length of both alleles on a STR locus as the genotype to test STR genotype-phenotype association. Each STR locus can have more than three genotypes due to the multiallelic nature of STR length. To ensure that we were only using high quality STR calls, we further filtered the 1,465,954 aforementioned STR set to exclude STRs with call rate < 90% across all samples and STRs with minor allele frequency less than 10%. After filtering, we obtained 442,509 STRs used as input for Matrix-eQTL analysis.

STR-QTL mapping was performed with the same inputs and parameters as the SNP-QTL mapping analysis described above. As with the SNP-QTL analysis, we accounted for unmeasured-surrogate confounders by PCA on the age and batch corrected expression matrices. The number of PCs chosen for each data type empirically led to the identification of the largest QTL in each condition for the STR mapping analysis are reported in Table 2-4 below.

eQTL	Non-infected	1 to 3
	Flu-infected	1 to 7
caQTL	Non-infected	1 to 5
	Flu-infected	1 to 5
H3K27acQTL	Non-infected	1 to 3
	Flu-infected	1 to 2
H3K27me3QTL	Non-infected	1 to 3
	Flu-infected	1 to 3
H3K4me1QTL	Non-infected	1 to 4
	Flu-infected	1 to 5
H3K4me3QTL	Non-infected	1 to 3
	Flu-infected	1 to 4
meQTL	Non-infected	None
	Flu-infected	None

Table 2-4. Principal components regressed for STR-QTL mapping.

SNP v. STR analysis

We used a linear model to evaluate the proportion of variance explained (PVE) by the top SNP and top STR on a feature for each genomic phenotype using the R package relaimpo[140]. We used a model analogous to QTL mapping adapted to the requirements of relaimpo. The expression values and regressors used to model PVE are closely matched to the ones used for QTL mapping. We used the same batch corrected counts as described in the section "*Infection effects: Infection-Related Differential Effects"*, moving age to the regressor in the model such that the relative importance of the variants could be compared in situations that a feature/gene was only associated

with either an SNP- or STR-QTL (i.e., relaimpo requires at least two variables to be included in the model). For DNA methylation data, we used unadjusted, quantile normalized expression value and model it with batch, age, genotype of best associated SNP, and genotype of best associated STR. The relative importance of each regressor to the total variance of the linear model was then reported using the calc.relimp function. We chose to report the lmg relative importance metric as recommended by the package, which outputs the R^2 of each regressor partitioned by averaging over orders.

Identification of condition-specific and shared QTL

As in the popDE analysis, we use both qvalue and lfsr to determine if QTL are condition-specific (i.e., only showing an effect in the non-infected or flu-infected conditions) or shared (i.e., showing an effect in both conditions). Specifically, we require SNP/STR-feature pairs to have both FDR <.10 and lfsr < .10 in only one condition to be considered condition specific. If either the SNP-feature or STR-feature pair was found to be condition specific using these thresholds, the feature was classified as condition specific. QTL were classified as shared if FDR < .10 and lfsr <.10 in both conditions for either the SNP or STR.

Enrichment of TF binding sites among condition-specific SNP QTL

To investigate if condition-specific SNP QTL overlap transcription factor footprints at a significantly higher rate than non-QTL SNP-feature pairs, we used transcription factor footprints detailed in the previous section "Transcription factor activity scores". Briefly, we overlapped TF footprints with TF motifs and corresponding cluster information. For each data type and condition, we extracted the best SNP for each condition-specific QTL (detailed in "Identification of

condition-specific and shared QTL") and marked if it overlapped with a TF footprint for each of the 200 clusters. This resulted in a matrix containing either 0 (no overlap) or 1 (overlap) for each of the 200 clusters. We collected the same information for the best SNP of all SNP-feature pairs that were not significant (FDR ≥ .10) for that condition, which we used as the background set (for WGBS we randomly sampled 500k from the non-significant pairs). We then performed a logistic regression in R for each cluster to determine if there is a relationship between QTL type (condition-specific v. non-significant SNP-feature pairs) and if the SNP falls within a TF footprint.

QTL integration across the data types

To determine if SNPs that are QTL (FDR <.10) in one data type are also QTL for other data types we performed the following steps for each data type in each condition: i) collect all significant SNP-gene pairs FDR<.10 (not just the "best" SNPs) for the data type ii) for each of the 6 additional data types, select the top p-value for each feature using the list of significant SNPs for each feature iii) use this top p-value to determine if the SNP is a QTL for the additional data types. By performing this analysis from the perspective of each data type, we ask what percent of QTL are specific to data type or shared among patterns of data types.

To derive a null expectation for the observed overlaps we did the following. Take as an example the expected overlap between eQTLs and the other 6 additional epigenetic QTLs. First, we collected the list of all SNPs that are eQTL (FDR<.10) in the original data. Then, we asked how many of these are also significant for each of the 6 additional data types but using the p-values derived from the permuted results (described in "SNP genotype-phenotype association analysis"). Ultimately, by doing so, we are testing how often we expect to see an overlap between, in this example, an eQTL and other epigenetic QTL just based on the number of association tests

performed. These analyses were performed from the perspective of each data type separately. The null expectation represents the probability of finding an overlap by random chance simply based on the number of tests performed. The fact that it is not zero does not indicate statistical inflation, but instead the number of significant associations identified for each of the marks.

Enrichment of TF Binding Sites among shared QTL

To test if shared QTL are more likely to disrupt TF binding than those that are not shared, we modified the QTL integration pipeline to collect *all* p-values for each feature using the list of significant SNPs rather than just the top p-value. We then used each p-value to determine if the SNP is a QTL for the additional data types. For each condition, we took the union across all data types of the SNPs that were shared in 0, 1, etc. data types. For each union set, we marked if each SNP overlapped with any TF footprint. We also collected this information for all SNPs that were tested for QTL mapping to use as the background set. We then performed a logistic regression in R for each union set against the background set to determine if there is a relationship between the number of QTL a SNP is shared across and if the SNP falls within a TF footprint.

Relationship between epigenetic QTL at baseline and transcriptional response

For each data type separately, we tested the relationship between epigenetic QTL at baseline and transcriptional response. We created a meta genotype for each QTL (using the best SNPs only). We extracted the direction of the effect size from the QTL mapping results at baseline and categorized the 3 possible genotypes as either low/low, low/high, or high/high depending on the direction of the effect. Using the closest gene for the feature, we matched the meta genotype for each individual to the gene expression fold changes quantile normalized to a standard normal

distribution. We also tested the relationship between epigenetic QTL and gene expression at baseline and after flu infection using the same steps as above but matching with quantile normalized expression at baseline or after flu infection.

Enrichment of QTL within popDE features

We tested for an enrichment of QTL among popDE features within each condition. For each condition, we created two vectors: i) a popDE feature vector, where significant features (FDR<.10) were coded as 1 and non-significant features were coded as 0, and ii) a QTL vector, where we extracted the top SNP and STR for each feature and indicated the presence of a significant QTL if either the top SNP or STR feature pair was significant (FDR<.10) by coding a 1. Non-significant variant-feature pairs were coded as 0. A logistic regression was performed using the popDE feature and QTL vectors using glm in R (popDE_status[0,1] ~ QTL_status[0,1]). The odds ratios were converted to log2 fold enrichments with a 95% confidence interval for plotting.

Calculating DeltaPVE of Admixture

To evaluate the impact of genetic variation on population differences we calculated ΔPVE of admixture for the significant popDE features (FDR<.10) that have both an associated SNP and STR (i.e., a SNP or STR within the QTL mapping window size for each data type). To calculate ΔPVE of admixture we first calculated the effect of admixture towards the total variance for each batch-corrected feature using the R package relaimpo[140] and the following model:

$$R^2 = PVE_{age} + PVE_{admixture}$$

Age was included in the model so that the relative importance of admixture could be compared. We then regressed out the effects of the top SNP and STR:

$$R^2 = PVE_{age} + PVE_{admixture} + PVE_{top\ SNP\ for\ feature} + PVE_{top\ STR\ for\ feature}$$

We then calculated the ΔPVE of admixture, which is PVE original - PVE variants regressed / PVE original.

Calculation of predicted and observed population differences

We estimated the predict cis-genetic population differences across the data types by comparing predicted and observed population differences. For each data type and condition, we extracted significant (FDR <.10) popDE features. We ran a model with inputs analogous to QTL mapping for each data type, but including both SNPs and STRs in the model, in addition to PC1 of the SNP genotype data. Only features that had both a SNP and STR tested were included. For WGBS data, age and batch are also included as regressors since they are not adjusted for in the input file. We regressed out the same number of expression PCs as in "STR PCs reg table". This resulted in QTL effect sizes with both variants considered. We then computed the predicted "expression" of each feature considering the QTL effect size of the top *cis* SNP for that feature from both the SNP and STR mapping analyses and an individual's genotype dosage (a vector of 0, 1, or 2 for SNP effect size, and a vector denoting the total length of the STR for the STR analysis). For each feature i, individual j:

$predicted\ expression_i = SNP\ QTL\ effect\ size_i * genotype_j + STR\ QTL\ effect\ size_i * genotype_j$

We modeled the predicted expression values using a model analogous to the popDE model (Y ~ Admixture) in each condition separately since the same features are not always popDE in both conditions. The previously described popDE outputs were used as the observed population differences.

Evaluating the impact of the top SNP and STR on popDE effects through GSEA

For each data type, we extracted popDE features that were significant in either the NI or flu condition and ran the same model as described in M2 but adding the top SNP and STR for each condition within the model. We used 5 permutations obtained by randomly reshuffling admixture estimates in order to estimate FDR, as described in detail above. We then ran GSEA as described in "GSEA of Infection effects and popDE effects" to compare the size and direction of ancestry-associated effects both before and after regressing out the top SNP and STR for each data type.

Colocalization analysis

We tested colocalization between the molecular QTLs and 14 well-powered GWAS including 6 unique autoimmune diseases and 4 unique inflammatory diseases (average N: 120132) listed below in Table 2-5.

Trait	N	PMID
Allergy and eczema	456899	29892013
Multiple sclerosis	115635	31604244
Rheumatoid arthritis	103638	24390342
Allergic diseases	242569	29083406
Adult-onset Asthma	327253	30929738
Atopic dermatitis	40835	26482879
IIBDGC – Crohn's disease	20883	26192919
IIBDGC - Inflammatory bowel disease	34652	26192919
IIBDGC - Ulcerative colitis	27432	26192919
Systemic lupus erythematosus	10995	27399966
Asthma	142486	29273806
Crohn's disease	40266	28067908
Inflammatory bowel disease	59957	28067908
Ulcerative colitis	58341	28067908

Table 2-5. Immune GWAS used in this study.

We first identified each lead GWAS SNP with P-value below 1e-05 and defined a locus as a 1Mb window centered around the lead GWAS SNP. A GWAS locus was moved from COLOC analysis if its lead SNP overlaps the HLA region (chr6: 25Mb-35Mb). Colocalization test was only performed when the most significant QTL of a feature falls within a defined window (100Kb for RNA-seq, ATAC-seq, H3K27ac, H3K27me3, H3K4me1 and H3K4me3, and 5Kb for WGBS) around the lead GWAS SNP. We used "coloc.signals" function from COLOC package (v5.1.0) with default priors[107]. We defined colocalization as PP3+PP4 > 0.5 and PP4/(PP3+PP4) > 0.8.

Imputation of SNPs for heritability analysis

We performed imputation using the same 7,383,243 SNPs for QTL mapping in order to eliminate missing genotypes, as required by the heritability analyses described below. Briefly, we used Genotype harmonizer[126] to harmonize the strand with the 1000 Genomes reference panel. We used SHAPEIT to phase the haplotypes[157] prior to imputation with IMPUTE v2 using one phased reference panel (1000 Genomes)[158]. We imputed each chromosome in 5 MB intervals and used the "pgs_miss" flag to replace only the missing genotypes.

Fine-mapping molecular QTLs

To better identify likely causal variants, we performed fine-mapping of molecular QTLs using the Bayesian statistical fine-mapping tool SuSiE[159]. We used SuSiE with individual-level phenotype and genotype data and set the maximal causal variants per region parameter L = 3.

We fine-mapped molecular QTLs with distance-based informative prior inclusion probability, so that a SNP close to a gene or peak would have a higher prior probability of being a causal variant. In specific, we separated molecular QTLs into distance bins, with six distance bins

(<500bp, 500bp-1kb, 1kb-2kb, 2kb-5kb, 5kb to 10kb, and 10kb-100kb) for eQTL, caQTL and histone QTLs, and four distance bins (<500bp, 500bp-1kb, 1kb-2kb, and 2-5kb) for methylation QTL. We used the Bayesian statistical tool Torus[160] to estimate the enrichment for the distance bins, compute SNP-level priors using the estimated distance enrichment estimates for each locus.

Heritability and enrichment analysis of GWAS summary statistics using S-LDSC

To partition the heritability of complex traits and estimate heritability enrichment for each type of molecular QTLs we used Stratified LD score regression which assesses how the heritability of a complex trait is partitioned among functional features, while controlling for LD, allele frequency and other baseline features (S-LDSC, baseline v2.2)[110–112]. S-LDSC estimates the heritability enrichment as a ratio of the proportion of heritability explained by an annotation divided by the proportion of SNPs in that annotation.

For the enrichment analysis, we constructed a continuous annotation using the posterior inclusion probability (PIP) from SuSiE fine-mapping with distance-based prior. We applied S-LDSC separately for each type of molecular QTL annotations. In our S-LDSC analysis, we adjusted for various baseline annotations of SNPs using a generic baseline LD model, including gene annotations (coding, UTRs, intron, promoter), MAF bins and LD-related annotations. We did not include functional annotations such as enhancer markers in our baseline model, because these annotations are likely correlated with our QTL features of interest and may bias our estimated enrichment. That said, we did run two other versions of the s-LDSC analysis ensure that our conclusions were robust to changes in the baseline model: 1) adjusting for histone mark annotations included in the S-LDSC full baseline derived from multiple cell types as defined by the Roadmap Epigenomics project, 2) adjusting for histone mark data collected in the current

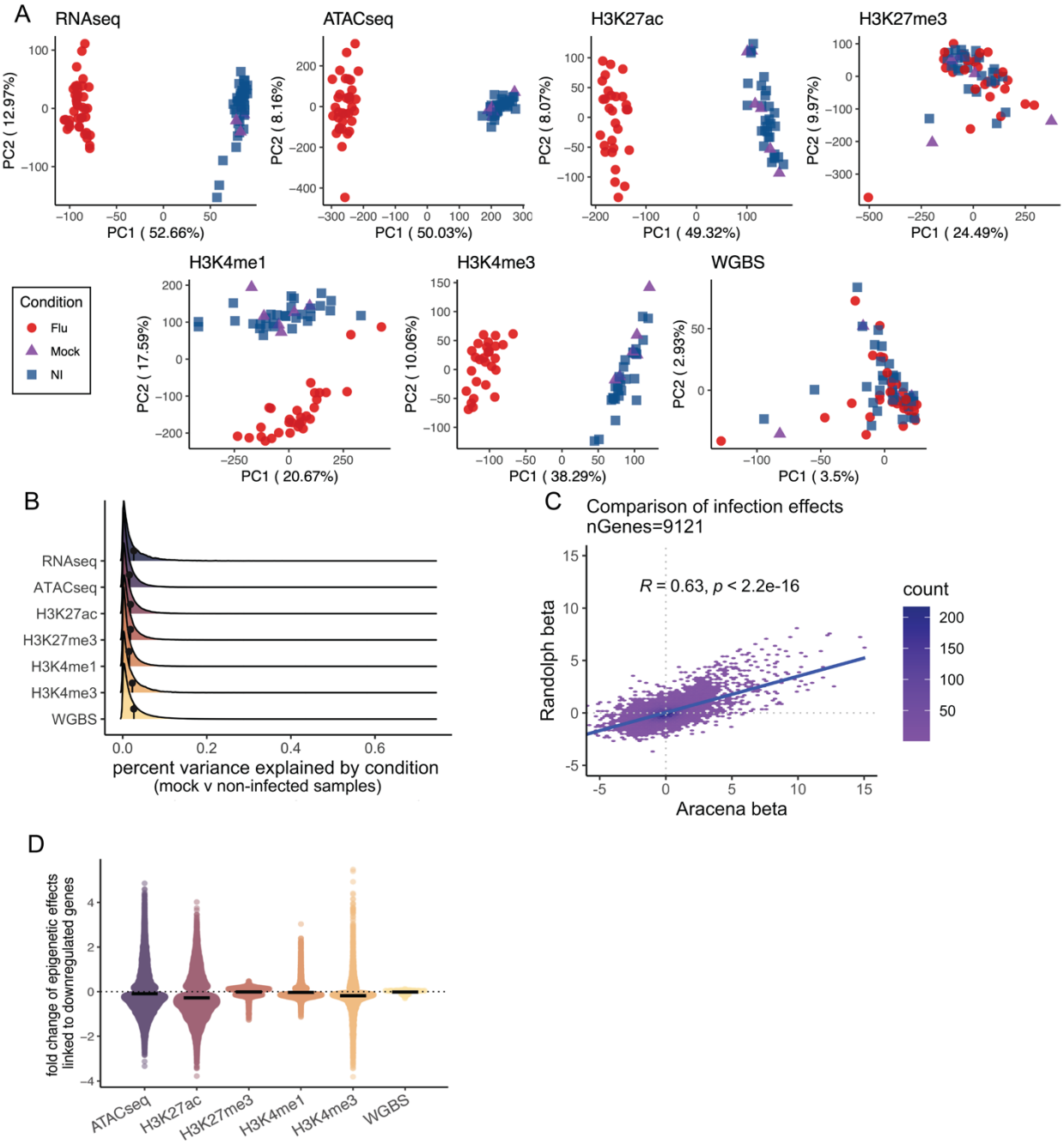
study. As shown in Figure S2-10C, the enrichments are somewhat reduced after adjusting for histone marks in our study, but the overall conclusions remain the same. We caution that the enrichment of heritability in histone marks is partially mediated through the QTLs we found. Thus, adjusting for histone marks when testing enrichment in QTLs will inevitably result in a conservative estimate of the true enrichment.

To estimate heritability explained by molecular QTLs, we constructed a binary annotation containing all SNPs with SNP-level FDR < 10% since the GWAS (same as detailed above in "Colocalization analysis") used have only been performed on SNPs. We note that the exact values of the heritability estimates may be biased as we have only 35 individuals from a mixture of European and African populations, but the relative heritability estimates should reflect the relative contributions of different molecular QTLs to these complex traits.

Estimation of the association between genomic marks and immune disease

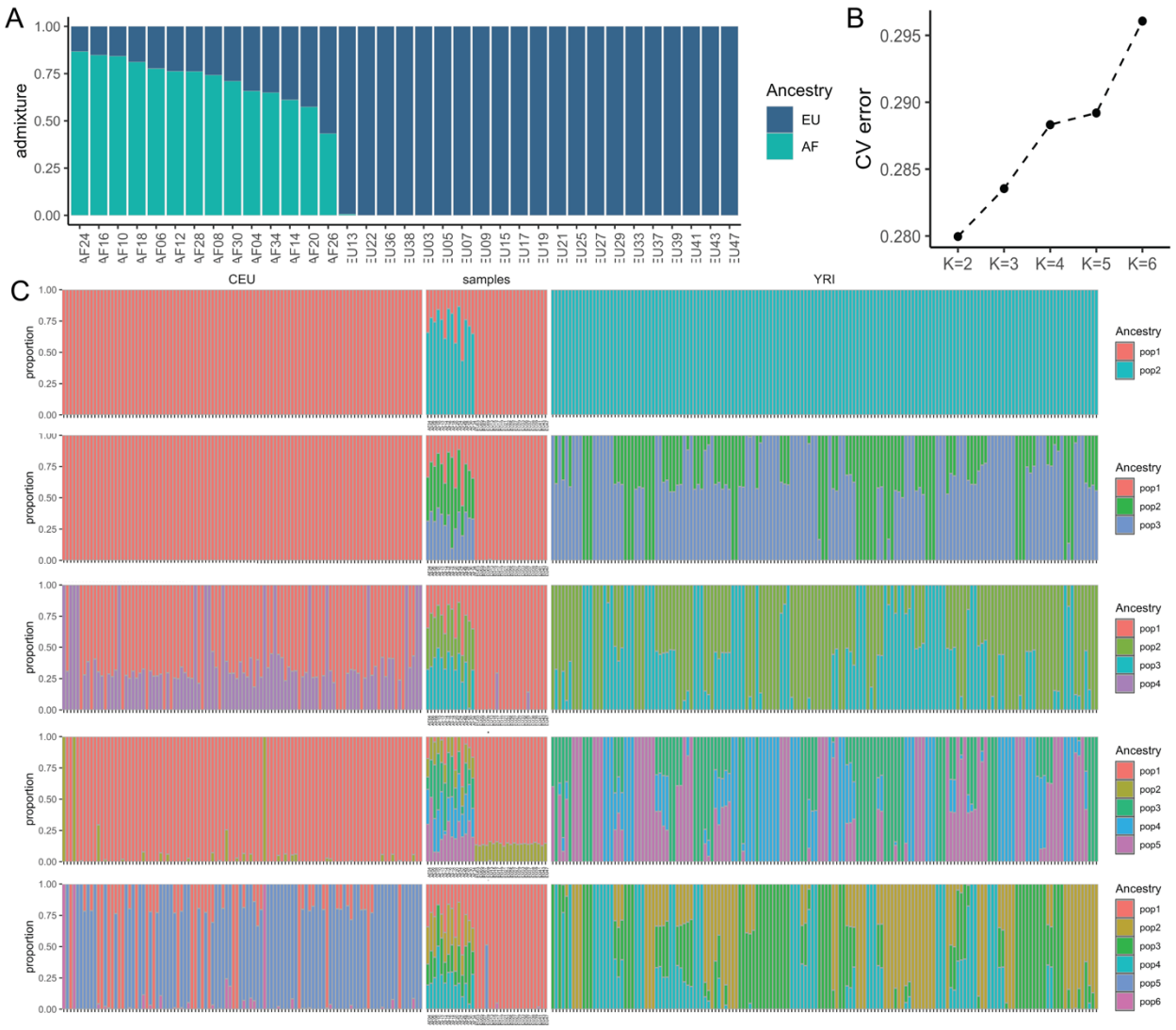
We used S-PrediXcan[113] to estimate the association between immune system disorders and the expression of genes and epigenetic marks. S-PrediXcan requires prediction models that describe the association between an aggregate of SNPs and the expression of nearby genomic marks. However, instead of explicitly predicting the genetically determined component of expression, it requires only the summary statistics of GWAS studies to assess the association between a genomic mark and a disorder. We trained a set of prediction models of gene and epigenetic mark expression in both non-infected and flu-infected conditions using the same genotype and phenotype data used for QTL mapping. We used summary statistics from the same 14 GWAS studies previously described to identify the genes and epigenetic marks involved in these disorders. We used the beta and p-value of SNPs from the GWAS summary statistics when available in order to compute the

association between the molecular traits and disorders. Otherwise, we used odd ratios. We apply Bonferroni correction to determine the condition and data type specific p-value cut off and identify genes and epigenetic marks that are significantly associated with the immune disorders. We mapped the epigenetic marks to their closest genes using the annotatePeak function of CHIPseeker using the same parameters previously described.

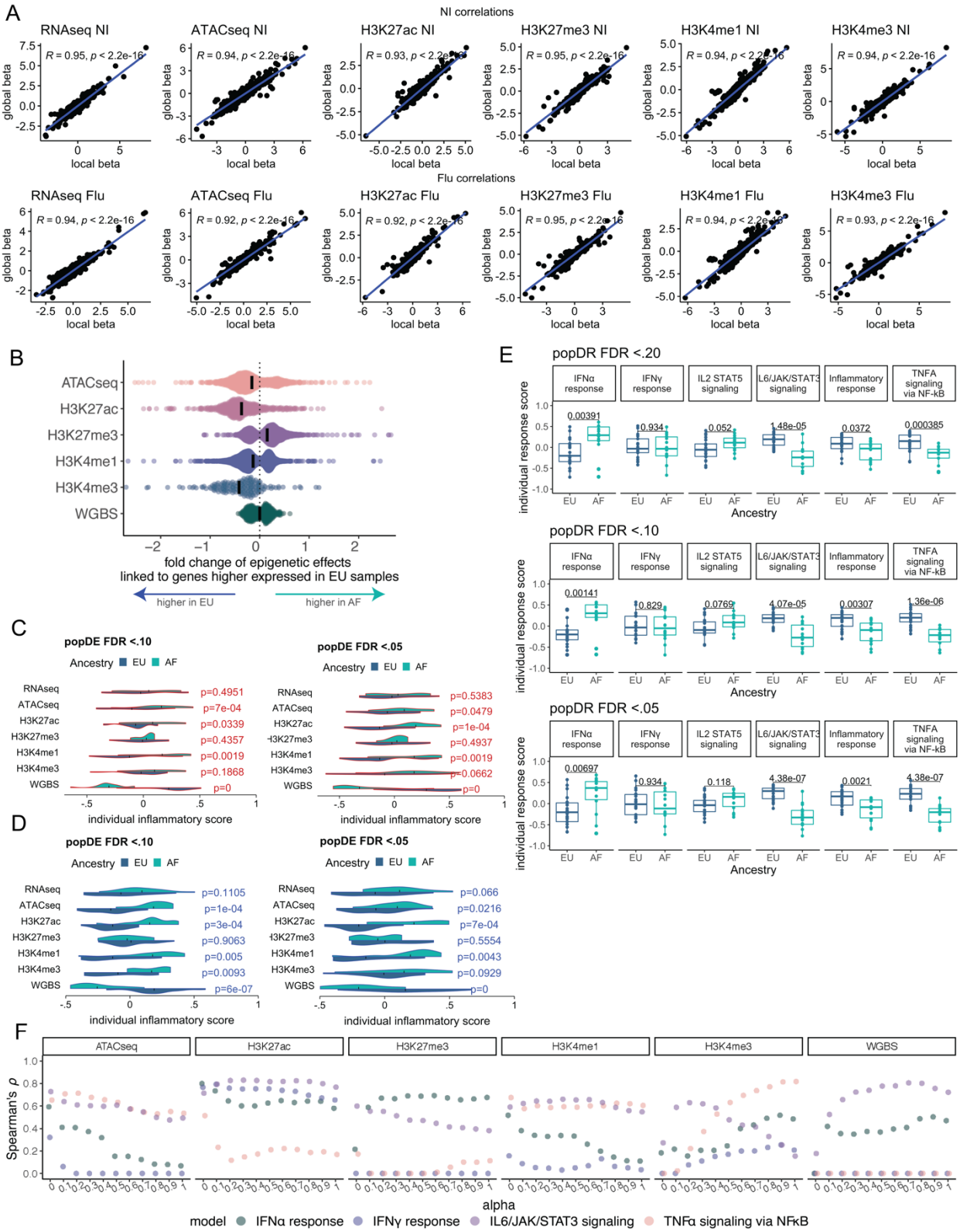


Supplementary Figure 2-1. Genome-wide impact of flu infection across regulatory marks. (A) Principal Component Analysis read counts showing for all data types the separation of NI and Flu samples along the two main axes of variation. Mock samples cluster with NI samples, revealing little technical variation. (B) PVE by mock versus NI samples. (C) Comparison of infection effects for the union of genes tested in our study and Randolph et al., 2021. (D) Distribution depicting the relationship between gene expression changes and epigenetic changes in response to flu infection as seen in Fig. 2-1E but here focusing on epigenetic changes nearby genes that are downregulated in response to infection. Downregulated genes are defined as genes

Supplementary Figure 2-1, continued. with beta < -0.5 and FDR<.01. Epigenetic changes are those with FDR<.01, except for methylation changes (FDR<.20).

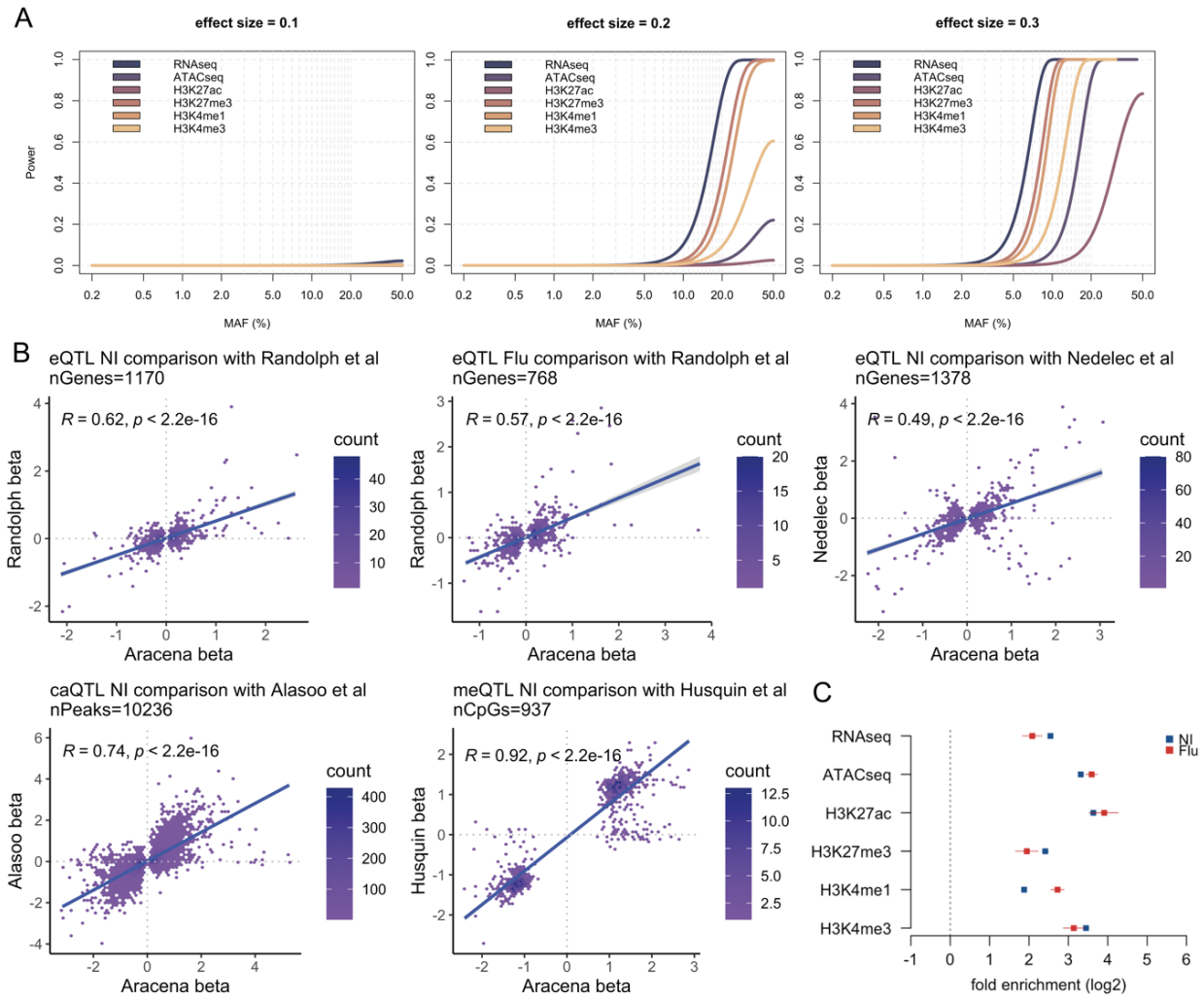


Supplementary Figure 2-2. Estimation of genetic ancestry. (A) Quantitative genetic ancestry proportions partitioned into European (dark blue) and African (turquoise) components for each individual. (B) Plot of cross validation errors from ADMIXTURE analysis. (C) Varying K from 2 to 6 in ADMIXTURE analyses.

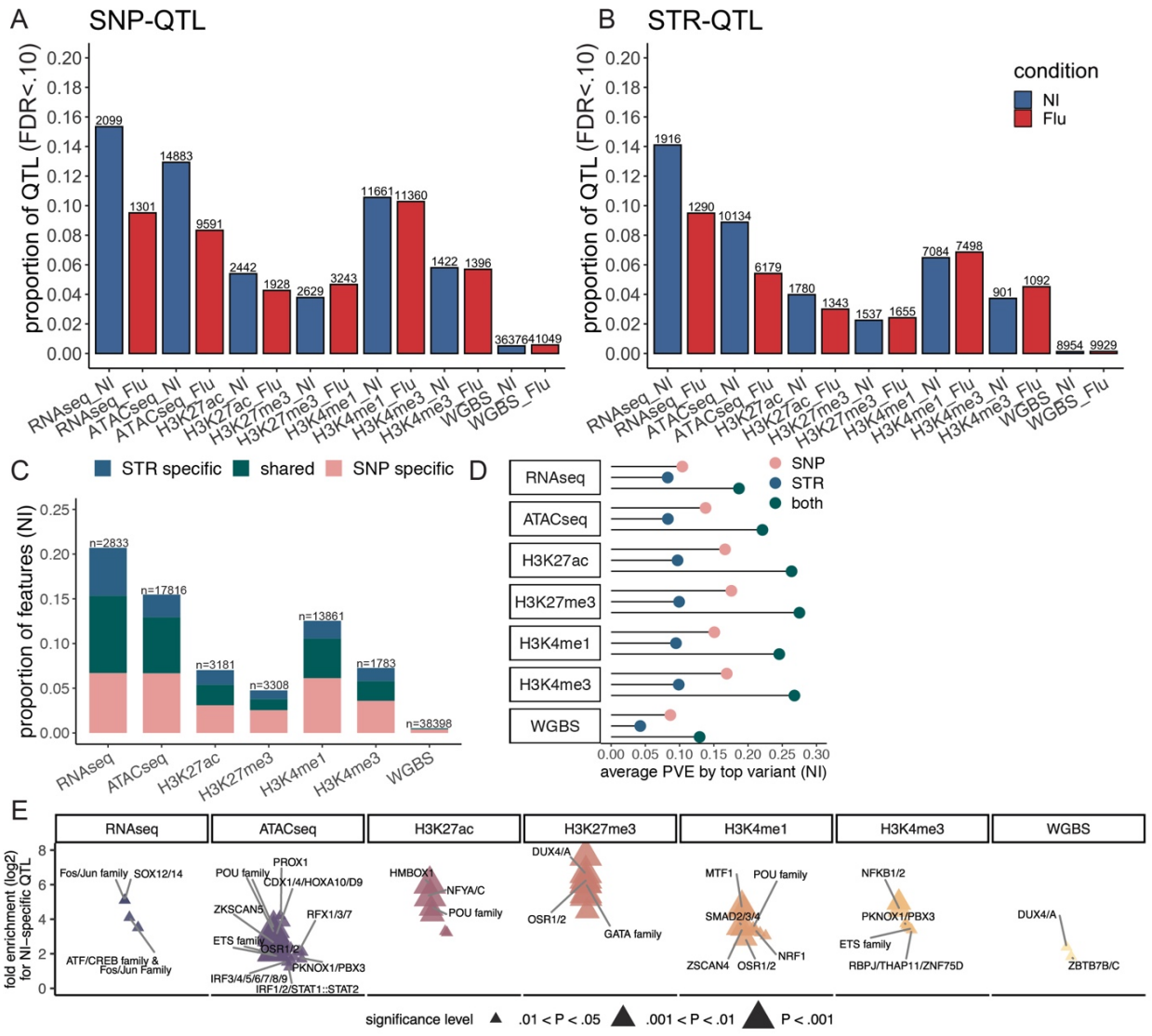


Supplementary Figure 2-3. Classification of ancestry-associated differences. (A) Correlation of population differentially expressed (popDE) effects calculated with global or local ancestry effects. (B) Distribution depicting the relationship between popDE genes and popDE epigenetic

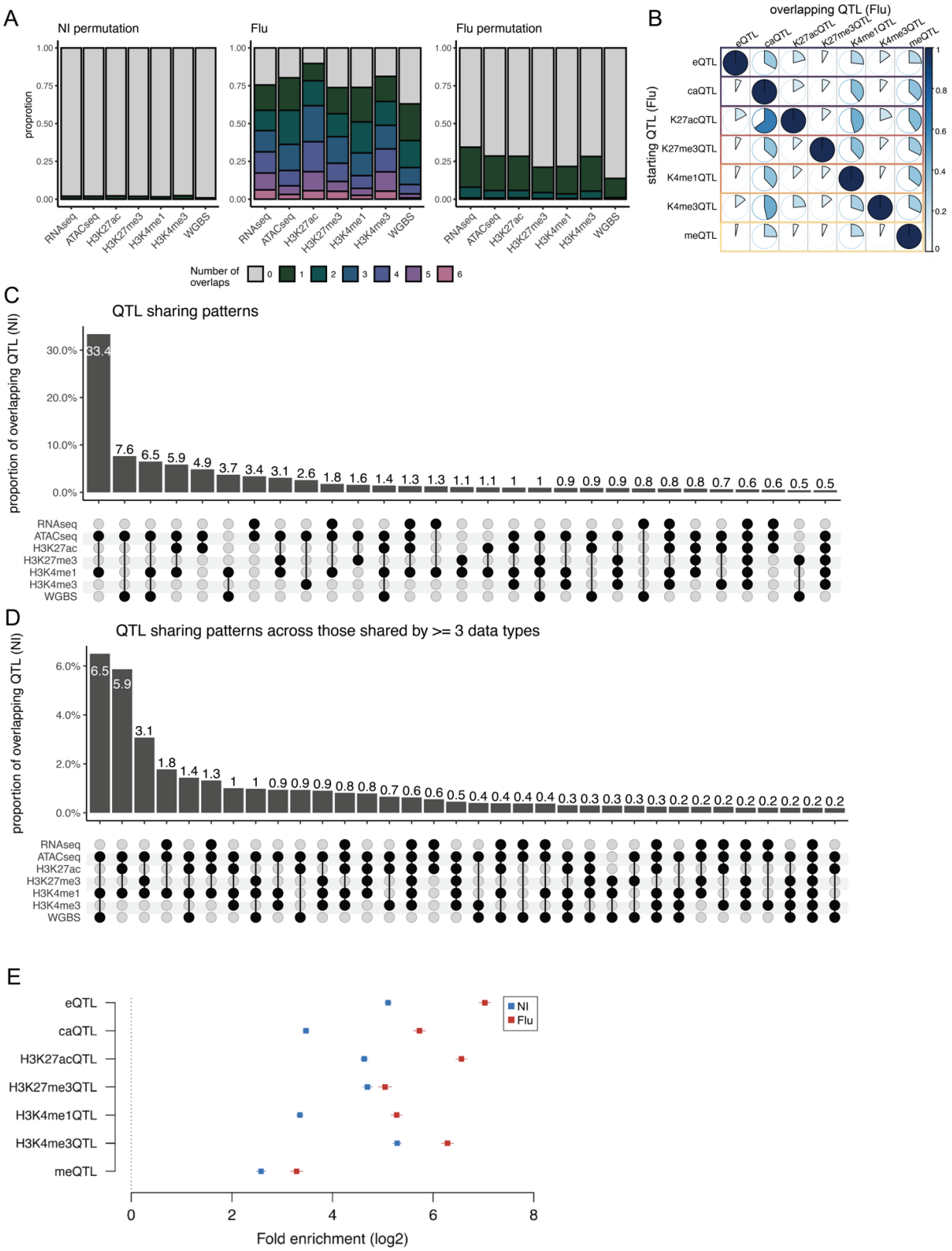
Supplementary Figure 2-3, continued. changes across both conditions. Genes more highly expressed in individuals with high proportions of European ancestry (fold change < -0.5, FDR< 0.10) are nearby popDE epigenetic regions (FDR <.10) that show increased levels of chromatin accessibility, H3K27ac, H3K4me1 and H3K4me3 in individuals with increased European ancestry levels. Black lines represent means. (C) Distributions of individual mean scores of inflammatory pathways in the flu-infected condition comparable to Fig. 2-2C which shows non-infected condition distributions. A higher score indicates a strong expression of genes or epigenetic marks nearby genes within the Hallmark inflammatory response pathway. (D) Individual mean score differences between the population-groups for the Hallmark "inflammatory pathway" in the non-infected and (C) flu-infected conditions remain consistent when reducing popDE effects FDR from 10% to 5% (E) Population-group differences utilizing popDR effects to calculate individual transcriptional response score across 6 immune pathways remain consistent with varying FDR thresholds (20%, 10% and 5%). (F) The distribution of Spearman's correlation between the predicted and observed mean scores for the various pathways using different alphas.



Supplementary Figure 2-4. Power calculations and validation of the QTL identified using external data sets. (A) Power calculations for QTL with effect sizes ranging from 0.1-0.3. Power to detect QTL increases as the effect size of the variant increases. (B) Validation of significant FDR <.10 QTL in our dataset. First row: Left- Comparison of non-infected eQTL with Randolph et al. Middle- Comparison of flu-infected eQTL with Randolph et al. Right- Comparison of non-infected eQTL with Nedelec et al. Second row: Left- Comparison of non-infected caQTL with Alasoo et al. Right- Comparison of non-infected meQTL with Husquin et al. (C) ASE hits are enriched for QTL.

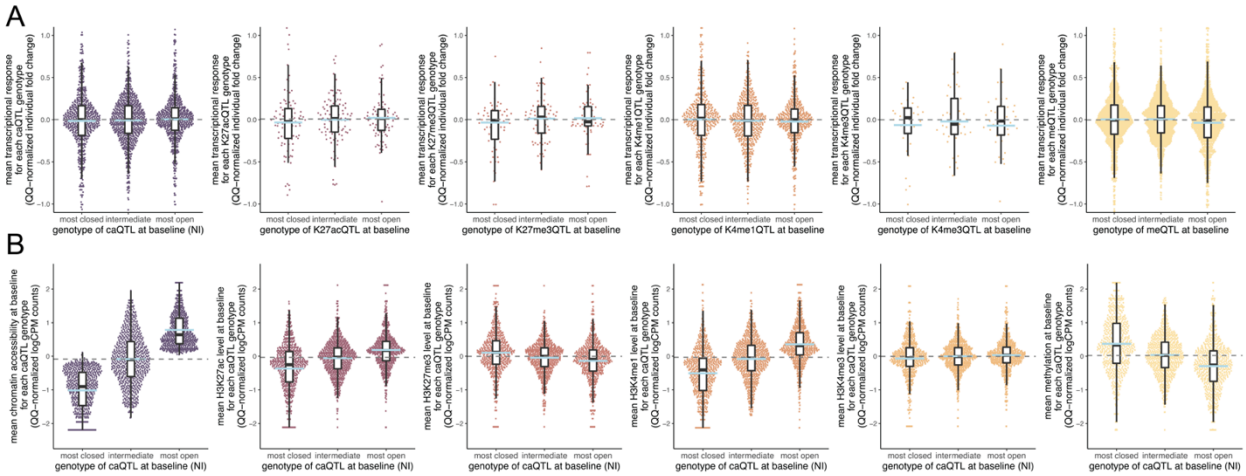


Supplementary Figure 2-5. QTL mapping of the different molecular traits. (A) Proportion and number of SNP-QTL at a significance threshold of FDR <.10 in each condition (B) Proportion and number of STR-QTL at a significance threshold of FDR<.10 in each condition. (C) Proportion and number of genes/features associated with at least one SNP or STR QTL in non-infected macrophages. Shared QTL were defined as those genes/features associated with a QTL at an FDR<.10 when performing the QTL mapping against SNPs and STRs separately. SNP- or STR-specific are those only identified as significant (FDR<0.1) against either SNPs or STRs. (D) The mean percent variance explained by the top SNP and STR across all features in the non-infected condition. Both is the sum of the PVE of the top SNP and top STR (E) The enrichment of TF binding sites across non-infected specific SNP-QTL. TF clusters are shown.

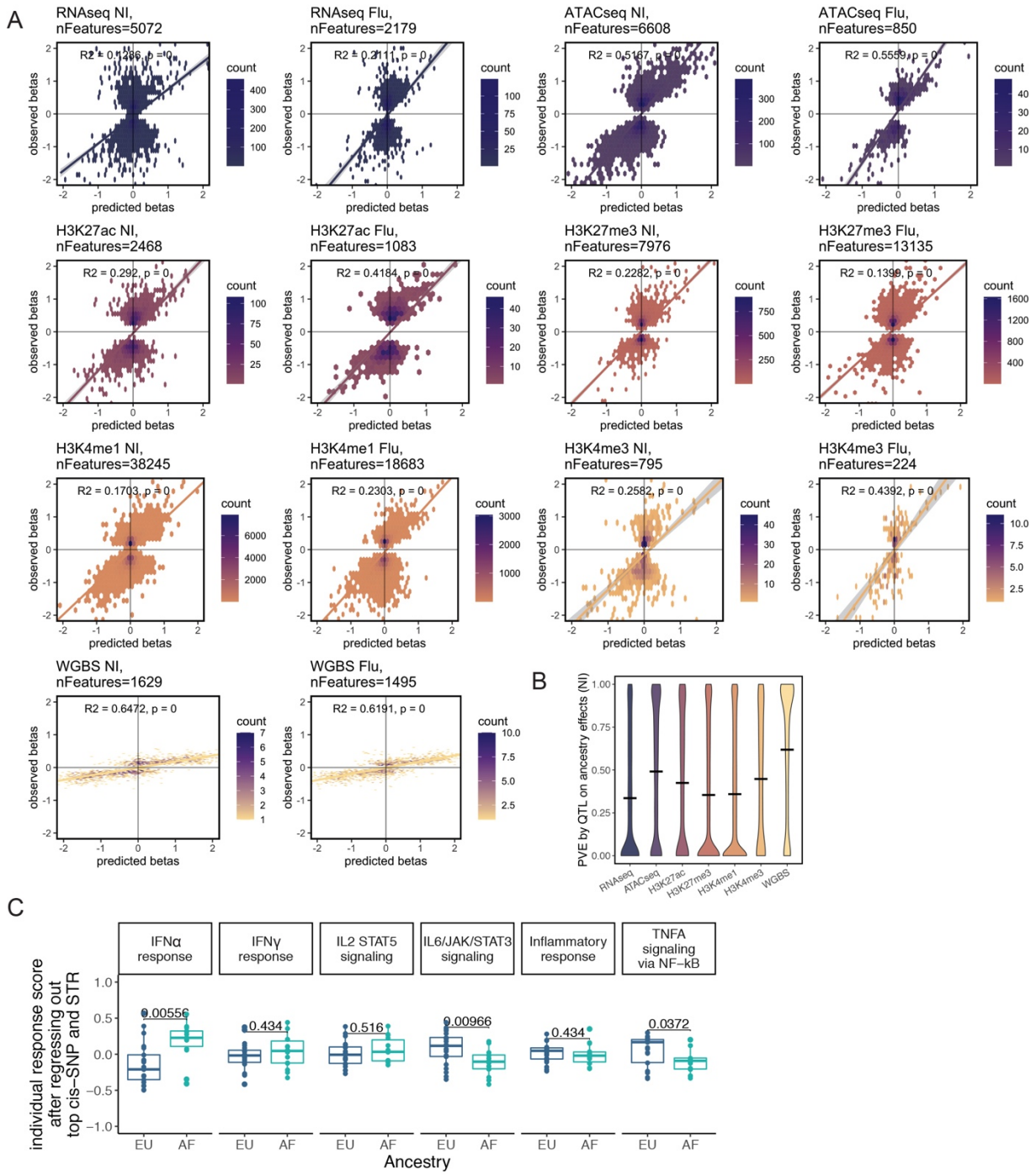


Supplementary Figure 2-6. Overlap of QTL across molecular traits. (A) *Left*: The number of overlaps for each QTL type for the permuted analysis in the non-infected condition. More than

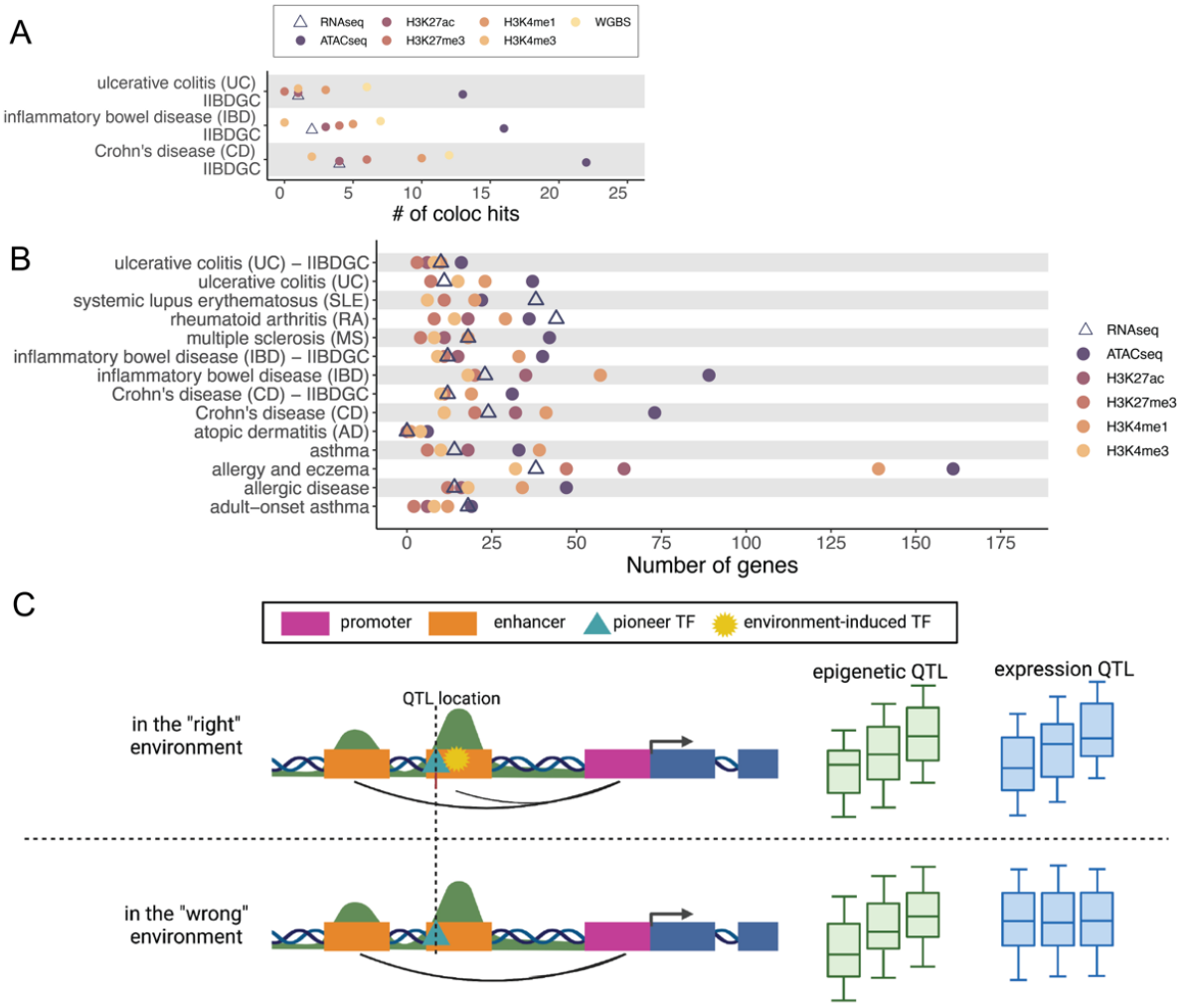
Supplementary Figure 2-6, continued. one overlap indicates the QTL is shared with at least one other datatype. *Center*: The number of overlaps for each QTL type in the flu-infected condition. *Right*: The number of overlaps for each QTL type for the permuted analysis in the flu-infected condition. (B) The percentage of QTL in one data type that are also QTL for another data type in the flu-condition. The starting QTL (rows) are the QTL that are tested for sharing while the overlapping QTL (columns) are the percentage of each starting QTL that are shared with that datatype. The color of each circle corresponds to the percentage of sharing. (C) QTL sharing patterns for those QTL overlapping 2≥ data types) in the non-infected condition. Y axis the proportion of overlapping QTL (i.e., the denominator is the number of QTL that are shared in at least 2 or more data types). (D) QTL sharing patterns for those QTL overlapping 3≥ data types) in the non-infected condition highlighting that caQTL, K4me1 QTL and meQTL are the most commonly shared. The Y axis is the same as described in (C) above. (E) eRNA-QTLs are highly enriched for other regulatory QTL.



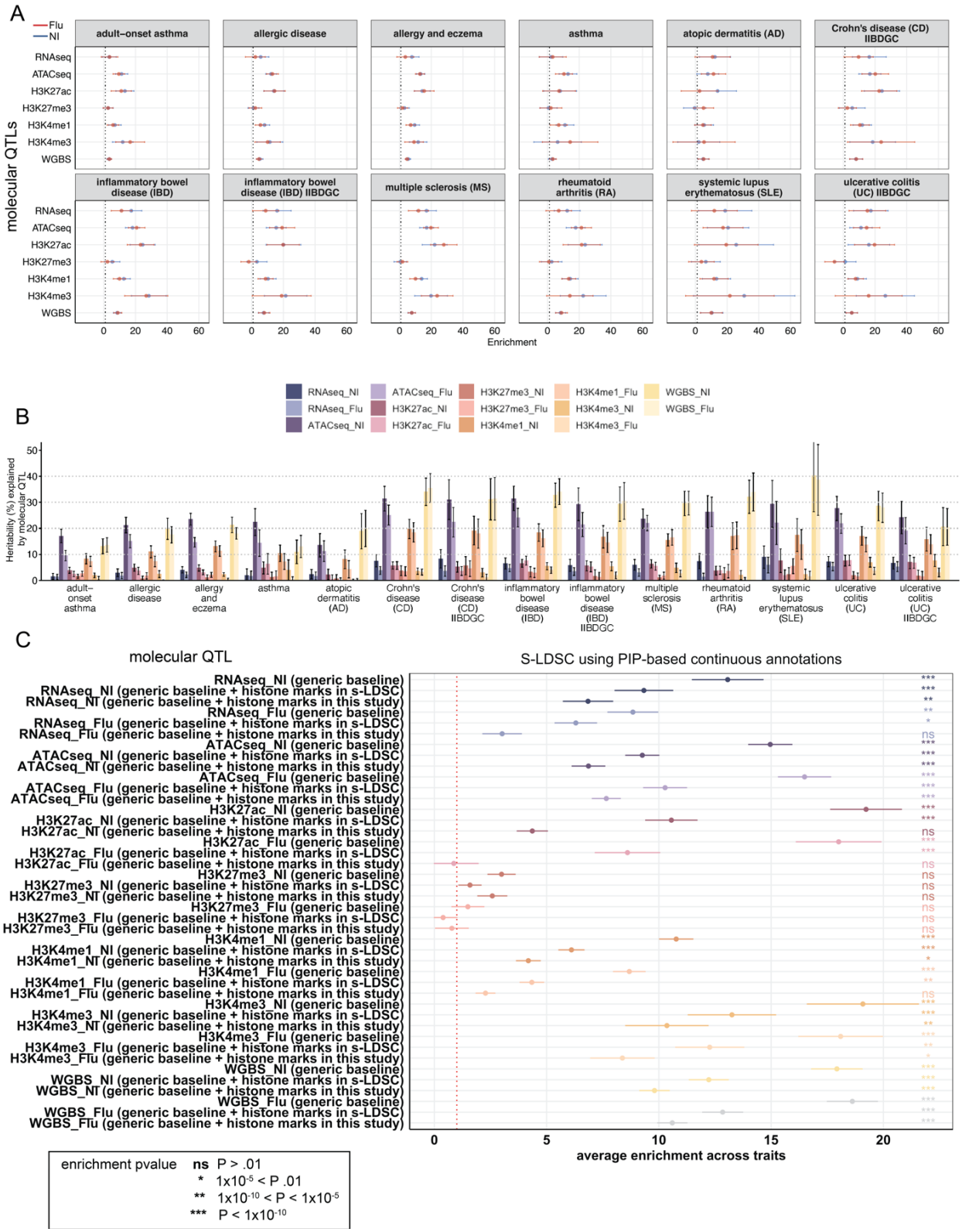
Supplementary Figure 2-7. Genetically driven variation in epigenetic levels has no impact on the magnitude of transcriptional responses upon IAV infection. (A) Genotypes for epigenetic QTL at baseline have no impact on the transcriptional response of nearby genes. The light blue marks the mean for each genotype and gray the median across all genotypes. As detailed in Fig. 2-4F, we restricted to QTL nearby upregulated genes that are not eQTL. (B) Association between genetically encoded baseline differences in chromatin accessibility and baseline differences in other epigenetic marks. *Left*- Meta caQTL plot (at baseline condition) across caQTLs for accessibility regions associated with up-regulated genes (n=681 caQTLs associated with 506 genes). Individuals with genotypes associated with increased chromatin accessibility also show significantly increased levels of H3K4me1 and H3K27ac ($P<2.2x10^{-16}$), and to a lesser extent, a reduction in the repressive mark H3K27me3 ($P<1.15x10^{-10}$).



Supplementary Figure 2-8. Calculating the contribution of cis-acting regulatory variants to ancestry-associated differences. (A) Correlations between the observed and predicted betas for significant population differentially expressed (popDE) features (FDR<.10) for each of the data types in both conditions (Pearson's correlation coefficient reported). (B) Boxplot of the ΔPVE of admixture for each feature in each data type in the non-infected condition (flu-infected condition shown in Fig. 2-5C). (C) Boxplots of individual transcriptional response scores after regressing out the effects of the top SNP and STR in each condition for the 6 immune response pathways.

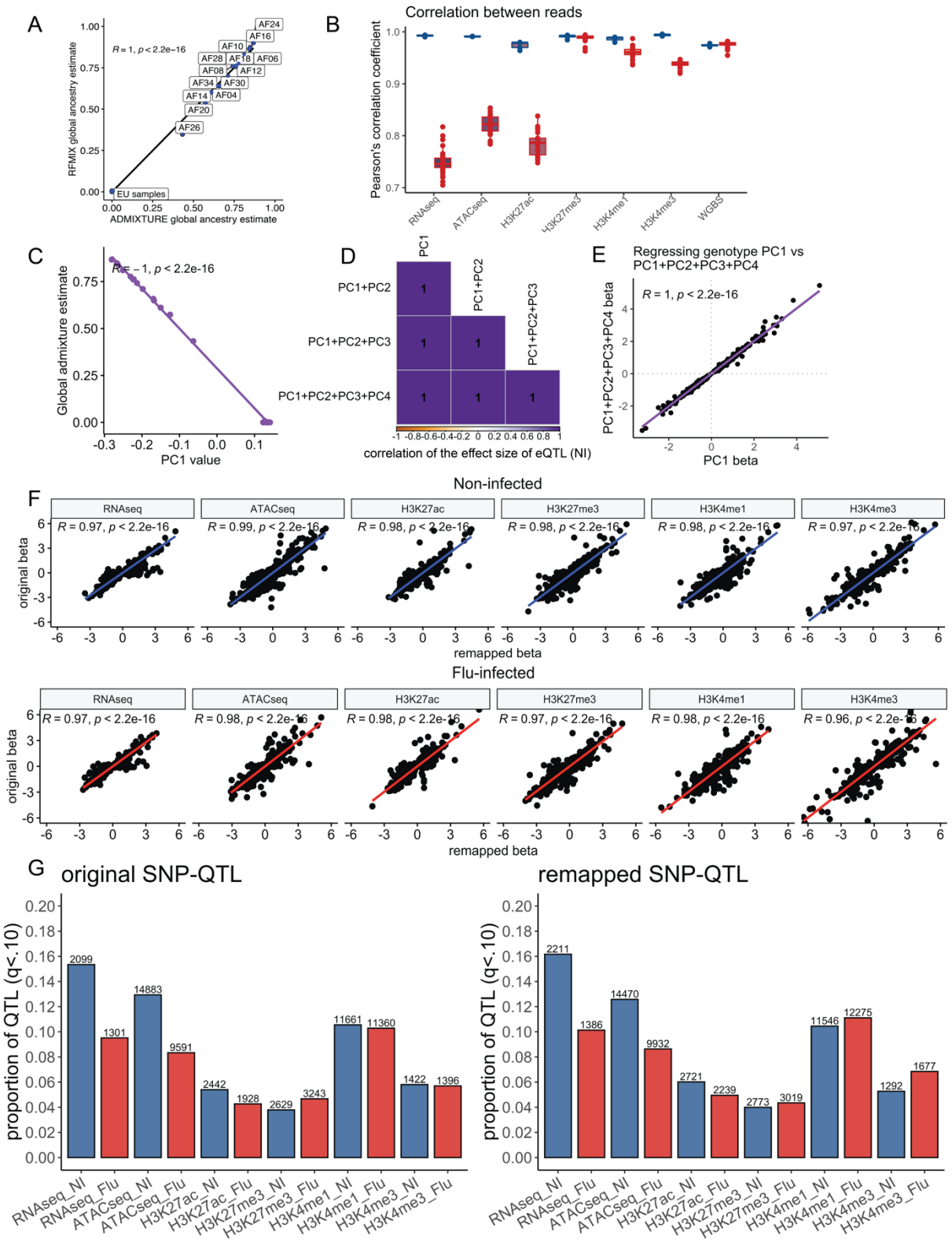


Supplementary Figure 2-9. Epigenetic QTLs overlap with genetic variants associated with immune-related diseases. (A) Summary of colocalization results for duplicated immune related diseases (11 diseases were investigated through 14 GWAS). Points represent the number of significant hits defined as PP3+PP4 > 0.5 and PP4/(PP3+PP4) > 0.8 in either condition. (B) Summary of PrediXcan results. Each point represents the total number of genes (Bonferroni corrected p=0.05) associated with the disease trait in either condition. A gene is only counted once even if multiple peaks are associated with the gene. (C) Schematic depicting the proposed hypothesis that epigenetic QTL may act as a proxy for genetic variation that under particular environmental conditions has an impact on gene expression levels. Blue boxes represent gene exons and green peaks represent ATACseq peaks. A genetic variant at the QTL location impacts TF binding, such that differential binding of the TF is associated with variation in chromatin accessibility (i.e., an caQTL). If the activity of this enhancer requires the recruitment of an additional TF (here labelled "environment-induced TF") only induced in response to specific environmental/developmental conditions, the caQTL will not be associated with variation in gene expression levels. Yet, this caQTL will be a proxy for a genetic variant that on the "right environment" will ultimately be associated with an eQTL. Under this model, epigenetic QTLs that colocalize with GWAS variants (but not with eQTLs) can be thought of as a means to identify genetic variants that have an impact on gene expression in a yet unmeasured environment.



Supplementary Figure 2-10. Heritability explained by molecular QTL. (A) Heritability enrichment results for all 14 GWAS. A 95% confidence interval is displayed. (B) Bar plots, with standard error, representing the percent of heritability explained by each of the molecular QTL in

Supplementary Figure 2-10, continued. all conditions. (C) Average heritability enrichment across 14 GWAS studies, comparing s-LDSC results using generic baseline, generic baseline and adjusting for s-LDSC's histone marks, as well as generic baseline and adjusting for histone marks from the current study. s-LDSC analysis was conducted on finemapped molecular QTLs (using fine-mapping tool SuSiE), treating PIPs from fine-mapping results as continuous annotations. The average enrichments across all 14 GWAS studies (error bars represent standard errors) are plotted with the p-value of enrichments from meta-analysis of the 14 GWAS studies using Fisher's method.



Supplementary Figure 2-11. Validation of technical effects, correction for population structure and mapping bias. (A) Correlation of ADMIXTURE and RFMIX global ancestry

Supplementary Figure 2-11, continued. estimates revealing a 1:1 correlation. (B) Boxplots of the Pearson's correlation coefficient with mock samples showing high correlation between voom normalized mock and non-infected reads. (C) Correlation of PC1 and global ancestry estimates across the samples. (D) Correlation of the effect size of eQTL (FDR < .10) with 1-4 genotype PCs as covariates. (E) Example correlation of (D) between PC1 and PCs 1-4. (F) Effect sizes of reported results highly correlate with QTL mapping performed with remapped reads. (G) Comparison of proportion of QTL at FDR <.10 performed with original reads (left) and remapped reads (right).

Supplementary Tables

Supplementary tables for this chapter are described in Appendix: Supplementary Tables and available online.

Chapter III: DNA methylation-environment interactions in the human genome

Note:

The following section (*Chapter III*) is a summary of a project to which I contributed, titled "DNA methylation-environment interactions in the human genome" (Johnston et al. 2023). This paper was published on BioRxiv on May 19, 2023 and is pending publication in a peer-reviewed journal. My contribution evaluated the relationship between DNA methylation and transcriptional response to environmental challenge *in vivo* which is described in detail in the following chapter.

Authors:

Rachel A. Johnston, Katherine A. Aracena, Luis B. Barreiro, Amanda J. Lea, Jenny Tung

Introduction

Epigenetic changes, particularly DNA methylation are heritable, passed from parent to daughter cell across cell divisions[15,17]. Typically, the presence of DNA methylation is indicative of gene repression, as the presence of the methyl group can block the cellular machinery required for transcription. While DNA methylation can be maintained for cell generations, it is also dynamic and adaptable to varied environments, and can be unique to cell types, stages of development, and environmental exposures[46]. Despite specific loci-environment contexts being well-characterized[161,162], it remains unclear how DNA methylation and the environment interact to enact functional changes on a genome-wide scale. That is, it remains unknown if DNA methylation-environment interactions directly cause changes in gene expression, likely through altering of transcription factor binding, or if DNA methylation changes are downstream consequences of gene regulatory differences[41,163]. In either scenario, methylation changes from previous environmental interactions can remain functionally relevant if they affect subsequent environmental challenges, for example in the case of reinfection[51,164].

Characterizing DNA methylation and environment interactions may provide greater clarity into the genomic regions and environmental conditions in which DNA methylation functionally impacts gene regulation. To that end, *Johnston et al*[165] uses mSTARR-seq to construct a genome-wide map of DNA-methylation dependent enhancer activity at baseline and in response to two environmental challenges: synthetic glucocorticoid dexamethasone (Dex), which inhibits inflammatory response and therefore has an anti-inflammatory effect, and interferon alpha (IFNA), a cytokine involved in the immune response to viral infections. The massively parallel reporter assay, mSTARR-seq, simultaneously tests for both enhancer-like activity and DNA methylation-dependent enhancer activity for millions of loci in a single experiment, allowing the causal effects

of DNA methylation to be directly tested (Fig. 3-1). In total, over 27.3 million CpG sites across the human genome were queried for methylation-dependent regulatory activity, revealing a subset of regions that were IFNA specific (n=1033), methylation dependent (n=2146), and both (n=881). To investigate if these DNA methylation-interactions replicate *in vivo*, we leveraged the dataset previously described in *Chapter II* to investigate mSTARR-seq identified IFNA regulatory regions and DNA methylation dependent IFNA regulatory regions, allowing us to test if DNA methylation levels at baseline affect transcriptional response to environmental challenge as previously hypothesized[116,166,167].

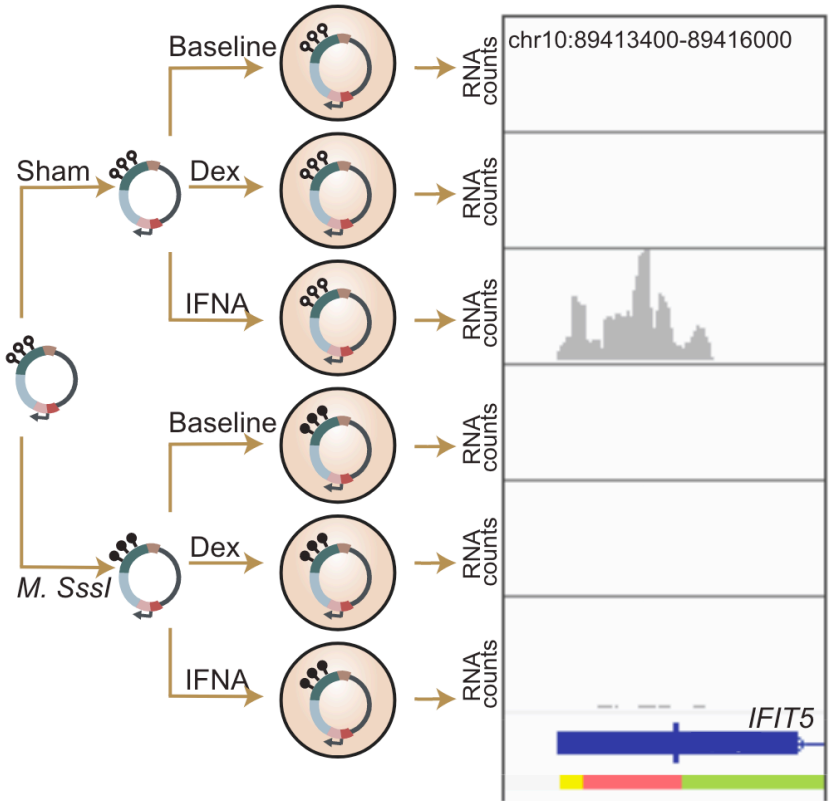


Figure 3-1. Diagram of the mSTARR-seq design adapted from Johnston et al. mSTARR-seq simultaneously tests for enhancer-like activity and DNA methylation-dependent enhancer activity. Sheared DNA from the GM12878 cell line was enriched for CpG loci and 100,000 randomly distributed control regions. Captured loci were cloned into the mSTARR-seq vector *pmSTARRseq1* and treated with either the CpG methylating enzyme *M. SssI* or a sham treatment, and then transfected into K562 cells. K562 cells were treated with Dex, IFNA or no challenge (baseline). An example of a DNA methylation-environment interaction is shown overlapping the

Figure 3-1, continued. interferon-induced gene *IFIT5* and an ENCODE-annotated weak promoter (pink denotes weak promoter, yellow denotes heterochromatin, and green denotes weak transcription. Three consecutive 600 bp windows have interaction FDR < 1 x 10^{-4} in this region. Panels depict non normalized, raw read pileups for mSTARR-seq RNA replicates, with all y-axis maximums set to 14000. No methylation-dependent activity is detectable in the baseline condition because this enhancer element is inactive. Upon IFNA stimulation, only unmethylated enhancer elements can respond.

Results

To investigate the relationship between DNA methylation and transcriptional response we drew on whole genome bisulfite sequencing (WGBS) and RNA-seq data from monocyte-derived macrophages from 35 individuals before and after flu-infection (Fig. 3-2A; previously described in *Chapter II*). Importantly, IFNA and flu infection can be compared since the IFNA pathway is directly upregulated upon flu infection, and thus the challenges illicit a similar phenotypic response[87]. We first investigated if DNA methylation levels at baseline predict transcriptional response to flu across mSTARR-seq windows with detectable enhancer activity in the IFNA condition (n=2769). Of these enhancers, 1033 are IFNA-specific and 1736 shared (detected in IFNA as well as in at least one other condition). We find that within each individual, the mean DNA methylation levels of all IFNA-detected enhancers significantly predict mean transcriptional response to flu (mean Pearson's r = -0.105 ± 0.006 s.d., all Bonferroni-corrected p < 3 x 10-5; Supplementary Table S3-1). However, this effect is largely driven by the subset of mSTARR-seq IFNA-specific enhancers rather than the shared enhancers as the correlation between baseline DNA methylation and transcriptional response is 3.44-fold stronger in IFNA-specific enhancers than for shared enhancers (Fig. 3-2B, IFNA-specific mean Pearson's r = -0.170 ± 0.009 s.d., all Bonferroni corrected p < 2 x 10-5; shared enhancers: mean Pearson's r = -0.049 ± 0.01 s.d., all Bonferroni-corrected p > 0.1). This led us to investigate the difference in methylation dependence between gene expression levels of non-infected and flu-infected cells as a stronger correlation in

IFNA regions may be driven by a stronger methylation dependence in flu-infected cells. Indeed, in flu-infected cells the average within-individual correlation between baseline DNA methylation and gene expression is 2.44 times larger after infection ($r = -0.261 \pm 0.006$) than at baseline ($r = -0.106 \pm 0.008$) in IFNA-specific mSTARR-seq enhancers (Fig. 3-2C, Supplementary Table S3-2). Together, these results suggest that in these regions pre-existing levels of methylation are an important factor in the transcriptional response to subsequent challenges.

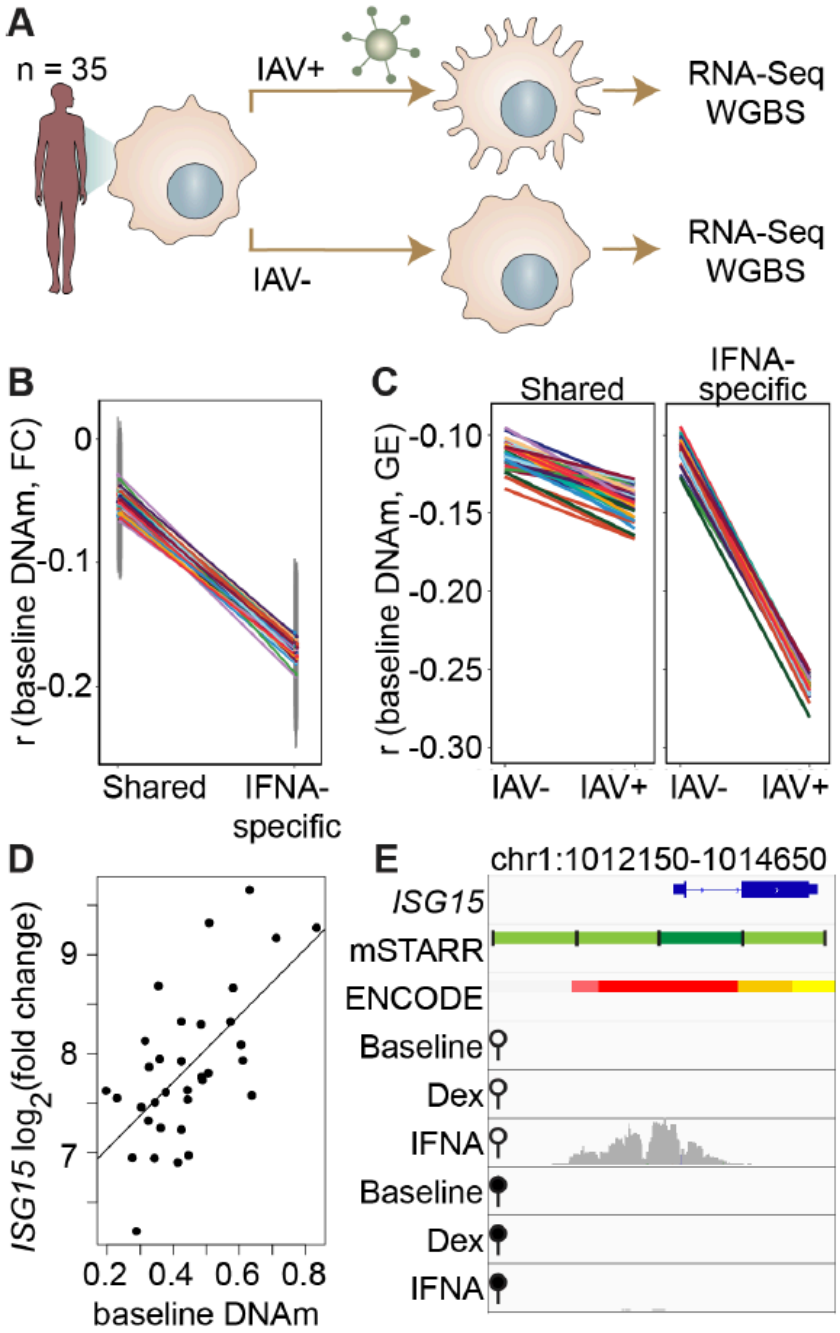


Figure 3-2. DNA methylation in mSTARR-seq enhancers predicts *in vivo* gene expression in macrophages. (A) Study design of the in *vivo* experiment, in which matched macrophage samples from 35 individuals were either left non-infected or infected with influenza A virus (IAV) for 24 hours and processed for RNA-Seq and whole genome bisulfite sequencing (WGBS; (Aracena *et al.*, 2022)). (B) Within individuals, DNA methylation (DNAm) levels at mSTARR-Seq enhancers in non-infected cells are negatively correlated with the nearest genes' transcriptional responses to IAV, but only in mSTARR-seq enhancers that were specific to the IFNA condition (IFNA-specific enhancers: n = 1033, mean Pearson's r = -0.170 ± 0.009 s.d., all Bonferroni-corrected $p < 2 \times 10^{-5}$; shared enhancers: n = 1736, mean Pearson's r = -0.049 ± 0.01 s.d., all Bonferroni-corrected $p > 0.1$). Each colored line represents an individual, and vertical gray lines represent 95% confidence

Figure 3-2, continued. intervals (see Supplementary Table S3-1 for full results). (C) The average within-individual correlation (r) between DNA methylation and gene expression (GE) is 2.44 times as large after infection ($r = -0.261 \pm 0.006$) than at baseline ($r = -0.106 \pm 0.008$) in IFNA-specific mSTARR-seq enhancers (right panel) but much less affected by infection at shared mSTARR-seq enhancers (left panel). Each colored line represents an individual (see Supplementary Table S3-2 for full results). (D) Across individuals, the *ISG15* transcriptional response to IAV is significantly correlated with average DNAm at the mSTARR-Seq enhancer chr1:1013400-1014000 in non-infected cells ($R^2 = 0.381$, $p = 6.05 \times 10^{-5}$, $q = 0.084$). Each dot represents an individual (see Supplementary Table S3-3) for full results and Figure S3-1 for condition-specific results). (E) The mSTARR-seq enhancer predictive of *ISG15* response to IAV (dark green bar) is located in the active promoter of *ISG15* (as defined by ENCODE (ENCODE, 2012); red denotes active promoter, pink denotes weak promoter, orange denotes strong enhancer, yellow denotes weak enhancer). Three adjacent, methylation-dependent, IFNA-specific mSTARR-seq enhancers were identified (light green), but do not significantly predict *ISG15* response to IAV ($q > 10\%$). The bottom 6 tracks depict non-normalized, raw read pileups for mSTARR-seq RNA replicates in either the unmethylated (open circle) or methylated (filled circle) condition, with all y-axis maximums set to 20,000.

We next sought to identify locus-specific examples in which inter-individual variation in DNA methylation levels drives differences in transcriptional response. Despite our relatively small samples size and the low DNA methylation variation across individuals (mean = 0.004, standard deviation = 0.0008), across 1382 testable mSTARR-seq enhancers (regions with at least 1 CpG with interindividual variance > 0.01), we identified one IFNA-specific, methylation dependent mSTARR-seq enhancer where variation in baseline DNA methylation levels across individuals predicts the transcriptional response to flu (Fig. 3-2D, 3-2E; $p = 6.05 \times 10^{-5}$, q-value = 0.0837; Supplementary Table S3-3). The mean methylation of the mSTARR-seq enhancer chr1:1013400-1014000, which overlaps the promoter of interferon-stimulated gene 15 (*ISG15*; Fig. 3-2E), explain 38% of the variance in *ISG15* transcriptional response to flu across individuals. The gene *ISG15* is involved in regulating the type I interferon response and modulating host immunity to viral and bacterial infections, thus previous evidence supports the biological relevance of regulating this locus[168]. Upon further exploration, we found that the direction of the effect on *ISG15* response is driven by the enhancer's effect on gene expression at baseline. Specifically, at

baseline, lower enhancer methylation is associated with higher *ISG15* expression (Fig. S3-1A), resulting in a shallower fold change response to flu (Fig. S3-1B, S3-1C). These data suggest that in a minority of regions, inter-individual variation in baseline DNA methylation levels significantly effects transcriptional response.

Discussion

The genome-wide application of the massively parallel reporter assay mSTARR-seq in combination with the human macrophage data indicates that while DNA methylation-environment interactions are widespread in the human genome, their functional role varies. While the vast majority of DNA methylation was not highly variable across individuals in our dataset, we identified one region (the *ISG15* locus) in which variation in DNA methylation explains a large proportion of variance of transcriptional regulation. We expect that given a larger sample size of individuals we would have greater power to detect smaller effects and uncover additional loci in which baseline DNA methylation across individuals predicts transcriptional response.

Despite the limitation of inter-individual variability, within-individual comparisons across loci indicate a clear correlation between baseline DNA methylation levels and transcriptional levels and response, particularly in specific environmental contexts, such as flu infection. Across thousands of genomic regions, regulatory activity and methylation-dependent effects on regulatory activity could only be detected in the flu-infected condition, highlighting the importance of assaying biologically relevant conditions[164,166]. Applying mSTARR-seq in additional cell types and to repeated challenges will be important to further understand the contexts in which DNA methylation impacts transcriptional responses.

Materials and Methods

Endogenous gene expression and methylation in human macrophages

To assess effects of DNA methylation-environment interactions on gene expression *in vivo*, we evaluated endogenous methylation and gene expression from matched whole genome bisulfite sequencing (WGBS) and RNA-seq data collected from human monocyte-derived macrophages (n = 35 donors), with and without infection with the influenza A virus (IAV)[68]. Unsmoothed methylation counts were obtained for 19,492,906 loci in both non-infected and IAV-infected samples (total n = 70) as described[68]. We filtered loci to require coverage of ≥ 4 sequence reads in at least half of the non-infected or IAV-infected samples. In the RNA-seq dataset for the same 35 individuals, we excluded any genes that did not have an average RPKM > 2 in non-infected or IAV-infected samples. This resulted in a total of 19,041,420 CpG sites and 14,122 genes used in downstream analyses.

We calculated normalization factors using calcNormFactors in edgeR (v 3.28.1)[137] to scale the raw library sizes. We then used the voom function in limma (v 3.42.2)[138,169] to apply the normalization factors, estimate the mean-variance relationship, and convert raw read counts to logCPM values. Sequencing batches were regressed out using ComBat from the sva Bioconductor package (v 3.34.0)[170], which fit a model that also included age (mean centered) and admixture. We subsequently regressed out age effects using limma. Individual-wise fold-change (FC) matrices were built by subtracting non-infected counts from IAV-infected counts for each individual using weights calculated as in (Harrison *et al.*, 2019; Aracena *et al.*, 2022)[68,147].

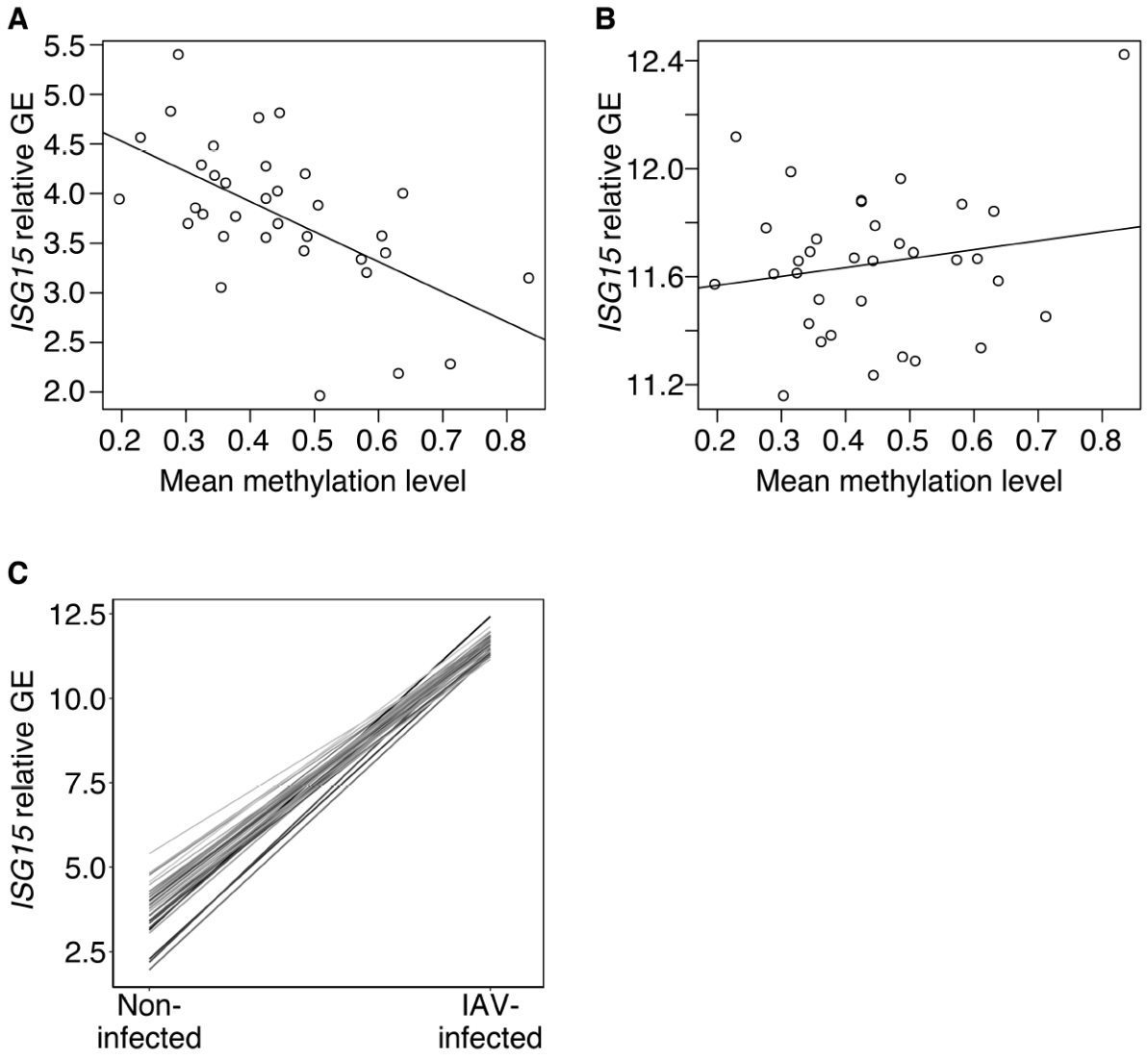
For comparison of the mSTARR-seq dataset to the dataset from Aracena *et al.,* 2022[68], GrCh38/hg38 coordinates were lifted over to GRCh37/hg19 using the UCSC liftOver tool. We

required a 0.95 minimum ratio of bases that remap, excluded loci that output multiple regions, set the minimum hit size in query to 0, and set the minimum chain size in the target to 0. The GRCh37/hg19 coordinates were used in the following analyses.

We first sought to assess, within each of the 35 individuals, the correlation between mSTARR-seq enhancer methylation levels at baseline (i.e., in non-infected cells) and transcriptional responses of their nearest genes to IAV. To find overlaps between enhancer regions and CpG loci, we used the *bedtools* intersect function (v2.29.1) with the "left outer join" option. For each 600 bp mSTARR-seq enhancer region in each individual, we calculated the mean methylation level across all overlapping CpGs. Each enhancer was linked to its nearest gene, with a maximum distance of 100 kb, as described above. If an enhancer had multiple linked genes (e.g., an enhancer that overlapped more than one gene), we took the mean transcriptional response of all linked genes. For each individual, we calculated the Pearson's correlation coefficient (r) between the DNA methylation levels within the mSTARR-seq-defined enhancer windows, in non-infected cells, and transcriptional responses of their linked genes to IAV infection. We also investigated the correlation between mean mSTARR-seq enhancer window methylation and gene expression *within* non-infected and flu-infected samples separately.

Finally, for each mSTARR-seq enhancer-gene pair, we sought to test the extent to which average methylation level in the enhancer in non-infected samples explained the gene's transcriptional response to IAV, across individuals. Here, we calculated the average CpG methylation level for each 600 bp enhancer, after excluding CpGs with methylation variance less than 0.01 across individuals. Thus, enhancers that did not contain any CpGs with appreciable interindividual variation in DNA methylation levels were excluded from the analysis. This filtering step resulted in 1,382 enhancer-gene pairs for this analysis. For each enhancer-gene pair, we

calculated the Pearson's correlation coefficient (and R^2) between the average methylation level of the enhancer and the transcriptional response of the linked gene. P-values were corrected for multiple hypothesis testing using the q-value method in R[97,171].



Supplementary Figure 3-1. Across individuals, methylation in the mSTARR-se annotated enhancer chr1:1013400-1014000 predicts the ISG15 gene expression response to flu. (A) Across individuals, average methylation within the mSTARR-seq annotated enhancer chr1:1013400-1014000 in non-infected baseline macrophages significantly predicts ISG15 gene expression (GE) in the non-infected condition (R2 = 0.324, p = 2.66 x 10-4), but (B) not in the IAV-infected condition (R2 = 0.001, p = 0.316). Each dot represents an individual. Relative GE is log(CPM) after regressing out the effects of sequencing batch and age. (C) Individuals with relatively low methylation in the mSTARR-seq chr1:1013400-1014000 enhancer (indicated by lighter line color) in the baseline, non-infected condition tend to have higher ISG15 gene expression in the non-infected condition, ultimately resulting in a shallower ISG15 transcriptional response to IAV infection (as indicated by slope of the line). Each line represents an individual.

Supplementary Tables

Supplementary tables for this chapter are described in Appendix: Supplementary Tables and available online.

Chapter IV: Discussion

In this thesis, I expand our understanding of how genetic variation impacts transcriptional and epigenetic inter-individual variation in response to influenza infection across individuals with varying proportions of African and European ancestry. I distinguish between genetically driven variation in molecular phenotypes and those that correlate with genetic ancestry but may be driven by other variables such as environmental factors. Furthermore, I emphasize the pervasiveness of gene x environment interactions and how epigenetic marks and gene expression act together to engage a coordinated response to infection. By highlighting the interactions between the same genetic variant and multiple molecular traits, I demonstrate the utility of generating multi-molecular trait data from the same individual. Lastly, I discuss the importance of considering molecular phenotypes beyond gene expression in disease-focused studies. While these contributions significantly advance the field, this work is not exhaustive and does not capture the full picture of how gene regulatory mechanisms affect disease risk. Exploring remaining questions appropriately contextualizes these advances and prioritizes future directions for follow-up.

In *Chapter II*, our findings support a driving role for differential transcription factor binding in many of the molecular QTL identified[44,79]. Future efforts to develop large scale datasets of TF-binding QTL can formally test this hypothesis, for example, using pooled ChIP-seq

methods.[115] To test if genetic variation is driving differential TF binding, the frequency of alleles can be compared pre- and post-ChIP to identify TF-binding QTL. If a genetic variant affects binding affinity of a TF, the measured frequency of the high-affinity allele will be enriched post-ChIP, relative to its pre-ChIP frequency. Overlaying these TF-binding QTL with molecular QTL identified herein will allow the formal testing of the hypothesis that differential TF binding at polymorphic sites primarily drives the high amount of epigenetic variation explained by genetic variation. Establishing the mechanism by which genetic variants influence molecular traits will get us closer to understanding the order of gene regulatory changes.

Though we identify genetic variants that effect multiple molecular traits, we have yet to discern the causal order of these regulatory changes. Namely, we are unable to determine the "first responder" in the gene regulatory cascade and which molecular traits follow. In *Chapter II* we find that epigenetic state at baseline strongly predicts transcriptional response of immune-related pathways, yet we cannot rule out that baseline epigenetic levels may still be reflective of baseline gene expression. Identifying if there is a direct causal path between the regulatory differences identified will improve insight into the gene regulatory cascade and subsequently allow more targeted and effective therapeutics. Nevertheless, we are limited in our ability to apply statistical methods developed to assess causal effects (e.g., mediation analysis) due to our limited sample size and the large number of molecular traits we wish to disentangle. Though mediation analysis can provide evidence in support of a causal relationship, it is not well suited to determine the direction of the relationship[172]. We hope the ongoing development of statistical methods will eventually allow the study of the causal relationships in our dataset and the biological mechanisms by which genetic variation mediates these in the context of immune response.

Thus far, the vast majority of immune response studies have only evaluated a single time point, essentially a snapshot of response. To gain a more complete picture of the response to infection over time, dense-time course experiments should be prioritized to evaluate inter-individual and inter-population speed and magnitude of response. With these datasets, dynamic QTL mapping can be performed to evaluate the genotype effect on response over time to identify loci associated with the magnitude of immune response or loci that only have an effect at early or late time points. These insights will be crucial to more closely replicate how gene by environmental interactions occur *in vivo* and increase the clinical translatability of findings.

It is well known that gene–environment (G × E) interactions are essential in determining risk and progression of disease, particularly infectious and immune diseases[16,28,33]. Thus, it is not surprising that inter-individual differences in response to pathogens are not only influenced by genetics, but also the environment. In *Chapter III*, we find that within-individual comparisons across loci show a clear relationship between baseline DNA methylation levels and transcriptional response, which is accentuated in the appropriate environmental context (IFNA stimulation). In *Chapter II*, we find that interferon and inflammatory responses are among the pathways most strongly associated with ancestry-associated differences. Interestingly, previous work has found these same pathways to be associated with social status[173].

At this time, the vast majority of immunogenetic studies are conducted with anonymized samples, eliminating our knowledge of an individual's environment, which include an individual's lifestyle, demographics and medical (such as infection and vaccination) history[174]. Evaluating environmental contributions to immune variation is necessary to frame the impact of potential clinical interventions derived from immunogenomics studies, as environment is closely tied to social determinants of health, such as access to healthcare and socioeconomic status, which impact

health outcomes. Social-environment triggered chronic inflammation may lead to a higher burden of infectious disease in individuals in poor social conditions[175]. In particular, low socioeconomic status and low social integration/support have been linked to higher prevalence of infectious disease, such as cytomegalovirus (CMV) and Epstein-Barr Virus (EBV).[175–178] Latent CMV fundamentally changes the adaptive and innate immune profiles of infected individuals, which in turn drives T cell immunosenescence, chronic inflammation, and increased vulnerability to other pathogens and some non-infectious diseases.[28,179] Despite this evidence, social environment is largely ignored as a contributing factor in immune gene regulation, as these data are often not available for individuals used in molecular studies.

Regrettably, we do not have the ability to formally test for social environment driven differences due to the use of anonymized samples in our study, and thus, it remains unknown if differences in social environment may contribute to the ancestry-associated differences we identified. We can only say with confidence that these differences are not driven by the top (most significant for each feature) *cis* variants identified in our study. Generating datasets that can directly test the relationship between genetics and social environment is challenging but crucial to determine the relative effect sizes of each on immune response. Interdisciplinary longitudinal studies would be the gold standard, but even collecting survey data on individuals' socioeconomic and health history would allow us to begin to better assess environment. In parallel, environment-adjacent data from the metabolome and virome or evaluation of antibodies with VirScan[180], can be collected to provide an additional quantitative approximation of an individuals' environmental exposures. Pairing these two approaches may allow the development of metrics to compare the relative accuracy of survey data and efficacy of different environmental determinant proxies. Performing experimental challenges on cell's derived from these individuals will allow us to

evaluate the interaction between social environment and genetics in immune response and advance our understanding of GxE in a disease context. To ensure these insights equitable advance insights for all, we must ensure these studies are representative of the global diversity of genetic variation and environments.

One of the most critical gaps in biomedical research today is the lack of representation of non-European ancestry individuals in genomics studies. We contribute to filling this gap by generating a large array of data on admixed individuals with African ancestry which are often missing from genomic studies. Yet, individuals from all over the world are required to truly understand the genetic diversity of human populations and equitably advance medical research. Incorporating individuals from varied genetic ancestral backgrounds will provide a more complete picture of the genetic basis of disease. Furthermore, while studying individuals from all over the world will increase environmental variation, it will likely decrease environmental confounders, as genetic ancestry may not necessarily correlate with the same set of environmental variables for individuals from different geographic locations. However, a study of global populations will also come with challenges. For example, populations have different allele frequencies and linkage disequilibrium patterns which can lead to false associations with disease risk. Additionally, it has also been debated if causal effects of alleles remain the same across populations, as previous cross-continental population comparisons have revealed discernable differences in the causal effects of alleles[181]. However, others argue these results are confounded by variation in environment or unstandardized categorizing of disease, suggesting that ancestry effects are virtually the same in admixed indviduals[182]. Future studies must carefully consider and test for potential sources of bias. Despite the complexity of studying and comparing global populations, an increase in diverse samples will promote the detection of population-specific variants associated with disease risk that

are rare in individuals of European descent. This work is essential to improve our ability to detect disease mechanisms in human populations and ensure that clinically transferrable findings are relevant to all people.

In human genetics research sample descriptors are used to describe genetic ancestry of the relatedness of people. Because they shape researchers' and the public's view of genetic relatedness, sample descriptors must accurately and clearly communicate how a person is related to others. At worst, inconsistent or overly simplified sample descriptors can cause confusion among researchers or even become weaponized by people seeking to racialize human diversity. Human geneticists must talk about their findings with accurate, clear, and meaningful descriptors of how a person is related to others. There is much debate in the field on how to describe genetic ancestry. Genetic similarity, or relatedness to a known set of samples, has been proposed to emphasize that individuals are only similar, not the same as a population or group of others. With this terminology, sample descriptors are not erroneously oversimplified as when using genetic ancestry groups, but the clarity of genetic similarity decreases as more admixture occurs[183]. Though the best terminology for communicating the genetic relatedness of people is still being debated, we must acknowledge that people's identities are shaped by many forces. To effectively collaborate with community members, researchers must approach people with cultural understanding, empathy, and communicate findings in an understandable way. Furthermore, there must be mutual understanding and engagement when collecting study samples, especially from historically marginalized and understudied populations, for which there has been historic mistrust of science.

If the goal of human genetics research is to increase understanding of genetic and trait variation and disease mechanisms, we must ensure future studies capture the full spectrum of

genetic diversity that exists across the globe. Furthermore, we must continue to investigate how genetic variation interacts with various environmental factors that closely recapitulate *in vivo* disease and gene regulatory mechanisms. Ultimately, understanding the relative contribution of genetics and environmental factors will help prioritize policy to ensure improved health equity and outcomes for all.

References

1. The Cost of Sequencing a Human Genome (2022). Genome.gov. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

2. Edwards, A.O., Ritter, R., Abel, K.J., Manning, A., Panhuysen, C., and Farrer, L.A. (2005). Complement factor H polymorphism and age-related macular degeneration. Science *308*, 421–424. 10.1126/science.1110189.

3. Mozzi, A., Pontremoli, C., and Sironi, M. (2018). Genetic susceptibility to infectious diseases: Current status and future perspectives from genome-wide approaches. Infect. Genet. Evol. *66*, 286–307. 10.1016/j.meegid.2017.09.028.

4. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22. 10.1016/J.AJHG.2017.06.005.

5. Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. *51*, 1339–1348. 10.1038/s41588-019-0481-0.

6. Parkes, M., Cortes, A., Van Heel, D.A., and Brown, M.A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat. Rev. Genet. *14*, 661–673. 10.1038/nrg3502.

7. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res. *22*, 1748–1759. 10.1101/gr.136127.111.

8. Barbeira, A.N., Bonazzola, R., Gamazon, E.R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., et al. (2021). Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Genome Biol. *22*, 49. 10.1186/s13059-020-02252-4.

9. Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., et al. (2021). Genome-wide enhancer maps link risk variants to disease genes. Nature *593*, 238–243. 10.1038/s41586-021-03446-x.

10. Morris, J.A., Daniloski, Z., Domingo, J., Barry, T., Ziosi, M., Glinos, D.A., Hao, S., Mimitou, E.P., Smibert, P., Roeder, K., et al. (2021). Discovery of target genes and pathways of blood trait loci using pooled CRISPR screens and single cell RNA sequencing. 2021.04.07.438882. 10.1101/2021.04.07.438882.

11. GWAS Diversity Monitor https://gwasdiversitymonitor.com/.

12. Soil-transmitted helminth infections https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections.

13. Douglas, G.M., Bielawski, J.P., and Langille, M.G.I. (2020). Re-evaluating the relationship between missing heritability and the microbiome. Microbiome *8*, 87. 10.1186/s40168-020-00839-4.

14. Augustine, T., Al-Aghbar, M.A., Al-Kowari, M., Espino-Guarch, M., and van Panhuys, N. (2022). Asthma and the Missing Heritability Problem: Necessity for Multiomics Approaches in Determining Accurate Risk Profiles. Front. Immunol. *13*, 822324. 10.3389/fimmu.2022.822324.

15. Knudsen, T.M., Rezwan, F.I., Jiang, Y., Karmaus, W., Svanes, C., and Holloway, J.W. (2018). Transgenerational and intergenerational epigenetic inheritance in allergic diseases. J. Allergy Clin. Immunol. *142*, 765–772. 10.1016/j.jaci.2018.07.007.

16. Hunter, D.J. (2005). Gene–environment interactions in human diseases. Nat. Rev. Genet. *6*, 287–298. 10.1038/nrg1578.

17. Trerotola, M., Relli, V., Simeone, P., and Alberti, S. (2015). Epigenetic inheritance and the missing heritability. Hum. Genomics *9*, 17. 10.1186/s40246-015-0041-3.

18. Karlsson, E.K., Kwiatkowski, D.P., and Sabeti, P.C. (2014). Natural selection and infectious disease in human populations. Nat. Rev. Genet. *15*, 379–393. 10.1038/nrg3734.

19. Barreiro, L.B., and Quintana-Murci, L. (2010). From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat. Rev. Genet. *11*, 17–30. 10.1038/nrg2698.

20. Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., and Nielsen, R. (2011). Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. PLoS Genet. *7*, e1002355. 10.1371/journal.pgen.1002355.

21. Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z.A., Pacis, A., Dumaine, A., Grenier, J.-C., Freiman, A., Sams, A.J., Hebert, S., et al. (2016). Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell *167*, 657-669.e21. 10.1016/J.CELL.2016.09.025.

22. Brinkworth, J.F., and Barreiro, L.B. (2014). The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. 10.1016/j.coi.2014.09.008.

23. Pennington, R., Gatenbee, C., Kennedy, B., Harpending, H., and Cochran, G. (2009). Group differences in proneness to inflammation. Infect. Genet. Evol. *Evol. 9*, 1371–1380. 10.1016/j.meegid.2009.09.017.

24. Richardus, J.H., and Kunst, A.E. (2001). Black-white differences in infectious disease mortality in the United States.

25. Oliva, M., Muñoz-Aguirre, M., Kim-Hellmuth, S., Wucher, V., Gewirtz, A.D.H., Cotter, D.J., Parsana, P., Kasela, S., Balliu, B., Viñuela, A., et al. (2020). The impact of sex on gene expression across human tissues. Science *369*, eaba3066. 10.1126/science.aba3066.

26. Whitacre, C.C. (2001). Sex differences in autoimmune disease. Nat. Immunol. *2*, 777–780. 10.1038/ni0901-777.

27. Bakker, O.B., Aguirre-Gamboa, R., Sanna, S., Oosting, M., Smeekens, S.P., Jaeger, M., Zorro, M., Võsa, U., Withoff, S., Netea-Maier, R.T., et al. (2018). Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. Nat. Immunol. *19*, 776–786. 10.1038/S41590-018-0121-3.

28. Brodin, P., and Davis, M.M. (2017). Human immune system variation. Nat. Rev. Immunol. *17*, 21–29. 10.1038/nri.2016.125.

29. Piasecka, B., Duffy, D., Urrutia, A., Quach, H., Patin, E., Posseme, C., Bergstedt, J., Charbit, B., Rouilly, V., MacPherson, C.R., et al. (2018). Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. Proc. Natl. Acad. Sci. U. S. A. *115*, E488–E497. 10.1073/PNAS.1714765115/-/DCSUPPLEMENTAL.

30. Schirmer, M., Smeekens, S.P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E.A., Jansen, T., Jacobs, L., Bonder, M.J., Kurilshikov, A., et al. (2016). Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. Cell *167*, 1125-1136.e8. 10.1016/J.CELL.2016.10.020.

31. Snyder-Mackler, N., Sanz, J., Kohn, J.N., Brinkworth, J.F., Morrow, S., Shaver, A.O., Grenier, J.C., Pique-Regi, R., Johnson, Z.P., Wilson, M.E., et al. (2016). Social status alters immune regulation and response to infection in macaques. Science *354*, 1041–1045. 10.1126/science.aah3580.

32. Snyder-Mackler, N., Burger, J.R., Gaydosh, L., Belsky, D.W., Noppert, G.A., Campos, F.A., Bartolomucci, A., Yang, Y.C., Aiello, A.E., O'Rand, A., et al. (2020). Social determinants of health and survival in humans and other animals. Science *368*. 10.1126/SCIENCE.AAX9553.

33. Cole, S.W. (2014). Human social genomics. PLoS Genet. *10*. 10.1371/JOURNAL.PGEN.1004601.

34. Quintana-Murci, L. (2019). Human Immunology through the Lens of Evolutionary Genetics. Cell *177*, 184–199. 10.1016/J.CELL.2019.02.033.

35. Quach, H., Rotival, M., Pothlichet, J., Loh, Y.-H.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic Adaptation and Neandertal

Admixture Shaped the Immune System of Human Populations. Cell *167*, 643-656.e17. 10.1016/J.CELL.2016.09.024.

36. Aguet, F., Alasoo, K., Li, Y.I., Battle, A., Im, H.K., Montgomery, S.B., and Lappalainen, T. (2023). Molecular quantitative trait loci. Nat. Rev. Methods Primer *3*, 4. 10.1038/s43586-022-00188-6.

37. Çalışkan, M., Baker, S.W., Gilad, Y., and Ober, C. (2015). Host Genetic Variation Influences Gene Expression Response to Rhinovirus Infection. PLoS Genet. *11*. 10.1371/journal.pgen.1005111.

38. Ye, C.J., Feng, T., Kwon, H.-K., Raj, T., Wilson, M.T., Asinovski, N., McCabe, C., Lee, M.H., Frohlich, I., Paik, H. -i., et al. (2014). Intersection of population variation and autoimmunity genetics in human T cell activation. Science *345*, 1254665–1254665. 10.1126/science.1254665.

39. Sanz, J., Randolph, H.E., and Barreiro, L.B. (2018). Genetic and evolutionary determinants of human population variation in immune responses. Curr. Opin. Genet. Dev. *53*, 28–35. 10.1016/J.GDE.2018.06.009.

40. Randolph, H.E., Fiege, J.K., Thielen, B.K., Mickelson, C.K., Shiratori, M., Barroso-Batista, J., Langlois, R.A., and Barreiro, L.B. (2021). Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. Science *374*, 1127–1133. 10.1126/SCIENCE.ABG0928/SUPPL_FILE/SCIENCE.ABG0928_MDAR_REPRODUCIBILITY_CHECKLIST.PDF.

41. Pacis, A., Mailhot-léonard, F., Tailleux, L., Randolph, H.E., and Yotova, V. (2019). Gene activation precedes DNA demethylation in response to infection in human dendritic cells. 10.1073/pnas.1814700116.

42. Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S.B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., et al. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife *2013*. 10.7554/eLife.00523.

43. Pai, A.A., Pritchard, J.K., and Gilad, Y. (2015). The Genetic and Mechanistic Basis for Variation in Gene Regulation. PLoS Genet. *11*, e1004857. 10.1371/journal.pgen.1004857.

44. McVicker, G., Van De Geijn, B., Degner, J.F., Cain, C.E., Banovich, N.E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J.K. (2013). Identification of genetic variants that affect histone modifications in human cells. Science *342*, 747–749. 10.1126/science.1242429.

45. Li, Y.I., Van De Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. Science *352*, 600–604. 10.1126/science.aad9417.

46. Rozek, L.S., Dolinoy, D.C., Sartor, M.A., and Omenn, G.S. (2014). Epigenetics: Relevance and Implications for Public Health. Annu. Rev. Public Health *35*, 105–122. 10.1146/annurev-publhealth-032013-182513.

47. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–329. 10.1038/nature14248.

48. Allis, C.D., and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. Nat. Rev. Genet. *17*, 487–500. 10.1038/nrg.2016.59.

49. Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318. 10.1038/ng1966.

50. Jin, B., Li, Y., and Robertson, K.D. (2011). DNA Methylation. Genes Cancer *2*, 607–617. 10.1177/1947601910393957.

51. Sun, S., and Barreiro, L.B. (2020). The epigenetically-encoded memory of the innate immune system. Curr. Opin. Immunol. *65*, 7–13. 10.1016/j.coi.2020.02.002.

52. Bekkering, S., Dominguez-Andres, J., Joosten, L.A.B., Riksen, N.P., and Netea, M.G. (2021). Trained Immunity: Reprogramming Innate Immunity in Health and Disease. Annu. Rev. Immunol. *39*, 667–693. 10.1146/ANNUREV-IMMUNOL-102119-073855.

53. Zhang, Q., and Cao, X. (2021). Epigenetic Remodeling in Innate Immunity and Inflammation. Annu. Rev. Immunol. *39*, 279–311. 10.1146/ANNUREV-IMMUNOL-093019-123619.

54. Quintin, J., Saeed, S., Martens, J.H.A., Giamarellos-Bourboulis, E.J., Ifrim, D.C., Logie, C., Jacobs, L., Jansen, T., Kullberg, B.-J., Wijmenga, C., et al. (2012). Candida albicans infection affords protection against reinfection via functional reprogramming of monocytes. Cell Host Microbe *12*, 223–232. 10.1016/j.chom.2012.06.006.

55. Cheng, S.-C., Quintin, J., Cramer, R.A., Shepardson, K.M., Saeed, S., Kumar, V., Giamarellos-Bourboulis, E.J., Martens, J.H.A., Rao, N.A., Aghajanirefah, A., et al. (2014). mTOR- and HIF-1α-mediated aerobic glycolysis as metabolic basis for trained immunity. Science *345*, 1250684. 10.1126/science.1250684.

56. Netea, M.G., Joosten, L.A.B., Latz, E., Mills, K.H.G., Natoli, G., Stunnenberg, H.G., O'Neill, L.A.J., and Xavier, R.J. (2016). Trained immunity: a program of innate immune memory in health and disease. Science *352*, aaf1098. 10.1126/science.aaf1098.

57. Pacis, A., Tailleux, L., Morin, A.M., Lambourne, J., MacIsaac, J.L., Yotova, V., Dumaine, A., Danckært, A., Luca, F., Grenier, J.C., et al. (2015). Bacterial infection remodels the

DNA methylation landscape of human dendritic cells. Genome Res. *25*, 1801–1811. 10.1101/gr.192005.115.

58. Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S., and Natoli, G. (2013). Latent enhancers activated by stimulation in differentiated cells. Cell *152*, 157–171. 10.1016/j.cell.2012.12.018.

59. Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell *167*, 1398-1414.e24. 10.1016/j.cell.2016.10.026.

60. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nat. 2012 4827385 *482*, 390–394. 10.1038/nature10808.

61. Husquin, L.T., Rotival, M., Fagny, M., Quach, H., Zidane, N., McEwen, L.M., MacIsaac, J.L., Kobor, M.S., Aschard, H., Patin, E., et al. (2018). Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation. Genome Biol. *19*. 10.1186/S13059-018-1601-3.

62. Carja, O., MacIsaac, J.L., Mah, S.M., Henn, B.M., Kobor, M.S., Feldman, M.W., and Fraser, H.B. (2017). Worldwide patterns of human epigenetic variation. Nat. Ecol. Evol. *1*, 1577–1583. 10.1038/S41559-017-0299-Z.

63. Waszak, S.M., Delaneau, O., Gschwind, A.R., Kilpinen, H., Raghav, S.K., Witwicki, R.M., Orioli, A., Wiederkehr, M., Panousis, N.I., Yurovsky, A., et al. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell *162*, 1039–1050. 10.1016/J.CELL.2015.08.001.

64. Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A.J., Mann, A.L., Kundu, K., Hale, C., Dougan, G., and Gaffney, D.J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat. Genet. *50*, 424–431. 10.1038/S41588-018-0046-7.

65. GTEx Consortium (2017). Genetic effects on gene expression across human tissues GTEx consortium*. 10.1038/nature24277.

66. Chun, S., Casparino, A., Patsopoulos, N.A., Croteau-Chonka, D.C., Raby, B.A., De Jager, P.L., Sunyaev, S.R., and Cotsapas, C. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat. Genet. *49*, 600–605. 10.1038/ng.3795.

67. Mu, Z., Wei, W., Fair, B., Miao, J., Zhu, P., and Li, Y.I. (2021). The impact of cell type and context-dependent regulatory variants on human immune traits. Genome Biol. *22*, 122. 10.1186/s13059-021-02334-x.

68. Aracena, K.A., Lin, Y.-L., Luo, K., Pacis, A., Gona, S., Mu, Z., Yotova, V., Sindeaux, R., Pramatarova, A., Simon, M.-M., et al. (2022). Epigenetic variation impacts ancestry-associated differences in the transcriptional response to influenza infection. bioRxiv, 2022.05.10.491413. 10.1101/2022.05.10.491413.

69. Duffy, D., Rouilly, V., Libri, V., Hasan, M., Beitz, B., David, M., Urrutia, A., Bisiaux, A., LaBrie, S.T., Dubois, A., et al. (2014). Functional analysis via standardized whole-blood stimulation systems defines the boundaries of a healthy immune response to complex stimuli. Immunity *40*, 436–450. 10.1016/J.IMMUNI.2014.03.002.

70. Barreiro, L.B., Tailleux, L., Pai, A.A., Gicquel, B., Marioni, J.C., and Gilad, Y. (2012). Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc. Natl. Acad. Sci. U. S. A. *109*, 1204–1209. 10.1073/PNAS.1115761109/-/DCSUPPLEMENTAL/PNAS.201115761SI.PDF.

71. Aguirre-Gamboa, R., Joosten, I., Urbano, P.C.M., van der Molen, R.G., van Rijssen, E., van Cranenbroek, B., Oosting, M., Smeekens, S., Jaeger, M., Zorro, M., et al. (2016). Differential Effects of Environmental and Genetic Factors on T and B Cell Immune Traits. Cell Rep. *17*, 2474–2487. 10.1016/J.CELREP.2016.10.053.

72. Li, Y., Oosting, M., Smeekens, S.P., Jaeger, M., Aguirre-Gamboa, R., Le, K.T.T., Deelen, P., Ricaño-Ponce, I., Schoffelen, T., Jansen, A.F.M., et al. (2016). A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. Cell *167*, 1099-1110.e14. 10.1016/J.CELL.2016.10.017.

73. Bierne, H., Hamon, M., and Cossart, P. (2012). Epigenetics and bacterial infections. Cold Spring Harb. Perspect. Med. *2*. 10.1101/CSHPERSPECT.A010272.

74. Monticelli, S., and Natoli, G. (2013). Short-term memory of danger signals and environmental stimuli in immune cells. Nat. Immunol. *14*, 777–784. 10.1038/NI.2636.

75. Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., Gregory, L., Lonie, L., Chew, A., Wei, C.L., et al. (2010). Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. Immunity *32*, 317–328. 10.1016/J.IMMUNI.2010.02.008.

76. Qiao, Y., Giannopoulou, E.G., Chan, C.H., Park, S. ho, Gong, S., Chen, J., Hu, X., Elemento, O., and Ivashkiv, L.B. (2013). Synergistic activation of inflammatory cytokine genes by interferon-γ-induced chromatin remodeling and toll-like receptor signaling. Immunity *39*, 454–469. 10.1016/J.IMMUNI.2013.08.009.

77. Villagra, A., Cheng, F., Wang, H.W., Suarez, I., Glozak, M., Maurin, M., Nguyen, D., Wright, K.L., Atadja, P.W., Bhalla, K., et al. (2009). The histone deacetylase HDAC11 regulates the expression of interleukin 10 and immune tolerance. Nat. Immunol. *10*, 92–100. 10.1038/NI.1673.

78. Carja, O., MacIsaac, J.L., Mah, S.M., Henn, B.M., Kobor, M.S., Feldman, M.W., and Fraser, H.B. (2017). Worldwide patterns of human epigenetic variation. Nat. Ecol. Evol. *1*, 1577–1583. 10.1038/S41559-017-0299-Z.

79. Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive variation in chromatin states across humans. Science *342*, 750–752. 10.1126/SCIENCE.1242510.

80. The ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799. 10.1038/NATURE05874.

81. The ENCODE Project Consortium, Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., et al. (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature *489*, 57. 10.1038/NATURE11247.

82. Fernández, J.M., de la Torre, V., Richardson, D., Royo, R., Puiggròs, M., Moncunill, V., Fragkogianni, S., Clarke, L., Flicek, P., Rico, D., et al. (2016). The BLUEPRINT Data Analysis Portal. Cell Syst. *3*, 491. 10.1016/J.CELS.2016.10.021.

83. Meischel, T., Villalon-Letelier, F., Saunders, P.M., Reading, P.C., and Londrigan, S.L. (2020). Influenza A virus interactions with macrophages: Lessons from epithelial cells. Cell. Microbiol. *22*, e13170. 10.1111/CMI.13170.

84. Ichinohe, T., Pang, I.K., and Iwasaki, A. (2010). Influenza virus activates inflammasomes via its intracellular M2 ion channel. Nat. Immunol. 2010 115 *11*, 404–410. 10.1038/ni.1861.

85. Diebold, S.S., Kaisho, T., Hemmi, H., Akira, S., and Reis E Sousa, C. (2004). Innate Antiviral Responses by Means of TLR7-Mediated Recognition of Single-Stranded RNA. Science *303*, 1529–1531. 10.1126/SCIENCE.1093616/SUPPL_FILE/DIEBOLD.SOM.PDF.

86. Ernst, J., and Kellis, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216. 10.1038/nmeth.1906.

87. Killip, M.J., Fodor, E., and Randall, R.E. (2015). Influenza virus activation of the interferon system. Virus Res. *209*, 11. 10.1016/J.VIRUSRES.2015.02.003.

88. Ciancanelli, M.J., Abel, L., Zhang, S.-Y., and Casanova, J.-L. (2016). Host genetics of severe influenza: from mouse Mx1 to human IRF7. Curr. Opin. Immunol. *38*, 109. 10.1016/J.COI.2015.12.002.

89. Froggatt, H.M., Harding, A.T., Chaparian, R.R., and Heaton, N.S. (2021). ETV7 limits antiviral gene expression and control of influenza viruses. Sci. Signal. *14*, 1194. 10.1126/SCISIGNAL.ABE1194/SUPPL_FILE/SCISIGNAL.ABE1194_SM.PDF.

90. Pezzè, L., Meškytė, E.M., Forcato, M., Pontalti, S., Badowska, K.A., Rizzotto, D., Skvortsova, I.I., Bicciato, S., and Ciribilli, Y. (2021). ETV7 regulates breast cancer stem-like cell features by repressing IFN-response genes. Cell Death Dis. *12*. 10.1038/S41419-021-04005-Y.

91. Pham, D., Moseley, C.E., Gao, M., Savic, D., Winstead, C.J., Sun, M., Kee, B.L., Myers, R.M., Weaver, C.T., and Hatton, R.D. (2019). Batf Pioneers the Reorganization of Chromatin in Developing Effector T Cells via Ets1-Dependent Recruitment of Ctcf. Cell Rep. *29*, 1203-1220.e7. 10.1016/j.celrep.2019.09.064.

92. Wu, X., Kasmani, M.Y., Zheng, S., Khatun, A., Chen, Y., Winkler, W., Zander, R., Burns, R., Taparowsky, E.J., Sun, J., et al. (2022). BATF promotes group 2 innate lymphoid cell-mediated lung tissue protection during acute respiratory virus infection. Sci. Immunol. *7*. 10.1126/SCIIMMUNOL.ABC9934.

93. Lee, W., Kingstad-Bakke, B., Kedl, R.M., Kawaoka, Y., and Suresh, M. (2021). CCR2 Regulates Vaccine-Induced Mucosal T-Cell Memory to Influenza A Virus. J. Virol. *95*. 10.1128/JVI.00530-21.

94. Scott-Browne, J.P., López-Moyado, I.F., Trifari, S., Wong, V., Chavez, L., Rao, A., and Pereira, R.M. (2016). Dynamic changes in chromatin accessibility in CD8+ T cells responding to viral infection. Immunity *45*, 1327. 10.1016/J.IMMUNI.2016.10.028.

95. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664. 10.1101/gr.094052.109.

96. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044. 10.1126/science.1172257.

97. Storey, J., Bass, A., Dabney, A., and Robinson, D. (2019). qvalue: Q-value estimation for false discovery rate control. R Package Version 2180.

98. Urbut, S.M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. *51*, 187–195. 10.1038/s41588-018-0268-8.

99. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. *1*, 417–425. 10.1016/j.cels.2015.12.004.

100. Banovich, N.E., Lan, X., McVicker, G., van de Geijn, B., Degner, J.F., Blischak, J.D., Roux, J., Pritchard, J.K., and Gilad, Y. (2014). Methylation QTLs Are Associated with

Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. PLoS Genet. *10*. 10.1371/journal.pgen.1004663.

101. Huan, T., Joehanes, R., Song, C., Peng, F., Guo, Y., Mendelson, M., Yao, C., Liu, C., Ma, J., Richard, M., et al. (2019). Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. Nat. Commun. 2019 101 *10*, 1–14. 10.1038/s41467-019-12228-z.

102. Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. *5*, 435–445. 10.1038/NRG1348.

103. Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. Httpdxdoiorg101146annurev-Genet-072610-155046 *44*, 445–477. 10.1146/ANNUREV-GENET-072610-155046.

104. Lee, M.N., Ye, C., Villani, A.C., Raj, T., Li, W., Eisenhaure, T.M., Imboywa, S.H., Chipendo, P.I., Ran, F.A., Slowikowski, K., et al. (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science *343*. 10.1126/SCIENCE.1246980.

105. Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., et al. (2014). Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science *343*, 1246949. 10.1126/SCIENCE.1246949.

106. Zhang, Q., and Cao, X. (2019). Epigenetic regulation of the innate immune response to infection. Nat. Rev. Immunol. *19*, 417–432. 10.1038/s41577-019-0151-6.

107. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. *10*. 10.1371/journal.pgen.1004383.

108. Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genet. *16*, e1008720. 10.1371/journal.pgen.1008720.

109. Lovering, R.C., Camon, E.B., Blake, J.A., and Diehl, A.D. (2008). Access to immunology through the Gene Ontology. Immunology *125*, 154. 10.1111/J.1365-2567.2008.02940.X.

110. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228. 10.1038/NG.3404.

111. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., Gusev, A., et al. (2017). Linkage disequilibrium–

dependent architecture of human complex traits shows action of negative selection. Nat. Genet. 2017 4910 *49*, 1421–1427. 10.1038/ng.3954.

112. Bulik-Sullivan, B., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M., Corvin, A., et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. 2015 473 *47*, 291–295. 10.1038/ng.3211.

113. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. 2015 479 *47*, 1091–1098. 10.1038/ng.3367.

114. Ding, Z., Ni, Y., Timmer, S.W., Lee, B.K., Battenhouse, A., Louzada, S., Yang, F., Dunham, I., Crawford, G.E., Lieb, J.D., et al. (2014). Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. PLoS Genet. *10*. 10.1371/JOURNAL.PGEN.1004798.

115. Tehranchi, A.K., Myrthil, M., Martin, T., Hie, B.L., Golan, D., and Fraser, H.B. (2016). Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. Cell *165*, 730–741. 10.1016/J.CELL.2016.03.041.

116. Fanucchi, S., Domínguez-Andrés, J., Joosten, L.A.B., Netea, M.G., and Mhlanga, M.M. (2021). The Intersection of Epigenetics and Metabolism in Trained Immunity. Immunity *54*, 32–43. 10.1016/J.IMMUNI.2020.10.011.

117. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. *106*, 9362–9367. 10.1073/PNAS.0903103106.

118. GTEx Consortium (2017). Genetic effects on gene expression across human tissues GTEx consortium*. 10.1038/nature24277.

119. Võsa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310. 10.1038/S41588-021-00913-Z.

120. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. 10.1126/SCIENCE.AAZ1776/SUPPL_FILE/AAZ1776_TABLESS10-S16.XLSX.

121. Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods *14*, 959–962. 10.1038/nmeth.4396.

122. Jiang, H., Lei, R., Ding, S.W., and Zhu, S. (2014). Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics *15*, 182. 10.1186/1471-2105-15-182.

123. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. 10.1093/bioinformatics/btp324.

124. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. 10.1101/gr.107524.110.

125. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, S.K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res *29*, 308–311.

126. Deelen, P., Bonder, M.J., Van Der Velde, K.J., Westra, H.J., Winder, E., Hendriksen, D., Franke, L., and Swertz, M.A. (2014). Genotype harmonizer: Automatic strand alignment and format conversion for genotype data integration. BMC Res. Notes *7*, 1–4. 10.1186/1756-0500-7-901.

127. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288. 10.1016/j.ajhg.2013.06.020.

128. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. 10.1093/bioinformatics/btu170.

129. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. 10.1093/bioinformatics/bts635.

130. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169. 10.1093/bioinformatics/btu638.

131. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. 10.1186/gb-2008-9-9-r137.

132. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. 10.1093/bioinformatics/btt656.

133. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.

134. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. 2019 91 *9*, 1–5. 10.1038/s41598-019-45839-z.

135. Ernst, J., and Kellis, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216. 10.1038/nmeth.1906.

136. Krueger, F., and Andrews, S.R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572. 10.1093/bioinformatics/btr167.

137. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. 10.1093/bioinformatics/btp616.

138. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. 10.1093/nar/gkv007.

139. Park, Y., and Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics *32*, 1446–1453. 10.1093/bioinformatics/btw026.

140. Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. J. Stat. Softw. *17*, 1–27. 10.18637/jss.v017.i01.

141. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2021). Fast gene set enrichment analysis. bioRxiv, 60012. 10.1101/060012.

142. Yu, G., Wang, L.G., and He, Q.Y. (2015). ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics *31*, 2382–2383. 10.1093/bioinformatics/btv145.

143. Li, Z., Schulz, M.H., Look, T., Begemann, M., Zenke, M., and Costa, I.G. (2019). Identification of transcription factor binding sites using ATAC-seq. Genome Biol. *20*, 45. 10.1186/s13059-019-1642-2.

144. Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: An open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. *32*. 10.1093/nar/gkh012.

145. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, R24. 10.1186/gb-2007-8-2-r24.

146. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300. 10.2307/2346101.

147. Harrison, G.F., Sanz, J., Boulais, J., Mina, M.J., Grenier, J.C., Leng, Y., Dumaine, A., Yotova, V., Bergey, C.M., Nsobya, S.L., et al. (2019). Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. Nat. Ecol. Evol. *3*, 1253–1264. 10.1038/s41559-019-0947-6.

148. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinforma. 2013 141 *14*, 1–15. 10.1186/1471-2105-14-7.

149. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22. 10.18637/JSS.V033.I01.

150. Dong, X., Li, X., Chang, T.-W., Scherzer, C.R., Weiss, S.T., and Qiu, W. (2021). powerEQTL: an R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis. Bioinformatics *37*, 4269–4271. 10.1093/bioinformatics/btab385.

151. Shabalin, A.A. (2012). Gene expression Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *28*, 1353–1358. 10.1093/bioinformatics/bts163.

152. McClay, J.L., Shabalin, A.A., Dozmorov, M.G., Adkins, D.E., Kumar, G., Nerella, S., Clark, S.L., Bergen, S.E., Consortium, S.S., Hultman, C.M., et al. (2015). High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. Genome Biol. *16*. 10.1186/S13059-015-0842-7.

153. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915. 10.1038/s41587-019-0201-4.

154. Kalita, C.A., Moyerbrailean, G.A., Brown, C., Wen, X., Luca, F., and Pique-Regi, R. (2018). QuASAR-MPRA: accurate allele-specific analysis for massively parallel reporter assays. Bioinforma. Oxf. Engl. *34*, 787–794. 10.1093/bioinformatics/btx598.

155. Harvey, C.T., Moyerbrailean, G.A., Davis, G.O., Wen, X., Luca, F., and Pique-Regi, R. (2015). QuASAR: quantitative allele-specific analysis of reads. Bioinformatics *31*, 1235–1242. 10.1093/bioinformatics/btu802.

156. Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., and Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. Nat. Methods *14*, 590–592. 10.1038/nmeth.4267.

157. Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448. 10.1038/ng.3679.

158. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, 1000529. 10.1371/journal.pgen.1000529.

159. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Ser. B Stat. Methodol. *82*, 1273–1300. 10.1111/RSSB.12388.

160. Wen, X. (2016). Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. Httpsdoiorg10121416-AOAS952 *10*, 1619–1638. 10.1214/16-AOAS952.

161. Dolinoy, D.C., Weidman, J.R., Waterland, R.A., and Jirtle, R.L. (2006). Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. Environ. Health Perspect. *114*, 567–572. 10.1289/ehp.8700.

162. Kazachenka, A., Bertozzi, T.M., Sjoberg-Herrera, M.K., Walker, N., Gardner, J., Gunning, R., Pahita, E., Adams, S., Adams, D., and Ferguson-Smith, A.C. (2018). Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. Cell *175*, 1717. 10.1016/j.cell.2018.11.017.

163. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*. 10.1126/SCIENCE.AAJ2239.

164. Kamada, R., Yang, W., Zhang, Y., Patel, M.C., Yang, Y., Ouda, R., Dey, A., Wakabayashi, Y., Sakaguchi, K., Fujita, T., et al. (2018). Interferon stimulation creates chromatin marks and establishes transcriptional memory. Proc. Natl. Acad. Sci. U. S. A. *115*, E9162–E9171. 10.1073/pnas.1720930115.

165. Johnston, R.A., Aracena, K.A., Barreiro, L.B., Lea, A.J., and Tung, J. (2023). DNA methylation-environment interactions in the human genome. 2023.05.19.541437. 10.1101/2023.05.19.541437.

166. Sun, S., and Barreiro, L.B. (2020). The epigenetically-encoded memory of the innate immune system. Curr. Opin. Immunol. *65*, 7–13. 10.1016/j.coi.2020.02.002.

167. Provençal, N., Arloth, J., Cattaneo, A., Anacker, C., Cattane, N., Wiechmann, T., Röh, S., Ködel, M., Klengel, T., Czamara, D., et al. (2020). Glucocorticoid exposure during hippocampal neurogenesis primes future stress response by inducing changes in DNA methylation. Proc. Natl. Acad. Sci. U. S. A. *117*, 23280–23285. 10.1073/pnas.1820842116.

168. Perng, Y.-C., and Lenschow, D.J. (2018). ISG15 in antiviral immunity and beyond. Nat. Rev. Microbiol. *16*, 423–439. 10.1038/s41579-018-0020-5.

169. Smyth, G.K. (2005). limma: Linear Models for Microarray Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor Statistics for Biology and

Health., R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, eds. (Springer-Verlag), pp. 397–420. 10.1007/0-387-29362-0_23.

170. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883. 10.1093/bioinformatics/bts034.

171. Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A. *100*, 9440–9445. 10.1073/pnas.1530509100.

172. Pierce, B.L., Tong, L., Argos, M., Demanelis, K., Jasmine, F., Rakibuz-Zaman, M., Sarwar, G., Islam, M.T., Shahriar, H., Islam, T., et al. (2018). Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. Nat. Commun. *9*, 804. 10.1038/s41467-018-03209-9.

173. Simons, N.D., and Tung, J. (2019). Social Status and Gene Regulation: Conservation and Context Dependence in Primates. Trends Cogn. Sci. *23*, 722–725. 10.1016/j.tics.2019.06.003.

174. Cuomo, A.S.E., Nathan, A., Raychaudhuri, S., MacArthur, D.G., and Powell, J.E. (2023). Single-cell genomics meets human genetics. Nat. Rev. Genet., 1–15. 10.1038/s41576-023-00599-5.

175. Young Chung, H., Cesari, M., Anton, S., Marzetti, E., Giovannini, S., Young Seo, A., Carter, C., Pal Yu, B., and Leeuwenburgh, C. (2009). Molecular Inflammation: Underpinnings of Aging and Age-related Diseases. Ageing Res Rev *8*, 18–30. 10.1016/j.arr.2008.07.002.

176. Lantos, P.M., Hoffman, K., Permar, S.R., Jackson, P., Hughes, B.L., Kind, A., and Swamy, G. Neighborhood disadvantage is associated with high cytomegalovirus seroprevalence in pregnancy. 10.1007/s40615-017-0423-4.

177. Aiello, A.E., Dowd, J.B., Jayabalasingham, B., Feinstein, L., Uddin, M., Simanek, A.M., Cheng, C.K., Galea, S., Wildman, D.E., Koenen, K., et al. (2016). PTSD is associated with an increase in aged T cell phenotypes in adults living in Detroit HHS Public Access. Psychoneuroendocrinology *67*, 133–141. 10.1016/j.psyneuen.2016.01.024.

178. Gares, V., Panico, L., Castagne, R., Delpierre, C., and Kelly-Irving, M. (2019). The role of the early social environment on Epstein Barr virus infection: a prospective observational design using the Millennium Cohort Study On behalf of the Lifepath consortium. 10.1017/S0950268817002515.

179. Picarda, G., and Benedict, C.A. (2018). Cytomegalovirus: Shape-Shifting the Immune System. J. Immunol. *200*, 3881–3889. 10.4049/jimmunol.1800171.

180. Shrock, E.L., Shrock, C.L., and Elledge, S.J. (2022). VirScan: High-throughput Profiling of Antiviral Antibody Epitopes. Bio-Protoc. *12*, e4464. 10.21769/BioProtoc.4464.

181. Patel, R.A., Musharoff, S.A., Spence, J.P., Pimentel, H., Tcheandjieu, C., Mostafavi, H., Sinnott-Armstrong, N., Clarke, S.L., Smith, C.J., V.A. Million Veteran Program, et al. (2022). Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. Am. J. Hum. Genet. *109*, 1286–1297. 10.1016/j.ajhg.2022.05.014.

182. Hou, K., Ding, Y., Xu, Z., Wu, Y., Bhattacharya, A., Mester, R., Belbin, G.M., Buyske, S., Conti, D.V., Darst, B.F., et al. (2023). Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. Nat. Genet. *55*, 549–558. 10.1038/s41588-023-01338-6.

183. Coop, G. (2022). Genetic similarity versus genetic ancestry groups as sample descriptors in human genetics.

Appendix: Supplementary Tables

Supplementary Tables for Chapter II

Table S2-1. Description of the samples and libraries generated for *Chapter II.* Available as an excel file online. Description of samples, including demographic information about donors, technical batch/variable information, and experimental information.

Table S2-2. List of differentially expressed, accessible, and methylated features in response to flu infection. Available as an excel file online. IAV infection differential expression effect for each feature for each molecular trait is reported, including effect size estimate, p-value, q-value, and t-statistic.

Note: Only CpG sites with FDR<.50 in at least one condition are reported due to file size limitations. Full methylation analysis results available upon request.

Table S2-3. Transcription Factor activity scores and TF enrichment results in condition specific QTL. Available as an excel file online.TF activity scores including mean activity and p-value and TF enrichments for QTL with p-value.

Table S2-4. List of population differentially expressed and responsive features. Available as an excel file online. Population differential expression and response effect for each feature for each molecular trait is reported, including effect size estimate, p-value, q-value, and local false sign rate (lfsr).

Note: Only CpG sites with FDR<.50 in at least one condition are reported due to file size limitations. Full methylation analysis results available upon request.

Table S2-5. List of *cis* regulatory QTLs identified in non-infected and flu-infected macrophages using both SNPs and STRs. Available as an excel file online. QTL effect for the lead *cis*-SNP per feature is reported across molecular traits, including effect size estimate, p-value, q-value, and local false sign rate (lfsr).

Note: Only CpG sites with FDR<.10 in at least one condition are reported due to file size limitations. Full methylation analysis results available upon request.

Table S2-6. QTL integration results. Available as an excel file online. Frequency and percentage of all QTL overlap patterns.

Table S2-7. The best SNP-eRNA and corresponding molecular QTL. Available as an excel file online. For each eRNA, the feature and corresponding p-value for each overlapping QTL.

Table S2-8. Colocalization results for 14 immune related GWAS. Available as an excel file online. Colocalization results for each of the molecular QTL across immune-related diseases.

Table S2-9. LDSC-computed heritability results for 14 immune related GWAS. Available as an excel file online. Enrichments, percent heritability explained and meta enrichments for each of the molecular QTL.

Table S2-10. Predixcan results for 14 immune related GWAS. Available as an excel file online. Significant molecular features, with p-value and corresponding gene id across immune-related diseases.

Supplementary Tables for Chapter III

Table S3-1. Pearson's correlation results, within individuals, between mSTARR-seq enhancer DNA methylation in non-infected macrophages and transcriptional response of nearest genes upon flu infection. Available as an excel file online. Pearson's correlation, including confidence interval, p-value, and Bonferroni-corrected p-value.

Table S3-2. Pearson's correlation results, within individuals, between mSTARR-seq enhancer DNA methylation in non-infected macrophages and gene expression level in non-infected or IAV-infected macrophages. Available as an excel file online. Pearson's correlation, including confidence interval, p-value, and Bonferroni-corrected p-value.

Table S3-3. Table of Pearson's correlation results, across individuals, between mSTARR-seq enhancer DNA methylation in non-infected macrophages and transcriptional response of nearest genes upon IAV infection. Available as an excel file online. Adjusted R2, p-value and q-value for enhancers.