THE UNIVERSITY OF CHICAGO

COMMON GENETIC VARIATION INFLUENCES THE IMMUNE RESPONSE TO VIRAL INFECTIONS IN HUMAN POPULATIONS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY

HALEY ELIZABETH RANDOLPH

CHICAGO, ILLINOIS

JUNE 2023

This dissertation is dedicated to my mother, Elizabeth Randolph.

“Even Louis Vuitton makes mistakes.”

– Luann de Lesseps, The Real Housewives of New York City

Table of Contents

Tables S3-1 to 10 are available as supplementary files online. The List of Tables gives the page number for each table's caption.

List of Figures

List of Tables

Acknowledgements

It takes a village. First and foremost, I would like to thank my thesis advisor, Dr. Luis Barreiro. You have pushed me to become a better scientist, and I have learned so much more in your lab than I ever thought possible. Thank you for trusting me, giving me space to grow, and always being patient. I have never doubted that you have my back and that you genuinely care about those who join your lab, which is rare to find in an advisor, and I feel incredibly grateful to have experienced that. Not only do I look up to you immensely as a scientist for your creativity, rigor, and intellect, I also, after almost seven years of working with you, consider you to be a good friend. For me, the relationships that I have built in your lab have been the most invaluable part of my PhD. I know that we will be lifelong collaborators, and I really look forward to that.

Throughout my education, I have been immensely lucky to have had extremely supportive, passionate, and outstanding teachers. Specifically, I would like to thank my 9th and 11th grade biology teachers, John Politano and Brenda Frost, for fostering my curiosity in science well before I considered it my career, my 12th grade calculus teacher, Mrs. Zack, for proving to me that being good at math is cool and giving me a space to explore that idea, and my 10th grade English teacher, Brian Kocur, for teaching me to always unabashedly let your personality shine. Thank you to Dr. Stephen Schaeffer, my undergraduate advisor, for teaching me the fundamentals of being a scientist, instilling in me the belief that curiosity is the most important quality of a researcher, reminding me that your gut instinct is usually right, wholeheartedly supporting my decisions even though they were not necessarily the conventional ones, and continuing to support me throughout my graduate career. Thank you to Dr. Michele Ardolino for always offering sound advice with a sense of humor, being patient with me when I barely knew how to do anything in the lab (I am no

longer a “grandma” in the cell culture room!), and firmly cementing in my mind that I will never work with lab mice ever again.

To the current and past members of the Barreiro lab from the Université de Montréal and the University of Chicago, thank you for always providing an entertaining, welcoming, and hilarious work environment. Doctorates are difficult, and it is much easier to thrive when you genuinely enjoy the people around you. In particular, I would like to thank Dr. Joaquín Sanz for teaching me how to code and never getting frustrated with me even though I was frustrated with myself, Dr. Genelle Harrison for always letting me vent and be silly, Dr. Vania Yotova for collectively being our “lab mom”, Dr. João Barroso-Batista for also listening to punk music and attending many Riot Fests with me, and Dr. Raúl Aguirre-Gamboa for sharing advice on making beautiful figures and tequila. Finally, I would like to thank the “Only Child Club” rebranded as the “No Boys Club,” including Mari Shiratori, Katie Aracena, Veronica Locher, Dr. Sarah Sun, Bridget Chak, Cary Brandolino, and Ellen Ketter. All of you are badass, brilliant women – I am so grateful to have gotten to know you, and it has and will continue to be a pleasure to giggle with you.

Thank you to the many collaborators I have worked with along the way who have made my life considerably easier, including those in Dr. Ryan Langlois’s lab at the University of Minnesota, particularly Dr. Jessica Fiege and Dr. Beth Thielen, Dr. Daniel Kaufmann’s lab at the Centre Hospitalier de l'Université de Montréal, particularly Dr. Elsa Brunet-Ratnasingham, and Dr. Brent Richards’s lab at the Jewish General Hospital, particularly Dr. Tomoko Nakanishi. Thank you to my thesis committee members, Drs. Yoav Gilad, Patrick Wilson, and Xin He, for always engaging in valuable discussions and providing outside perspectives of my work. Thank you to Sue Levison for being the reason the Genetics, Genomics, and Systems Biology and Human

Genetics programs operate as well-oiled machines and for simultaneously being an incredibly kind and caring person.

To my ride or dies, Édgar Correa, Dr. Briana Mittleman, Rebecca Butler, Mari Shiratori, Andrew Stier, Dr. Genelle Harrison, and Samantha Lynch, thank you for being my chosen family. Thank you for the many nights at The Pub, brunches, charcuterie boards, vacations, theme parties, inside jokes, nights out in Boystown, and nights in watching Scooby Doo and RuPaul's Drag Race. Thank you for sharing your lives with me; I could not have asked for a better, funnier, smarter, more ridiculous, and more empathetic group of people to call my best friends. À la Sam: "There are good ships and there are wood ships, there are ships that sail the sea, but the best ships are friendships, and may they always be." Love you all and can't wait to see what we accomplish.

To Dr. Jeffrey Downey, I am so grateful to have met you. I never thought helping a collaborator analyze some RNA-seq data would turn into something that is much more than that but here we are. Thank you for always letting me lean on you, being unconditionally supportive and kind to me, and making me laugh uncontrollably and feel at peace. I could not be more excited to see what the future holds for us. Love you, forever and for always.

To my mom, Elizabeth Randolph, and late grandma, Helen Randolph, thank you for always supporting my dreams and allowing me to pursue my goals without question. From an early age, both of you taught me that you should never be afraid to speak your mind and break the mold, and I try to embody that every day. Mom, I am so proud of you, I love you so much, and I hope I continue to make you proud.

Abstract

Humans show remarkable differences in susceptibility to many infectious diseases, and in part, this heterogeneity arises from variation in the immune response following infection. The immune response to infection is a complex, dynamic process that involves the coordinated action of multiple cell types to recognize and eliminate a pathogen. Genome-wide association studies and expression quantitative trait loci (eQTL) mapping studies in immune cells have shown that certain polymorphisms drive variation in the response to viruses in specific contexts. Yet, little is known about how genetic ancestry and genetic variation affect the immune response to viral infection more broadly. We generated single-cell RNA-sequencing data in multiple disease contexts, including *in vitro* infection with influenza A virus and *in vivo* infection with SARS-CoV-2, and mapped eQTL to study the genetic architecture of gene expression in these various settings. Following influenza infection, we showed that genetic ancestry effects on gene expression were common, highly cell type-specific, and often explained by *cis*-eQTL. Among hospitalized COVID-19 patients, we demonstrated that a substantial proportion of all *cis*-eQTL identified showed a significant gene-environment interaction effect: either they were observed only in monocytes of SARS-CoV-2-infected individuals or were associated with functional cell state. Together, our findings establish common *cis*-regulatory variants as key determinants of the response to viral infection, highlight the ubiquitous nature of gene-environment interactions in the framework of an immune response, and underscore the need to study regulatory processes in relevant cell types and disease states.

Chapter I: Introduction

Human variation in immune responses and patterns of genetic diversity

Between-individual and between-population immune response variation is common and presumably contributes to varying susceptibility to infectious and autoimmune diseases across individuals (Duffy et al. 2014; Pennington et al. 2009; Brinkworth and Barreiro 2014). Both genetic and nongenetic factors (Piasecka et al. 2018; Brodin et al. 2015) shape this heterogeneity and defining the relative contribution of these components to immune response diversity is a principal goal of human immunogenomics research. Through the study of human genomes, the genetic underpinnings of complex immune-related diseases can be linked with molecular traits and clinically relevant variables to better define the genetic architecture of these phenotypes. Ultimately, a greater understanding of the features that give rise to immune response heterogeneity and immunological disorders may accelerate the field of personalized medicine, which promises to tailor medical decisions and treatment options based on genetic information (Ashley 2016).

Environmental factors

Environmental factors, such as age, sex, microbiome, previous exposure to pathogens, etc., are responsible for immune response variation across individuals to a large extent. Age-related effects on the immune system have been well characterized, and it is known that immune function

declines as a consequence of aging. Specifically, elderly individuals produce fewer B and T cells in primary lymphoid organs and harbor immune cells with reduced functional capacity, leading to overall weaker immune responses compared to younger individuals (Montecino-Rodriguez, Berent-Maoz, and Dorshkind 2013). Likewise, immunological differences associated with sex have been widely described. In general, adult females mount stronger immune responses compared to adult males, resulting in more rapid pathogen clearance, greater vaccine effectiveness, and increased susceptibility to autoimmune and inflammatory diseases in females (Klein and Flanagan 2016). Age and sex have also been shown to directly impact the transcriptional response of immune genes in a broad but cell type-specific manner, with $CD8^+$ T cells mediating age effects and $CD4^+$ T cells and monocytes mediating sex effects on expression (Piasecka et al. 2018).

In addition, an individual's prior infection history partly determines their subsequent immune responses to previously encountered and novel pathogens owing to adaptive immune memory, heterologous immunity, and trained immunity (Cooper and Alder 2006; Iwasaki and Medzhitov 2015; Netea et al. 2020). While these mechanisms of immune memory directly influence the immune response itself, pathogens may also exert effects that impact responses indirectly through other means, such as altering cell type composition. In particular, latent cytomegalovirus (CMV) infection has been shown to remodel the lymphoid compartment, accounting for up to 73% of population differences in lymphocyte cell type proportions between healthy donors originating from Central Africa and Western Europe (Aquino et al. 2022).

Other environmental factors closely linked with societal inequalities rather than biological traits, such as variation in access to healthcare and socioeconomic status, also contribute to immune response heterogeneity across individuals and populations. For example, in the United

States, racial and ethnic minorities are at much higher risk of significant morbidity and mortality due to influenza A infection and COVID-19 disease compared to non-Hispanic white Americans (J. Y. Ko et al. 2021; Chandrasekhar et al. 2017). Given the disproportionate access to healthcare and other health disparities in the United States, much of this imbalance can be attributed to health inequities caused by structural and social determinants. Although these biases likely lead to measurable differences between individuals that are not genetically controlled, it is difficult to tease apart their relative contributions to variation in immune response phenotypes because other nongenetic and genetic factors are often confounded.

Genetic factors

While a considerable amount of heterogeneity in the response to infection can be ascribed to environmental factors, a large proportion is also due to genetic variation at loci throughout the genome (Bakker et al. 2018; Piasecka et al. 2018; Nédélec et al. 2016; Quach et al. 2016). Pathogens constitute one of the most powerful sources of selection in human evolutionary history (Karlsson, Kwiatkowski, and Sabeti 2014). It is hypothesized that the geographic distribution of pathogens varied significantly between human populations as modern humans migrated out of Africa (Stephens Patrick R. et al. 2016). This variation in the magnitude and diversity of pathogen exposure across populations likely drove allele frequencies to diverge at loci impacting the host immune response via natural selection. Such past human evolution is expected to be reflected among individuals living today, and the study of modern human genomes coupled with functional immunological assays allows us to explicitly test this hypothesis.

Positive selection, a form of natural selection in which advantageous genetic variants sweep to high frequency in a population, has substantially influenced the evolution of the human

genome (Booker, Jackson, and Keightley 2017; Pardis C. Sabeti et al. 2007). Genes that play a central role in innate immunity and immune defense pathways exhibit clear signatures of positive selection in present-day human populations (Nielsen et al. 2005; Barreiro and Quintana-Murci 2010). Several studies have shown that regions targeted by positive selection are enriched for genes known to be involved in susceptibility to infectious diseases (Karlsson, Kwiatkowski, and Sabeti 2014; Fumagalli and Sironi 2014), indicating that genetic factors play a role in shaping the response to pathogens. RNA viruses, such as lentiviruses and orthomyxoviruses, have imposed some of the strongest evolutionary pressures on the human genome, with genomic footprints of these viruses present in modern genomes today (Enard and Petrov 2018; 2020). Specifically, introgressed segments of the human genome derived from ancient hominid populations, such as Neanderthals, are enriched for proteins known to interact with viruses, suggesting that these regions represent adaptively introgressed segments that likely conferred a selective advantage when introduced into modern human populations (Enard and Petrov 2018; 2020).

Finally, studies of gene-environment interactions that aim to define how genetic and environmental factors jointly affect response outcome or disease risk are becoming more common as cohort sample sizes rise (Virolainen et al. 2022). The importance of these non-additive effects in modifying the immune response to infectious diseases cannot be discounted, although they are generally less well-characterized in the context of human health at present due to the difficulty of identifying these loci.

Mapping the gene regulatory basis of immune response variation

Genome-wide association studies, or GWAS, allow us to identify regions of the genome that are associated with complex diseases and traits. These studies assess whether any genetic variants are overrepresented among individuals with a particular disease compared to healthy, control individuals. While GWAS have proven successful in identifying risk alleles for many complex diseases, including autoimmune, cardiovascular, metabolic, and neurodegenerative disorders, relatively few infectious disease GWAS have been performed in comparison (Mozzi, Pontremoli, and Sironi 2018). GWAS for various viral (human immunodeficiency virus, influenza A virus, hepatitis B/C virus, etc.) and bacterial (*Staphylococcus aureus*, *Mycobacterium tuberculosis*, etc.) infections exist, although these have attained only modest success, with few variants reaching genome-wide significance and minimal shared signals across studies considering the same pathogen (Mozzi, Pontremoli, and Sironi 2018). Of note, among the significant trait-associated loci identified by complex disease GWAS, the vast majority are located in non-protein coding regions of the genome (Edwards et al. 2013), pointing towards gene regulatory variation as a crucial factor in modulating disease risk.

In parallel with association studies, expression quantitative trait loci, or eQTL, studies can be used to map gene expression phenotypes to particular genomic loci by combining measures of gene expression with genome-wide genotyping data (Cheung et al. 2005; Gibson and Weir 2005; Morley et al. 2004). eQTL have been identified in an extensive variety of cell types and environmental contexts, and their study has significantly shaped our understanding of gene regulation and genetic architecture of gene expression (Pickrell et al. 2010; Lappalainen et al. 2013; Aguet et al. 2017). More specifically, eQTL mapping has proven to be uniquely powered to identify genetic factors that explain between-individual variation in the immune response to

pathogens (Barreiro et al. 2012; M. N. Lee et al. 2014; Nédélec et al. 2016; Manry et al. 2017; Kim-Hellmuth et al. 2017). Previous immune response eQTL studies have primarily focused on evaluating *cis*-regulatory genetic factors, or allele-specific genetic effects that act locally, in isolated immune cell types (M. N. Lee et al. 2014; Nédélec et al. 2016; Quach et al. 2016). This design inherently overlooks potential cell type-specific genetic effects that might influence immune response variation. Further, these studies have largely ignored the dynamic nature of the immune response, sampling gene expression at only a single time point following pathogen exposure or during the course of disease. To better understand the genetic basis of the immune response, we must characterize eQTL in various immune cell types and disease contexts, keeping temporal resolution and fluctuating functional cell states in mind.

Experimental human immune response models

Historically, mice have been utilized as a model to study human disease and health, especially considering the immune response to infection. While we have gleaned a considerable amount of mechanistic insight about the immune system from these studies, mice do not always faithfully reproduce human biology (Mestas and Hughes 2004). It is well-documented that considerable immunological differences between the two species exist, including variation in the composition of lymphocytes and granulocytes (Doeing, Borowicz, and Crockett 2003), differences in the response to exogenous immune challenges, such as lipopolysaccharide (LPS) (Copeland et al. 2005), and discrepancies in the function or expression of numerous immune cell receptors, signaling molecules, and cell type subsets (Mestas and Hughes 2004). Further, considering genetics, humans offer a distinct feature that inbred, isogenic laboratory mice do not: a standing

source of natural genetic variation. Because of these notable differences, mouse models are limited in their ability to advance the knowledge of basic human immunology. Ultimately, the consideration of human genetic diversity is necessary to understand immune response variation.

To characterize the human immune response more precisely, human models of disease and immunity have been adopted. Both *in vitro* and *ex vivo* models are commonly used, each with their own set of advantages and drawbacks. *In vitro* models involve the collection, isolation, and culture of primary immune cells or tissues from healthy or diseased donors or the culture of immortalized human cell lines (M. N. Lee et al. 2014; Quach et al. 2016; Nédélec et al. 2016). In culture, experimental manipulations can be performed to assess certain targeted questions. While these models preserve phenotypic differences between individuals, they mitigate the effects of environmental confounders, as cells from different donors can be cultured in identical laboratory environments. *Ex vivo* models involve the collection and characterization of fresh or frozen primary samples without cell culture or experimental manipulation from healthy donors or patients, which adds the element of environmental complexity but is arguably more relevant to disease phenotypes (Schulte-Schrepping et al. 2020). These human models serve to complement mouse models, as exhaustive experimental and mechanistic studies cannot be performed in the *in vivo* human setting, and both *in vitro* and *ex vivo* human models are often limited by the amount of tissue or blood that can be obtained from a subject at a given time.

RNA-sequencing technologies

Next-generation massively parallel sequencing technologies have facilitated the high-throughput profiling of whole genomes and transcriptomes at unprecedented scales.

Conventionally, gene expression phenotypes have been measured using bulk RNA-sequencing techniques, which provide an average gene expression measurement for all cells present in a sample (Trapnell 2015). Because information from an entire sample is aggregated into a single measurement, bulk RNA-sequencing will blur information if the underlying sample is heterogeneous, like peripheral blood mononuclear cells (PBMCs), which are comprised of multiple, distinct immune cell types at varying proportions across individuals. In particular, signals from rare immune cell types will be obscured by more abundant cell types, diminishing the nuances of the data.

Single-cell RNA-sequencing can dissect gene expression patterns at a finer resolution than bulk RNA-sequencing. While bulk RNA-sequencing relies on computing averages of expression measurements, single-cell RNA-sequencing outputs a gene expression measurement for each individual cell present in a sample (Macosko et al. 2015; Trapnell 2015). While it has been shown that lead *cis*-genetic effects are often shared across cell types and tissues (Aguet et al. 2017), a fraction of *cis*-eQTL exert cell type-specific effects, such that they are only present in a single cell type or their effect sizes are much larger in certain cell types (Urbut et al. 2019). Combining single-cell RNA-sequencing data with whole genome-sequencing data, we can map *cis*-eQTL within each cell type using pseudobulk data, or data that has been aggregated within a biological replicate, and then ask how *cis*-regulatory patterns change with cellular identity. More recently, methods have been described to map *cis*-eQTL using the true single-cell data itself, avoiding the loss of information that occurs with aggregated pseudobulk data (Nathan et al. 2022; Cuomo et al. 2022). These methods enable the mapping of *cis*-eQTL that fluctuate with fluid cell states within cell types or that correlate with disease-relevant variables, such as time. Single-cell RNA-sequencing

allows us to more precisely define the gene regulatory landscape within and between cell types while also considering how this may interact with a cell's functional state.

Summary

In this thesis, I investigate the genetic and evolutionary factors that give rise to human variation in the immune response to infection. First, I provide an in-depth summary of what is known about the evolutionary forces that influence between-individual and between-population variation in the immune response and the role that genetic variation plays in shaping these response differences. Then, I use single-cell RNA-sequencing to characterize heterogeneity in the response to influenza A across various immune cell types and tease apart how population variation and *cis*-genetic effects influence immune gene regulation. Finally, I measure the contribution of gene-environment interactions to differences in the response to SARS-CoV-2 infection among hospitalized COVID-19 patients. Together, this work highlights the utility of studies investigating gene regulatory processes in various cell type, cell state, and disease contexts, with the eventual goal of linking the path between disease-associated loci and mechanisms of disease susceptibility.

Chapter II: Genetic and evolutionary determinants of human population variation in immune responses

Note:

The following section (*Chapter II*) is reproduced verbatim, with the exception of figure numbering and reference labeling, from my co-first authored reference "Genetic and evolutionary determinants of human population variation in immune responses" (Sanz*, Randolph*, and Barreiro, 2018). This project was published in *Current Opinion in Genetics & Development* on December 1, 2018.[1]

Authors:

J. Sanz*, H.E. Randolph*, and L.B. Barreiro

*These authors contributed equally to this work.

[1] Sanz, J.*, Randolph, H.E.*, & Barreiro, L.B. (2018). Genetic and evolutionary determinants of human population variation in immune responses. Current Opinion in Genetics & Development, 53, 28–35.

Abstract

Humans display remarkable immune response variation when exposed to identical immune challenges. However, our understanding of the genetic, evolutionary, and environmental factors that impact this inter-individual and inter-population immune response heterogeneity is still in its early days. In this review, we discuss three fundamental questions concerning the recent evolution of the human immune system: the degree to which individuals from different populations vary in their innate immune responses, the genetic variants accounting for such differences, and the evolutionary mechanisms that led to the establishment of these variants in modern human populations. We also discuss how past selective events might have contributed to the uneven distribution of immune-related disorders across populations.

Full Text

Pathogens are one of the strongest selective pressures on the human genome (Fumagalli et al. 2011; Barreiro and Quintana-Murci 2010; Siddle and Quintana-Murci 2014). As modern humans migrated out of Africa ~60 kyr ago and traversed new territories, they encountered markedly different pathogenic environments, likely resulting in population-specific selection of immune phenotypes (Figure 2-1A) (Nielsen et al. 2017; Quach and Quintana-Murci 2017; Stephens Patrick R. et al. 2016). Consistent with this hypothesis, some of the most compelling evidence for local positive selection in the human genome has been detected among genes involved in immunity and host defense (Barreiro and Quintana-Murci 2010; Karlsson, Kwiatkowski, and Sabeti 2014; Quintana-Murci and Clark 2013). Yet, our understanding of the role that local

adaptation plays in shaping phenotypic variation in immune responses across populations is still in its infancy.

The innate immune system is the earliest immune defense mechanism activated upon pathogen invasion. Pathogen-induced signaling through innate immune receptors prompts pervasive changes in gene expression that subsequently trigger the activation of inflammatory and/or antiviral immune effectors involved in pathogen clearance (Smale 2010). Inter-individual differences in innate immune responses are common and presumably contribute to varying susceptibility to infection, inflammation, and autoimmune disorders (Figure 2-1B) (Pennington et al. 2009; Brinkworth and Barreiro 2014; Duffy et al. 2014). Although a substantial fraction of transcriptional heterogeneity in response to infection is likely attributable to environmental factors, a large portion is also due to host genetics. Recently, the contribution of host genetics to innate immune response diversity among individuals has been demonstrated using expression quantitative trait loci (eQTL) mapping (Fairfax and Knight 2014) on diverse subsets of immune cells both at baseline and after exposure to immune stimuli and live pathogens (Figure 2-2A,B) (Fairfax and Knight 2014; Fairfax et al. 2014; Kim-Hellmuth et al. 2017; Quach et al. 2016; Alasoo et al. 2018; Nédélec et al. 2016; Barreiro et al. 2012; M. N. Lee et al. 2014; Ferraro et al. 2014; Ye et al. 2014; Çalışkan et al. 2015; Piasecka et al. 2018). These "immune response eQTL" studies have identified a large number of host genetic variants that underlie differential innate immune responses to infection, some of which have been associated with increased susceptibility to sepsis, inflammatory bowel disease, viral hepatitis, typhoid fever, and tuberculosis (Barreiro et al. 2012; Wang et al. 2018; Alvarez et al. 2017; D. C. Ko et al. 2012). Recently, several studies have implemented similar eQTL mapping approaches to determine the extent to which disparities in

immune phenotypes between populations are due to genetically-controlled transcriptional response variation (Figure 2-2C) (Quach et al. 2016; Nédélec et al. 2016).

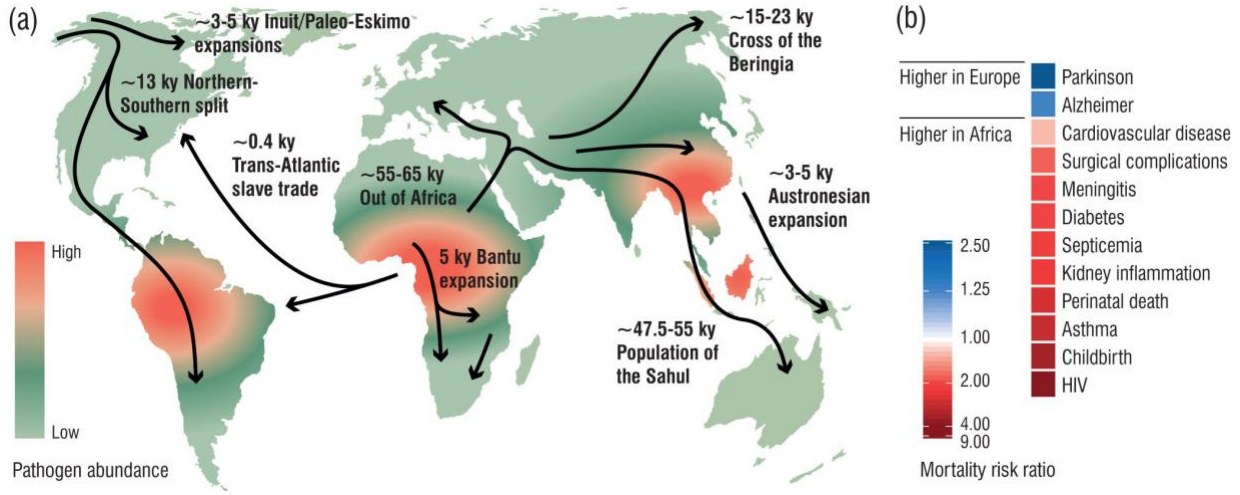


Fig. 2-1. Historical variation in pathogen exposure across different human populations. (A) Major migratory routes followed by the ancestors of present-day humans who originated in Africa around 200 kyr, adapted from Refs. (Nielsen et al. 2017; Quach and Quintana-Murci 2017). The spread of humans throughout the world led to the settlement of populations in geographical areas with variable environments. Ultimately, these migrations imposed novel and heterogeneous adaptive pressures on the human genome. The existence of varying levels of pathogen diversity (geographic distribution sketched from results summarized in Ref. (Stephens Patrick R. et al. 2016)) is linked with many significant signatures of adaptation found in relevant immune system genes (Barreiro and Quintana-Murci 2010; Karlsson, Kwiatkowski, and Sabeti 2014; Quintana-Murci and Clark 2013). (B) Age-adjusted death rate ratios associated with different diseases and fatality causes between individuals of African versus European ancestry in the USA, as summarized in Ref. (Pennington et al. 2009).

Here, we do not attempt to provide a comprehensive overview of the myriad of genetic and non-genetic determinants of inter-individual variation in immune phenotypes (see (Piasecka et al. 2018; Bakker et al. 2018)), for which outstanding reviews have been published elsewhere (e.g., (Brodin and Davis 2017)). Instead, we focus specifically on recent genomic studies characterizing the functional differences in immune response between populations as well as the genetic determinants partially controlling such differences. Further, we discuss the role of natural selection

in driving present-day disparities in immune response between populations, considering cases of local adaptation and adaptively-introgressed alleles.

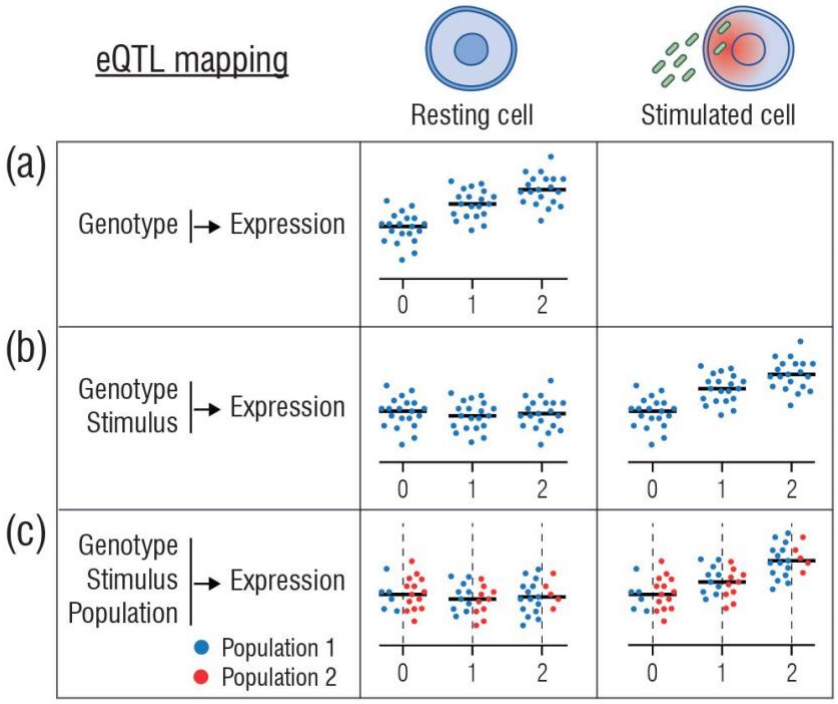


Fig. 2-2. eQTL studies aim to characterize single-nucleotide polymorphisms that significantly impact gene expression levels. (A) The first generation of eQTL studies characterized the extent of genetic control of gene expression levels at baseline in different tissues and cell types (reviewed in Ref. (Fairfax and Knight 2014)). (B) Further studies showed that eQTL effects are often dependent on infection context (Fairfax et al. 2014; Kim-Hellmuth et al. 2017; Alasoo et al. 2018; Barreiro et al. 2012; M. N. Lee et al. 2014; Ferraro et al. 2014; Çalışkan et al. 2015; Piasecka et al. 2018). These studies defined multiple gene-by-environment interactions by identifying immune response-eQTLs – genetic variants that impact the magnitude of response to infection of target genes. (C) More recently, a third round of eQTL studies addressed the role of eQTL-mediated regulation of inter-population differences in gene expression (Quach et al. 2016; Nédélec et al. 2016; Ye et al. 2014). Mapping eQTLs in different populations allows one to determine the proportion of differences in gene expression between populations that are under genetic control, whether as a result of disparities in allele frequencies of global regulatory variants or as a consequence of population-specific mechanisms of gene regulation.

Inter-population variation in disease susceptibility and immune phenotypes

Individuals from distinct regions of the world differ in their susceptibility to infectious diseases as well as chronic inflammatory and autoimmune disorders (reviewed in (Brinkworth and Barreiro 2014; Traherne 2008)). Recent studies suggest that such disparities in disease susceptibility can in part be explained by differences in immune response between individuals of varying genetic ancestry. Nédélec et al. (Nédélec et al. 2016) used RNA-sequencing to characterize the manner in which primary macrophages derived from a panel of 175 healthy individuals who self-identified as either African or European American responded to *Listeria monocytogenes* or *Salmonella typhimurium* infection. Using a similar approach, Quach et al. (Quach et al. 2016) tested for the effect of African versus European ancestry on monocyte response to several Toll-like receptor (TLR) ligands, including TLR1/2, TLR4, and TLR7/8, and the human seasonal influenza A virus. Both studies revealed marked ancestry-associated differences in gene expression and immune responses to infection. Across the experimental conditions tested, on average, 21.3% of the genes appeared to show differential expression between European and African individuals (i.e., population differentially expressed or popDE genes) (Figure 2-3A, left). In addition, up to 16.1% of the genes responding to immune stimulation showed a significant divergence in the intensity of response between European and African individuals (i.e., population differentially responsive or popDR genes) (Figure 2-3A, right).

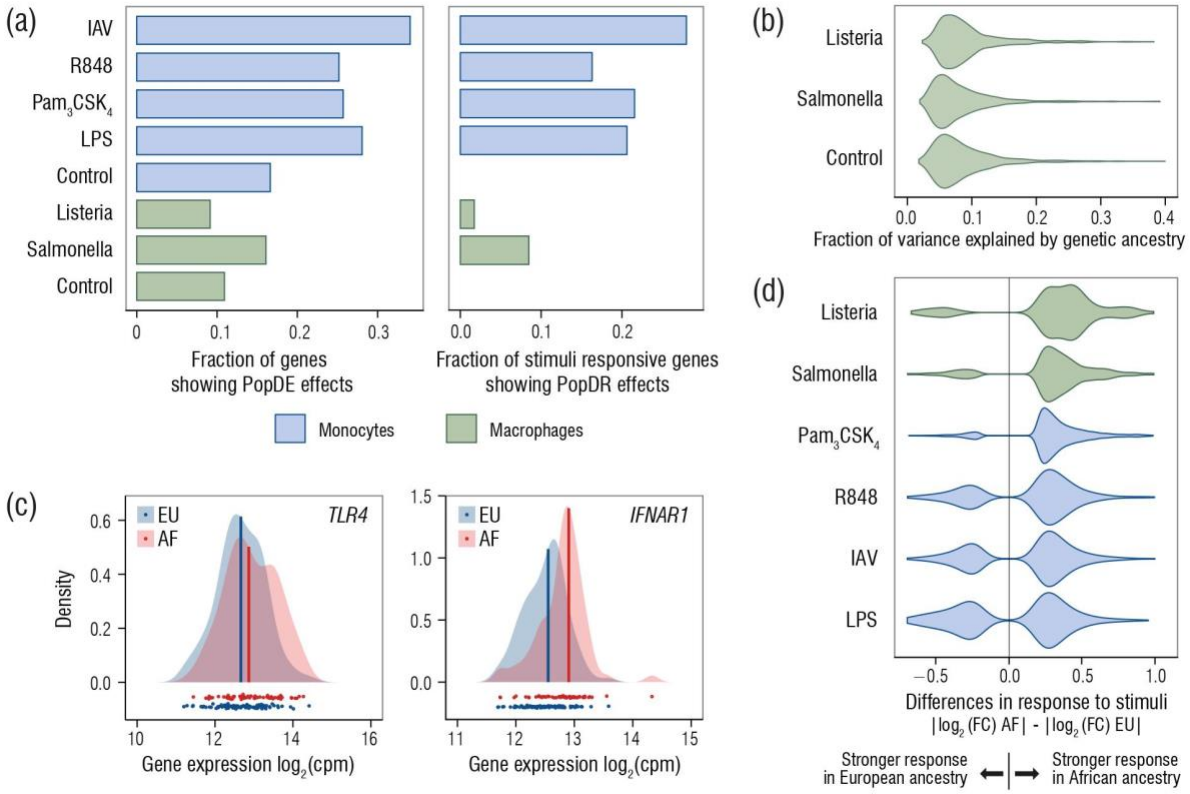


Fig. 2-3. Inter-population differences in immune responses between individuals of African and European ancestry. (A) Left: Fraction of genes expressed in monocytes (data from (Quach et al. 2016)) and macrophages (data from Ref. (Nédélec et al. 2016)) that show inter-population differences in gene expression (popDE genes: fdr<0.05, abs(logFC)>0.2). Right: Fraction of stimuli-responsive genes that exhibit differences in response across populations (popDR genes: fdr<0.05, abs(logFC)>0.2). Data from both studies was reanalyzed following the same analytical pipeline (Nédélec et al. 2016) to ensure that the amount of signal was comparable. (B) Proportion of variance explained by genetic ancestry (quantified as described in (Nédélec et al. 2016)) in genes showing popDE effects in macrophages at fdr<0.05 and abs(logFC)>0.2) (Median: control: 7.3%, *Listeria*: 7.8%, *Salmonella*: 6.8%). (C) Distribution of expression values upon *Salmonella* infection for two representative popDE genes: *TLR4* and *IFNAR1*. Despite the clear shift in the distribution of expression values between Africans and Europeans, the variance explained by genetic ancestry remains small: 4.44% for *TLR4* and 14.67% for *IFNAR1*. (D) Differences in the magnitude of response to immune stimuli between individuals of African and European descent for genes showing significant ancestry-dependent responses to stimulation (fdr<0.05 and abs(logFC)>0.2 both for stimulation and popDR effects).

The number of genes exhibiting significant differences in immune regulation between individuals of African and European descent is substantial. Yet, the proportion of immune response

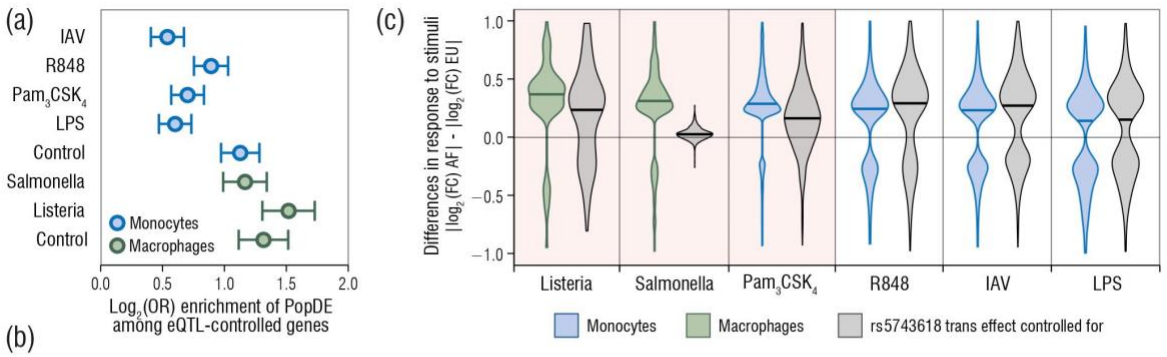
variation due to genetic ancestry remains modest – on average ~7% (Figure 2-3B). In other words, for most genes, two individuals from different populations tend to manifest more similar phenotypes than two individuals from the same population (Figure 2-3C). Despite this, the additive effect of subtle shifts in the distribution of many immune phenotypes (i.e., in the expression levels of many genes across the genome) between populations may be sufficient to explain the reported ethnic disparities in inflammatory and autoimmune disease susceptibility. Supporting this hypothesis, Nédélec et al. showed that genes differentially expressed between populations are significantly enriched for genes associated with immune-related disorders identified by genome-wide association studies. Such diseases include rheumatoid arthritis, systemic sclerosis, and ulcerative colitis, all of which have been reported to differ in incidence or disease severity between African American and European American individuals (Pennington et al. 2009; Richardus and Kunst 2001).

Several lines of evidence from genetic, epidemiological, and functional genomic studies indicate that individuals of African descent engage a stronger transcriptional response to immune stimulation as compared to individuals of European descent, particularly among genes related to the activation of inflammatory responses. As compared to European Americans, African Americans have higher frequencies of alleles associated with increased proinflammatory responses (Ness et al. 2004) and elevated levels of circulating C-reactive protein (Kelley-Hedgepeth et al. 2008). Additionally, African Americans possess immune cells capable of eliciting heightened responses, including T cells able to react more intensely upon TCR activation (Ye et al. 2014) and macrophages better able to control intracellular bacterial growth as well as engage a stronger proinflammatory response to infection (Nédélec et al. 2016). When reanalyzing the data from Quach et al. (Quach et al. 2016) for this Review, we found that monocytes from African individuals

living in Belgium also tend to show a stronger transcriptional response to Pam_3CSK_4 (TLR1/2 ligand) and R848 (TLR7/8 ligand) but not to flu or LPS (Figure 2-3D). This observation suggests that the increased inflammatory capacity often associated with individuals of African ancestry, although common, is not ubiquitous to all immune stimulations or cell types. In addition, it raises the possibility that the differences observed between the two studies are a reflection of unmeasured environmental and social factors that are differentially correlated with genetic ancestry in the US and Belgium.

Inter-population differences in immune responses are partially under genetic control

Both *cis*-acting and *trans*-acting regulatory variants have been implicated in driving differences in immune responses among populations. Indeed, *cis*-eQTL are enriched among genes displaying gene expression differences between populations (Figure 2-4A), and on average, they explain approximately 30% and 50% of the ancestry-related differences in expression in macrophages and monocytes, respectively (Quach et al. 2016; Nédélec et al. 2016). For hundreds of genes, a single *cis*-acting variant is sufficient to account for almost all ancestry effects contributing to immune response heterogeneity (Quach et al. 2016; Nédélec et al. 2016).



(b)

Locus/Hotspot (Closest gene or local gene affected)	Snp	Target genes	Conditions	Target genes and functional annotations	Ref.	MAF across 1K genomes super-populations					Fst (AFR-EUR)
						AFR	AMR	EAS	EUR	SAS	
Ch1: 207.7 Mb (CR1)	rs11117959 (70 SNPs hotspot)	35	S.aureus, C.albicans	IFN-γ response, cell surface proteins and signaling	[24]	0.15	0.27	0.32	0.18	0.5	0.0044
Ch2: 38.3 Mb (CYP1B1)	rs1800440	19 (at LPS) 14 (at IFN-γ)	LPS, IFN-γ	Non-synonymous variant of CYP1B1. Targets related to Cancer and inflammation	[14]	0.007	0.13	0.005	0.2	0.2	0.1983
Ch4: 185.3 Mb (IRF2)	rs13149699	102 (at LPS) 41 (at IFN-γ)	LPS, IFN-γ	Cis-eQTL of IRF2 whose action on the regulator propagates to IRF2 targets	[14]	0.03	0.38	0.41	0.39	0.32	0.3532
Ch4: 38.6-38.9 Mb (TLR1)	rs4833095 (254 SNPs hotspot)	154	S.aureus, E.Coli SEB, BCG	Inflammatory (e.g. IL6, IL1B, IL12B) and regulatory response (e.g. TGFB1, TGFBI, IL1RAP)	[24]	0.12	**0.53**	0.4	**0.72**	0.49	**0.5491**
	rs5743618	432	Pam_3CSK_4	Inflammation (e.g. CCL5, IL10, PTGS2, IL12B) and signaling (BID, IKBKE, PAK1)	[16]	0.04	0.28	0.006	**0.67**	0.07	**0.639**
Ch7: 26.2 Mb (NFE2L3)	rs2057763	5 (at LPS) 19 (at IFN-γ)	LPS, IFN-γ	Ubiquitin proteasomal system	[14]	0.13	0.13	0.05	0.21	0.12	0.021
Ch9: 21 Mb (IFNB1)	rs12553564	251 (at LPS) 175 (at Pam_3CSK_4)	LPS, Pam_3CSK_4	TLR4 and TLR1/2 mediated anti-viral geneset (IFIT1, IFIT3, OAS1, OAS2, CXCL10)	[16]	0.27	0.4	0.5	0.29	0.3	-0.0004
	rs2275888	15	LPS	IFN-β signaling (IRF9, IFI6, EPST11, LY6E, MX1)	[14]	0.3	0.4	0.5	0.29	0.3	-0.0004
Ch17: 33.8-34.1 Mb (AP2B1)	rs11080354 (155 SNPs hotspot)	5	NS, Influenza, C.albicans, S.aureus, E.coli, SEB, BCG	OAZ1, TFRC, LGALS3, CD36, TAL1	[24]	0.26	0.23	0.04	0.31	0.34	0.0054
Ch19: 10.22 Mb (P2RY11)	rs3745601	11	LPS	Non-synonymous coding variant of P2RY11 which modulates ligand-binding	[14]	0.11	0.16	0.33	0.11	0.22	-0.0007

Fig. 2-4. Differences in immune response between populations are under genetic control. (A) Genes showing popDE effects (at fdr<0.05 and abs(logFC)>0.2) are enriched for genes associated with *cis*-eQTLs (fdr<0.01). (B) Major *trans*-eQTLs known to impact immune responses to infection, as reported in Refs. (Fairfax et al. 2014; Quach et al. 2016; Piasecka et al. 2018). (C) Effect of the *TLR1 trans*-eQTL, rs5743618, on genome-wide distributions of inter-population response differences to immune stimuli. Introducing rs5743618 as a covariate in the linear models used to detect popDR effects demonstrates that the weaker proinflammatory response observed in Europeans is in part explained by this single *trans*-eQTL. Conditions for which the difference in response is significantly reduced after correction for the *trans*-effect are highlighted in pink (one-tail Mann-Whitney tests, p<0.005, 2.2E-16 and 2.2E-16 for *Listeria*, *Salmonella*, and Pam3CSK4, respectively).

Several genetic variants near the *TLR1* locus were identified as strong *trans*-eQTLs that regulate the expression levels of hundreds of genes upon stimulation with several immunogens, including *Escherichia coli*, BCG, and the TLR1/2 agonist Pam_3CSK_4 (Quach et al. 2016; Piasecka et al. 2018). The causal *trans*-eQTL in this locus is most likely a non-synonymous variant in *TLR1* (rs5743618), which was shown to cause dampened inflammatory immune responses due to hindered NF-κB signaling and activation (Quach et al. 2016; Barreiro et al. 2009). Interestingly, this SNP is almost absent in African populations but is found at very high frequency in European populations (Derived Allele Frequency $(DAF)_{Africa} = 0.04$, $DAF_{Europe} = 0.67$; $F_{st} = 0.639$; Figure 2-4B). To test whether this difference in allele frequency could explain the stronger transcriptional response to immune stimulation often associated with increased African ancestry, we included it as a covariate in the linear models used to quantify ancestry effects on immune response when reanalyzing the data from Quach et al. and Nédélec et al. Strikingly, we found that correcting for the effects of this *trans*-eQTL alone was enough to reduce the differences in the magnitude of response between African and European individuals by 26.5%, 91.1%, and 41.5% in macrophages infected with *Listeria* and *Salmonella* and monocytes stimulated with Pam_3CSK_4, respectively (Figure 2-4C, left). As expected, this same SNP has no effect on population differences in response when considering immune challenges that do not signal through TLR1 (Figure 2-4C, right). *IRF2* has also been identified as a major regulatory hub, as a single *cis*-eQTL was shown to moderate levels of the transcription factor IRF2, an effect which in turn was propagated to hundreds of *trans*-regulated IRF2 target genes following IFN-γ stimulation (Fairfax et al. 2014). This *trans*-eQTL is also extremely differentiated between African and European populations (Figure 2-4B) and, therefore, is likely to account for population differences in response to IFN-γ, an important cytokine associated with multiple autoimmune and autoinflammatory disorders (Schoenborn and

Wilson 2007). However, this hypothesis remains to be formally tested. Taken together, these findings indicate that host genetic factors markedly contribute to the differences in immune response observed between populations.

The contribution of natural selection and adaptive introgression to immune response diversity among populations

The importance of pathogen-mediated selection on human evolution can be characterized by utilizing population genetics approaches to study signatures of selection at functionally relevant sites throughout the genome (Fumagalli and Sironi 2014). Classical examples of adaptation to particular pathogenic environments include the selection for variants that confer reduced malaria risk in malaria-endemic regions (Kwiatkowski 2005; W.-Y. Ko et al. 2011) and for alleles exhibiting protective effects against African trypanosomiasis among African populations (W.-Y. Ko et al. 2013). Yet, the extent to which positive selection has contributed to inter-population variation in immune response remains relatively unexplored (Quach et al. 2016; Nédélec et al. 2016). In macrophages and monocytes, *cis*-regulatory variants identified both at baseline and in response to stimulation are enriched for signatures of positive selection, consistent with the significance of regulatory genetic variation in recent human evolution (Fraser 2013). More importantly, among genes showing the strongest signatures of selection, both Nédélec et al. and Quach et al. observed a pronounced enrichment of genes differentially regulated in response to immune stimulation between individuals of African and European descent. These findings provide the strongest empirical evidence to date that natural selection has crucially influenced present-day inter-population disparities in innate immune response to pathogenic infections.

Numerous immune response differences between populations are a direct consequence of local adaptation. Some examples include the variable expression of *CCR1*, a key chemokine receptor involved in the trafficking and proliferation of myeloid progenitor cells (Broxmeyer et al. 1999), *HLA-DQA1*, the major genetic factor associated with susceptibility to celiac disease (Abadie et al. 2011), and *IRF5*, a key transcription factor associated with susceptibility to systemic lupus erythematosus, rheumatoid arthritis, ulcerative colitis, and systemic sclerosis (Eames, Corbin, and Udalova 2016). The *trans*-eQTL in *TRL1* associated with an attenuated proinflammatory response in individuals of European descent is also associated with multiple signatures of positive selection (Quach et al. 2016; Barreiro et al. 2009). This observation raises questions about the possible evolutionary conflict between mounting a strong inflammatory response to effectively fight pathogens and avoiding the detrimental consequences of acute and chronic inflammation which can lead to tissue damage and the development of autoinflammatory and autoimmune diseases (Okin and Medzhitov 2012).

Neanderthal ancestry makes up ~2% of the ancestry of living humans found outside of Africa (Kelso and Prüfer 2014). Thus, these introgressed genetic variants may contribute to a portion of the observed ancestry-related differences in gene expression, especially if such variants enabled the ancestors of modern Europeans to more rapidly adapt to new pathogenic environments (Ségurel and Quintana-Murci 2014). In support of this theory, increased levels of Neanderthal admixture can be detected in innate immune system genes, as compared to the rest of the coding genome, in Europeans and Asians (Deschamps et al. 2016). Furthermore, regulatory variants introgressed into European genomes have been demonstrated to modulate immune responses of present-day Europeans, especially concerning responses to viral immune challenges (Quach et al. 2016; Nédélec et al. 2016). The *OAS* gene cluster provides one particular example of a genomic

region harboring adaptively-introgressed variants important for antiviral activity that reach frequencies above 40% in all Eurasian populations and are absent in Africans (Mendez, Watkins, and Hammer 2013; Sams et al. 2016). Positive selection of a Neanderthal haplotype in this region led to the reintroduction of an ancestral splice variant of *OAS1* associated with higher enzymatic activity (Sams et al. 2016) and reduced susceptibility to several flaviviruses (Bigham et al. 2011; El Awady et al. 2011). More broadly, recent studies suggest that RNA viruses acted as major drivers of adaptive introgression between Neanderthals and modern humans, further supporting a central role for pathogens in guiding recent human evolution (Alvarez et al. 2017).

Conclusions and perspectives

Our improved ability to combine immunogenomic approaches, population genetics tools, and classical immunological assays has provided unique insights into how natural selection has impacted the function of the human immune system. Still, much more work needs to be done before we can fully comprehend the role of natural selection in immune system evolution globally. Although the studies discussed herein argue for the importance of positive selection in shaping immune system differences among populations, their scope is limited to the search of footprints left by classical selective sweeps. However, many traits in humans, including most immune phenotypes, are highly polygenic (Pritchard, Pickrell, and Coop 2010; Boyle, Li, and Pritchard 2017). Consequently, this observation predicts that most selection in immune-related genes will be driven by polygenic adaptation. Future studies investigating signatures of polygenic evolution at the level of entire immune pathways, gene regulatory modules, and protein-protein interaction networks are needed to test this hypothesis.

In addition, to date, studies have focused on individuals of European or African ancestry. Broadening immune response eQTL studies to include a larger array of human populations is now required in order to capture the full extent of variation in immune responses across the globe, notably among populations that historically have faced distinct pathogen pressures. For example, many epidemiologists and anthropologists have hypothesized that the agricultural transition (beginning 10-12 kyr ago) precipitated novel infectious disease burdens (Wolfe, Dunavan, and Diamond 2007), which likely contributed to the strong signatures of recent adaptation detected in immunity genes (Karlsson, Kwiatkowski, and Sabeti 2014; Deschamps et al. 2016). Yet, given the lack of comparative functional studies interrogating immune responses between pairs of populations that previously differed in their modes of subsistence (i.e., hunter-gatherers (HG) versus agriculturalists (AGR)), the impact of agriculture on immune system evolution remains unclear. Future efforts aimed at characterizing immune phenotypes between HG and AGR populations will likely provide novel insights into how the agricultural transition influenced the evolution of immune responses to pathogens.

Finally, immune response eQTL studies have only assessed bulk cell populations at a limited number of time points post-infection. Therefore, they ignore two critical aspects of an immune response: 1) that response to infection is a dynamic process that cannot accurately be typified by a single event, and 2) that important cell-to-cell variation may exist within a population. By leveraging new technological advances in automated cell culture and single-cell sequencing, future studies will be able to generate a much more realistic and nuanced picture of how genetic variation shapes immune responses across human populations. Such studies will expand our current knowledge of the processes through which natural selection has and continues to contribute to the diversification of immune responses across human populations.

Acknowledgements

We thank Maitane Gutierrez for assistance with the figures. This work was supported by the Canadian Institute of Health Research (CIHR) operating grants (301538 and 232519) to LBB. LBB holds a Canada Research Chair (950-228993). JS was supported by the by a Banting Fellowship from CIHR. HER was supported by a CHU Sainte-Justine Foundation Fellowship. We thank Calcul Québec and Compute Canada for providing access to the supercomputer Briaree from the University of Montreal.

Chapter III: Genetic ancestry effects on the response to viral infection are pervasive but cell type specific

Note:

The following section (*Chapter III*) is reproduced verbatim, with the exception of figure numbering and reference labeling, from my first authored reference "Genetic ancestry effects on the response to viral infection are pervasive but cell type specific" (Randolph et al., 2021). This project was published in *Science* on November 26, 2021.[1]

Authors:

H.E. Randolph, J.K. Fiege, B.K. Thielen, C.K. Mickelson, M. Shiratori, J. Barroso-Batista, R.A. Langlois, L.B. Barreiro

[1] From Randolph, H.E., Fiege, J.K., Thielen, B.K., Mickelson, C.K., Shiratori, M., Barroso-Batista, J., Langlois, R.A., Barreiro, L.B. (2021). Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. Science, 374, 1127–1133. Reprinted with permission from AAAS.

Abstract

Humans differ in their susceptibility to infectious disease, partly due to variation in the immune response following infection. We used single-cell RNA-sequencing to quantify variation in the response to influenza infection in peripheral blood mononuclear cells from European- and African-ancestry males. Genetic ancestry effects are common but highly cell type-specific. Higher levels of European ancestry are associated with increased type I interferon pathway activity in early infection, which predicts reduced viral titers at later time points. Substantial population-associated variation is explained by *cis*-expression quantitative trait loci that are differentiated by genetic ancestry. Furthermore, genetic ancestry-associated genes are enriched among genes correlated with COVID-19 disease severity, suggesting that the early immune response contributes to ancestry-associated differences for multiple viral infection outcomes.

Full Text

Pathogenic viruses are among the strongest sources of selection pressure in human evolution (Fumagalli et al. 2011; Enard and Petrov 2018). Prior to the modern era, however, global pandemics on the scale of the 1918 Spanish influenza or the SARS-CoV-2 pandemic were probably rare due to the restricted potential for long-distance exchange (Enard and Petrov 2020). If past viral epidemics were geographically stratified, they may have driven population divergence in the frequencies of polymorphisms that mediate the immune response to viral infection. Testing this hypothesis is therefore valuable both for understanding human evolutionary history and for explaining differential susceptibility to viral epidemics in the present-day.

Indeed, genetic effects on the response to viruses are well-known in human populations

(Kenney et al. 2017). For example, over 120 genetic variants have been identified in humans that predict the gene regulatory response to influenza A virus (IAV) in dendritic cells (M. N. Lee et al. 2014). Variation in the transcriptional response to IAV *in vitro* is also correlated with genetic ancestry in monocytes derived from individuals of African and European descent (Quach et al. 2016). These results suggest that genetic divergence between human populations, especially at loci that are moderately differentiated by genetic ancestry, plays an important role in shaping the immune response to viral infection. However, because studies to date focus on isolated cell types (M. N. Lee et al. 2014; Quach et al. 2016), they fail to capture interactions between immune cell types necessary to mount an efficient antiviral response. They also leave unclear whether genetic ancestry effects are unique to, or generalize across, distinct immune cell types.

To address these limitations, here we combined single-cell RNA-sequencing with *in vitro* IAV infection assays in peripheral blood mononuclear cells from study subjects with varying degrees of European versus African genetic ancestry.

Single-cell profiling of the transcriptional response to influenza infection

We exposed peripheral blood mononuclear cells (PBMCs) from a diverse panel of humans (Table S3-1) to either a mock treatment or the pandemic H1N1 Cal/04/09 influenza A virus (IAV) strain (multiplicity of infection 0.5) (n = 180 samples, paired mock-exposed and IAV-infected samples from each of 90 males). We focused on males to avoid potential effects of sex-specific differences in expression (Oliva et al. 2020), which would reduce the power of our study. Following 6 hours of viral exposure, we performed single-cell RNA-sequencing on all samples (Fig. 3-1A). In total, we captured 255,731 single-cell transcriptomes across all individuals and conditions (n = 235,161 high-quality cells, Table S3-1). We also performed whole-genome

sequencing to estimate the proportion of African and European ancestry for each individual (n = 89 individuals who were successfully genotyped; Fig. S3-1A, Table S3-1). Clustering revealed eight distinct immune cell types (Fig. 3-1B), with five major cell clusters corresponding to the main PBMC cell types ($CD4^+$T cells, $CD8^+$T cells, B cells, natural killer (NK) cells, and monocytes).

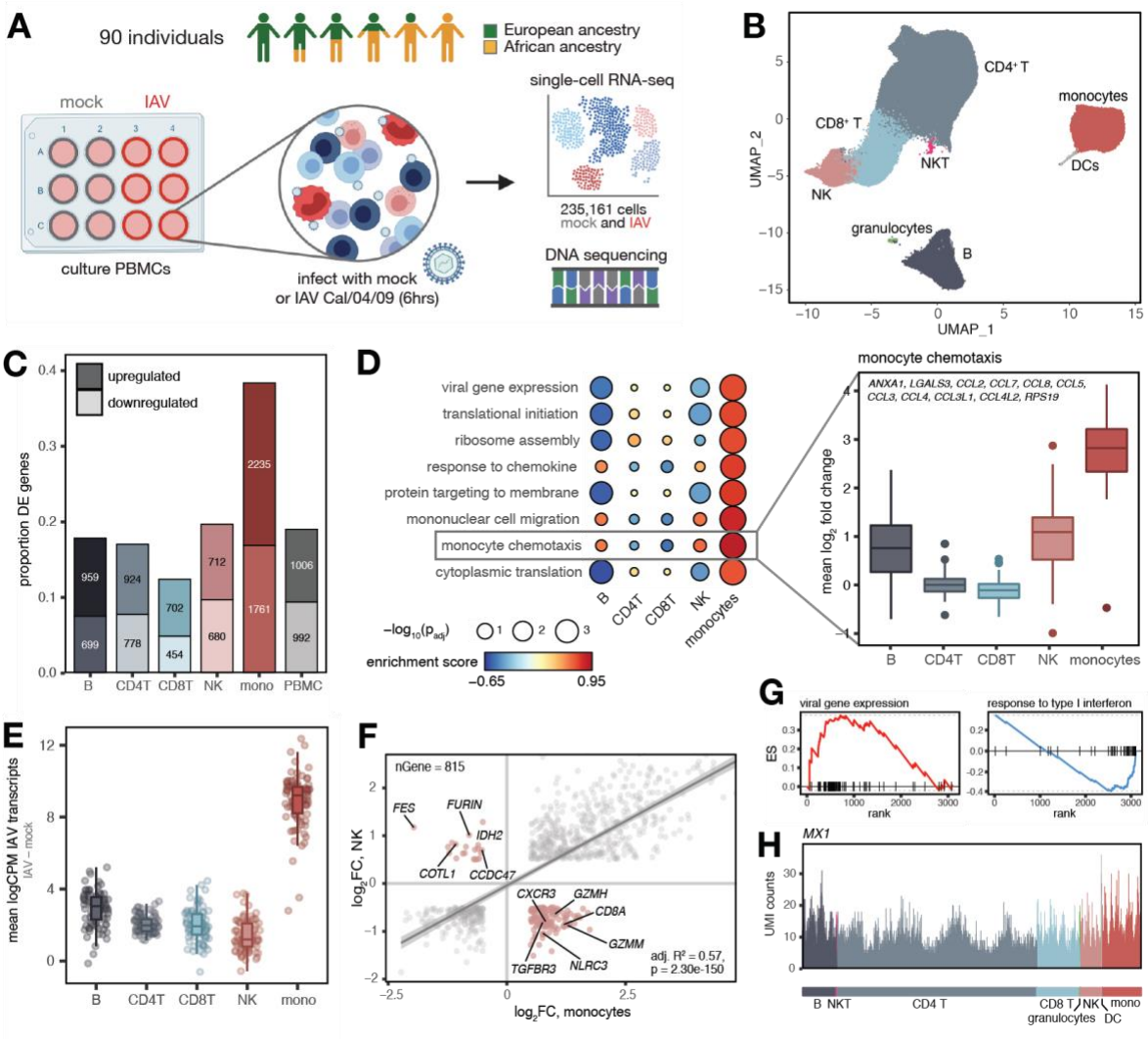


Fig. 3-1. Shared and cell type-specific responses to IAV infection. (A) Study design. (B) UMAP of 235,161 mock and IAV-infected cells across individuals. (C) Numbers and proportions of differentially expressed genes upon infection. (D) Upregulated (FDR < 0.10) monocyte-specific GO pathways following infection (Table S3-3). "Monocyte chemotaxis" genes display greater upregulation in monocytes (plotted means for each individual across genes in IAV minus mock

Fig. 3-1, continued. condition, t-tests, all p-values < $1x10^{-10}$ compared to each other cell type). (E) Distribution of IAV transcripts across cell types. (F) Correlation between global infection effect sizes in monocytes and NK cells among DE genes in both cell types (n = 815). P-value and best-fit slope was obtained from a linear regression model. Highlighted genes (pink) display discordant responses. (G) Example pathways enriched among genes with high (viral gene expression) and low (response to type I interferon) specificity scores. Genes are rank-ordered by specificity score (x-axis, highest to lowest). (H) UMI counts per cell in the IAV-infected condition for an example IFN-inducible gene (*MX1*) with a ubiquitous expression pattern.

We first investigated the overall signature of IAV infection by collapsing the single-cell gene expression values for each of the five main clusters and all cells together (i.e., "PBMCs") to generate pseudobulk estimates for each sample. Principal component analysis (PCA) of the PBMC pseudobulk data revealed a marked separation of mock and IAV-infected samples on PC1, which explains 43% of the variance in the dataset (Fig. S3-1B, paired t-test, p < $1x10^{-10}$). Monocytes were the most responsive to IAV infection (n = 3,996 differentially expressed (DE) genes identified using limma (Ritchie et al. 2015) [38.3% of those tested compared to 12.4 – 19.6% in other cell types], $|\log_2$ fold-change$|$ > 0.5, false discovery rate [FDR] < 0.05) (Figs. 3-1C and 3-1D, Tables S3-2 and S3-3). Monocytes also exhibited the highest levels of intracellular IAV transcripts (i.e., influenza-derived transcripts generated and processed by infected host cells; 3-6-fold increase in IAV transcript levels in monocytes relative to all other cell types, all t-test p-values < $1x10^{-10}$) (Fig. 3-1E). This observation is consistent with previous work showing that, among blood mononuclear cells, monocytes are particularly susceptible to viral infections (Hou et al. 2012).

We then explored the extent to which the infection response was shared across cell types. Overall, responses were strongly correlated (Pearson's r range 0.65 – 0.95 for pairwise effect size correlations across cell types among DE genes following IAV infection, Fig. S3-1C). However, discordant responses were also observed. For example, among differentially expressed genes

shared by monocytes and NK cells (n = 815), 135 genes (16.6%, Fig. S3-1D) responded to IAV infection in opposite directions (Fig. 3-1F). These findings underscore the importance of considering immune responses in a cell type-specific manner. Not only does this approach better capture the biological origins of variation in the response to viral infection, but it also avoids false negative or potentially misleading results that can emerge from bulk analysis.

To further dissect cell type-specific versus shared responses, we generated a specificity score based on variation in the strength of responses across cell types for all genes significantly differentially expressed in at least one cell type (Table S3-4, see Methods for details). Genes with highly cell type-specific response patterns were enriched for roles in translational initiation and viral gene expression (FDR < $1x10^{-10}$ for both terms, Fig. 3-1G, left, Table S3-4). In contrast, genes with low specificity scores were enriched for pathways related to type I interferon (IFN) signaling (FDR < $1x10^{-10}$) and response to type I IFN (FDR = $2.9x10^{-3}$) (Fig. 3-1G, right, Table S3-4). Thus, concordant with previous work in mice (Steuerman et al. 2018), our data show that the induction of IFN-related genes is a fundamental component of the antiviral response shared across immune cell types (Fig. 3-1H).

Increased European genetic ancestry predicts a stronger type I/II IFN response following IAV infection

We next identified genes for which expression levels are correlated with quantitative genetic ancestry estimates (i.e., proportion of estimated African ancestry) at baseline, following infection, or both (controlling for age, batch, and other technical covariates). To increase power and improve our effect size estimates for these "population differentially expressed" (popDE) genes, we applied a multivariate adaptive shrinkage method (mash) (Urbut et al. 2019), which

leverages the correlation structure of genetic ancestry effect sizes across cell types (see Methods for details of statistical models). Across conditions and cell types, we identified 1,949 unique popDE genes (local false sign rate [lfsr] < 0.10), ranging from 830 in NK cells to 1,235 genes in $CD4^{+}$T cells (Figs. 3-2A and S3-2A for distribution of effect sizes, Table S3-5). Within each cell type, most popDE genes were shared between conditions (52.9% in monocytes – 77.4% in $CD8^{+}$T cells, Fig. 3-2A). In contrast, across cell types, genetic ancestry effects on gene expression were highly cell type-specific, such that the majority of popDE genes were identified in only one or two cell types (52.2% in mock, 51.4% in IAV-infected, Figs. 3-2B. and S3-2B, left). Only 17.8% (mock) and 24.7% (IAV-infected) of popDE genes exhibited shared genetic ancestry effects across all five cell types (Figs. 3-2B. and S3-2B, right). Notably, despite differences in study subject country of origin, IAV strain, and experimental design, our popDE effect size estimates for monocytes were largely concordant with those derived from an independent bulk RNA-seq dataset of IAV-infected monocytes from European- and African-ancestry individuals (Quach et al. 2016) (Pearson's r = 0.662 [mock], Pearson's r = 0.499 [IAV], $p < 1.0x10^{-10}$ in both conditions, Figure S3-2C).

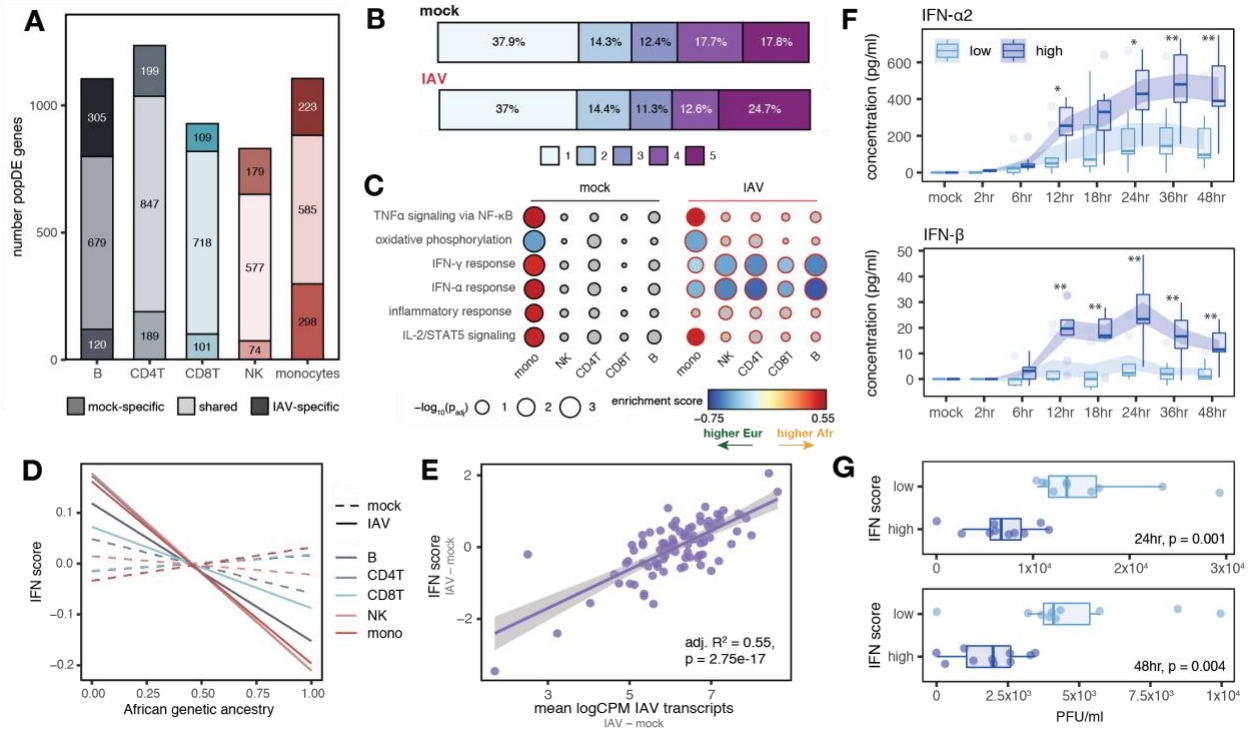


Fig. 3-2. Genetic ancestry influences the immune response to IAV infection. (A) Number of shared and condition-specific popDE genes. (B) Cell type sharing of popDE effects (1 = detected in a single cell type, 5 = detected across all cell types). (C) GO enrichments for popDE effects in the mock- and IAV-infected conditions. Colored circles represent pathways with FDR < 0.10. IFN pathways are among the most divergent between European and African-ancestry individuals in monocytes, with 26% (42 out of 163) of all IFN genes tested classified as popDE after infection. (D) Correlation between African genetic ancestry proportion and IFN score in mock (dotted lines) and IAV-infected conditions (solid lines). (E) IAV transcript levels are associated with IFN response in PBMCs. (F) Secreted IFN-α2 and IFN-β levels in low versus high IFN responders over a 48 hour time course. Shaded area represents the mean ± SE. *p < 0.02, **p < 0.009 (Mann-Whitney U tests). (G) Viral titers (plaque-forming units, PFU/ml) detected in supernatant 24 and 48 hpi. In (D) and (E), p-values and best-fit slopes were obtained from linear regression models.

To identify the functional pathways most closely associated with genetic ancestry, we performed gene set enrichment analysis for the MSigDB Hallmark gene sets (Liberzon et al. 2015) (Fig. 3-2C, Table S3-6). In monocytes, we identified significant enrichments for multiple immune pathways prior to infection, including IFN-α response (FDR = 1.9×10^{-3}), IFN-γ response (FDR = 5.4×10^{-4}), TNFα signaling via NF-κB (FDR = 6.1×10^{-4}), IL-2/STAT5 signaling (FDR = 2.1×10^{-3}),

and inflammatory response (FDR = 0.012) (Fig. 3-2C). In these cases, the enrichments were identified for genes more highly expressed at baseline in individuals with a greater proportion of African ancestry. Intriguingly, in IAV-infected monocytes, this pattern reversed: post-infection, we observed an enrichment of type I and II IFN pathways (IFN-α response FDR = 0.014, IFN-γ response FDR = 0.040 in monocytes) in genes more highly expressed with increasing European ancestry (Fig. 3-2C). Notably, this enrichment of type I/II IFN pathways among genes more highly expressed with greater European ancestry after infection was even more clear in the other four cell types (FDR range: 0.03 – 4.1×10^{-4}, Table S3-6). To further characterize genetic ancestry-associated differences in the IFN response, we constructed a per-sample score of interferon signaling activity, the "IFN score," which provides an estimate of the average expression of genes belonging to the hallmark IFN-α and IFN-γ gene sets for each individual (Methods). Again, increased European ancestry was strongly correlated with increased IFN score, but only following infection (mean Pearson's r across cell types = -0.26, Fisher's meta-p = 2.9×10^{-6} for IAV-infected; mean Pearson's r = -0.0045, Fisher's meta-p = 0.746 for mock) (Figs. 3-2D and S3-2D for cell type-specific associations).

These findings suggest that genetic ancestry may also predict the magnitude of the response to IAV infection. In support of this idea, we identified 445 genes for which genetic ancestry was associated with the magnitude of the response to infection (i.e., "population differentially-responsive" [popDR] genes, lfsr < 0.10). PopDR genes were found for all five cell types but were most common in monocytes (popDR genes: $n_{monocytes} = 272$ versus range = 53 – 181 in other cell types). A core set of 21 popDR genes was shared across all cell types (Fig. S3-3A, Table S3-7). Increased European genetic ancestry predicted a stronger type I/II IFN response (measured as the difference in IFN score between the IAV-infected and mock conditions per individual) across cell

types (mean Pearson's r = -0.23, Fisher's meta-p = 6.0×10^{-5}, Fig. S3-3B). This observation was not explained by baseline levels of Cal/04/09-specific serum IgG antibodies (a proxy for prior exposure to IAV), which were uncorrelated with genetic ancestry, the transcriptional response to IAV (Figs. S3-3C, D), and HLA genotype (Methods). However, stronger type I/II IFN responses predicted increased intracellular IAV transcript levels in PBMCs (adj. $R^2 = 0.55$, $p = 2.8 \times 10^{-17}$, Figs. 3-2E and S3-3E for cell type-specific effects). IAV transcript levels were also significantly higher in individuals with increased European ancestry (Pearson's r = -0.32, p = 0.002, Fig. S3-3F).

An early-induced type I IFN response is associated with decreased viral titers at later time points

To functionally validate our findings, we infected PBMCs from the 20 individuals with the strongest (n = 10, "high responders") and weakest (n = 10, "low responders") transcriptional type I/II IFN responses at 6 hours post infection (hpi) with IAV. We collected secreted cytokine measurements across 8 time points over 48 hours and viral titer measurements at 24 and 48 hpi. High responders produced significantly more secreted IFN-α2 (Fig. 3-2F, top) and IFN-β (Fig. 3-2F, bottom) than low responders beginning at 12 hpi, an effect that was exacerbated over time to 4-fold (IFN-α2) and 11.6 fold (IFN-β) more by 48 hpi ($p < 0.007$ for both cytokines, Mann-Whitney U tests). Viral titers quantified from supernatant at 24 and 48 hpi were also reduced in high responders compared to low responders (Mann-Whiney U tests, p = 0.001 for 24 hpi, p = 0.004 for 48 hpi, Fig. 3-2G). None of the 20 study subjects in this experiment harbored predicted loss-of-function mutations among genes associated with defects in type I IFN signaling (Zhang 2020; Zhang et al. 2020), suggesting that these results are not driven by rare genetic variants (Table

S3-1). Taken together, these results indicate that individuals better able to mount type I IFN responses shortly after infection also displayed a greater capacity to limit productive viral replication later in infection/at later time points. These observations are consistent with the finding that individuals with rare immunodeficiencies leading to defects in type I IFN signaling restrict viral replication poorly and, subsequently, are at increased risk for severe influenza (Ciancanelli et al. 2016; Thomsen et al. 2019).

Cis-regulatory genetic variation explains ancestry-associated differences in gene regulation

To assess the contribution of genetic variation to genetic ancestry-associated differences in the transcriptional response to IAV infection, we mapped expression quantitative trait loci (eQTL) in the mock and IAV-infected samples. We focused on *cis*-eQTL, which we defined as SNPs located either within or flanking (±100 kilobases) each gene tested. We identified at least one *cis*-eQTL for 2,234 genes (lfsr < 0.10, hereafter referred to as eGenes) across all cell types and conditions (Fig. 3-3A, Table S3-8). Independent bulk RNA-sequencing generated from the same samples validated our eGene discovery in the scRNA-seq data (Figs. S3-4A, B; average adj. R^2 = 0.71 for eGene effect sizes in the pseudobulk scRNA-seq and bulk RNA-seq datasets).

Although many variants are shared across cell types and conditions (45%, Fig. S3-4C), 13 – 24% of the eGenes identified within each cell type were only detected in one condition even after probing shared effects with mash (Urbut et al. 2019). A small set of 29 eGenes were also only detectable following infection, including the key IFN-inducible genes *OAS1* (Fig. 3-3B), *IFI44L*, *IFIT1*, *IRF1*, and *ISG15* (Fig. S3-4C).

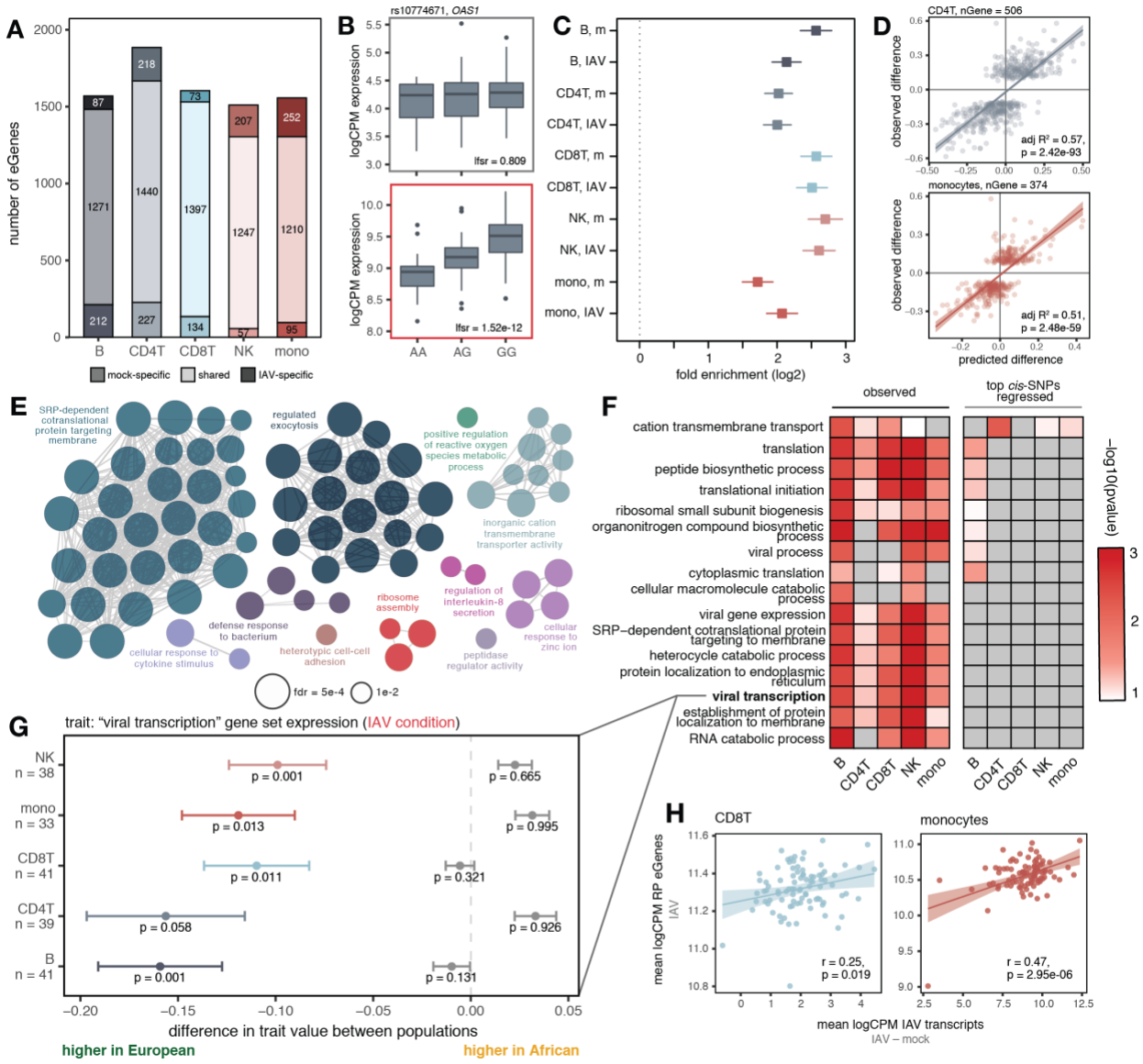


Fig. 3-3. *Cis*-regulatory variation drives differences in the antiviral response. (A) Number of shared and condition-specific eGenes. (B) Condition-specific eQTL example (rs10774671) in CD4^{+}T cells (top: mock, bottom: IAV-infected). (C) Enrichment of eGenes among popDE genes in each cell type/condition determined using logistic regression (log_2-fold enrichment with 95% confidence interval; "m" = mock). (D) Correlation of *cis*-predicted (x-axis) versus observed (y-axis) population differences in expression among popDE genes with an eQTL in CD4^{+}T cells and monocytes. (E) Significant ClueGO enrichments (hypergeometric test, FDR < 0.01) for popDE eGenes across cell types in the IAV-infected condition. (F) Heatmap of -log_{10} p-values in support of median ancestry-associated differences in gene expression among a subset of enriched GO terms (left) and a model estimating this effect after regressing out the effects of the top *cis*-SNPs for all genes contained in the term (right). (G) Example of a GO term for which patterns of population variation are compatible with polygenic selection. PopDE genes with an eQTL that belong to the GO term "viral transcription" (n range = 33 – 41 genes) show consistently higher expression levels

Fig. 3-3, continued. in European-ancestry individuals (median observed ancestry-associated difference (x-axis) < 0, colored points +/- SE). Following *cis*-SNP regression (gray points +/- SE), the overall trend for higher expression of viral transcription genes in European- compared to African-ancestry individuals is no longer significant. Empirical p-values were calculated using a permutation-based approach for (F) and (G) (Methods for details). (H) Correlation between IAV transcripts and ribosomal protein eGene expression in $CD8^{+}$T cells and monocytes. In (D) and (H), p-values and best-fit slopes were obtained from linear regression models.

We next tested whether eGenes were likely to be differentially expressed by genetic ancestry. Across cell types and conditions, eGenes (lfsr < 0.10) were 3.2 to 6.5-fold more likely to be classified as popDE (lfsr < 0.10) than expected by chance (Fig. 3-3C), and 1.3 to 5.0-fold more likely to specifically belong to the set of IFN-associated popDE genes (Fig. S3-4D). These enrichments suggest that ancestry-associated differences in gene expression are likely to have a substantial genetic component, perhaps due to divergence in allele frequencies at the causal eQTL. To test this hypothesis, we calculated the correlation between 1) the estimated genetic ancestry effect from our popDE analysis, and 2) the predicted genetic ancestry effect from the effect size of the top eQTL per eGene and the dosage genotype for this SNP across individuals (restricted to popDE genes that were also eGenes in at least one cell type, n = 835 genes; see Methods for details). The genotype and eQTL effect size for the top eQTL alone explained an average of 52.5% (mock) and 53.6% (IAV-infected) of the variance in genetic ancestry effect sizes across cell types (Figs. 3-3D and S3-4E). Thus, among popDE genes with an eQTL, over 50% of population differences are explained by differences in the frequency of *cis*-regulatory variants.

Polygenic selection on ribosomal protein gene expression

We next sought to evaluate if the intersection set of popDE genes and eGenes clustered into specific biological pathways. Among popDE genes where we also observed eQTL, we identified a strong enrichment for many Gene Ontology (GO) terms related to transcriptional and

translational processes, including ribosomal small subunit biogenesis and viral transcription (FDR < $3x10^{-10}$ in mock and IAV-infected, Fig. 3-3E, Table S3-9). Consistent population differences in the expression of genes within the same pathway/gene set could be explained by two hypotheses. First, genes in a given gene set may have evolved under relaxed evolutionary constraint, allowing *cis*-regulatory variants for these genes to diverge in frequency across populations due to genetic drift. Alternatively, if variants within a given pathway have been a repeated target of selection, they may have experienced directionally concordant shifts in allele frequencies across populations – a pattern consistent with polygenic selection.

We tested for such a pattern in each of the popDE eGene-enriched pathways in all cell type-condition combinations (n = 10: five cell types in the mock and IAV-infected conditions). To do so, we calculated the median genetic ancestry-associated effect on gene expression (i.e., popDE effect size) across all popDE genes in the gene set that also had an eQTL. Under the hypothesis of neutrality, we expect the direction of ancestry-associated effects to be randomly distributed: some genes will be more highly expressed in European-ancestry individuals whereas others will be more highly expressed in African-ancestry individuals. In contrast, under polygenic selection, we expect to find a directional effect, such that most genes for a given pathway show higher expression in one ancestry group versus the other (Pritchard, Pickrell, and Coop 2010). Consistent with a history of polygenic selection, most of the GO terms for ribosomal protein (RP)-related pathways (e.g., ribosomal biogenesis, viral transcription, etc.) show gene expression levels that are consistently higher in individuals with increased European ancestry across cell types (Figs. 3-3F, "observed"; 3-3G, colored bars). This pattern holds in both mock-exposed (Fig. S3-4F) and IAV-infected cells (Figs. 3-3F, 3-3G).

An alternative explanation for this observation is that global ancestry is correlated with consistent, directionally biased environmental effects on the expression of genes in RP-related pathways. If so, controlling for local genetic effects on gene expression (e.g., *cis*-eQTL where allele frequencies are not strongly correlated with ancestry) should not affect the ancestry-gene expression relationship. However, we find the opposite pattern. Specifically, when the effect of the top *cis*-eQTL for each gene is regressed out, the directional bias towards higher expression with increased European ancestry disappears for all RP-related enriched pathways (Figs. 3-3F, "top *cis*-SNPs regressed"; 3-3G, gray bars). Thus, our results suggest that the higher expression of transcription and translation-related pathways in European-ancestry individuals is driven by the cumulative effect of *cis*-regulatory variants that affect the regulation of genes within these pathways. This shift may in turn be explained by viral infection-induced selection pressures. In support of this possibility, we observed a strong correlation between the average expression of RP eGenes and IAV transcript expression in both $CD8^{+}$T cells (Pearson's $r = 0.32$, $p = 0.002$) and monocytes (Pearson's $r = 0.58$, $p < 1x10^{-10}$, Fig. 3-3H).

Genes differentially expressed between African- and European-ancestry individuals are enriched among genes associated with COVID-19 severity

The immune pathways activated in response to IAV largely overlap those triggered by other single-stranded RNA viruses (Jensen and Thomsen 2012). Thus, our dataset provides an opportunity to evaluate whether differences in COVID-19 susceptibility (caused by SARS-CoV-2, another single-stranded RNA virus) in African Americans and non-Hispanic white Americans (J. Y. Ko et al. 2021) could be partially explained by differences in population genetic history. We reasoned that if the genetic ancestry-associated differences in gene expression identified in our *in*

vitro infection model also affect susceptibility to COVID-19, those genes should be enriched among genes associated with COVID-19 disease severity *in vivo*. To test this hypothesis, we re-analyzed a publicly-available single-cell RNA-sequencing dataset consisting of 505,616 PBMCs across 129 COVID-19 patients with varying degrees of disease severity (Su et al. 2020) based on the World Health Organization Ordinal Scale (WOS) for Clinical Improvement (see Methods for details). Using a model adjusting for age, sex, and self-identified race and ethnicity, we identified genes where expression levels correlated with severity ("COVID severity-associated genes") within each of the five PBMC cell types included in the IAV data set. Monocytes, by far, displayed the largest number of genes associated with severity (n = 839, lfsr < 0.01) (Fig. 3-4A, Table S3-10).

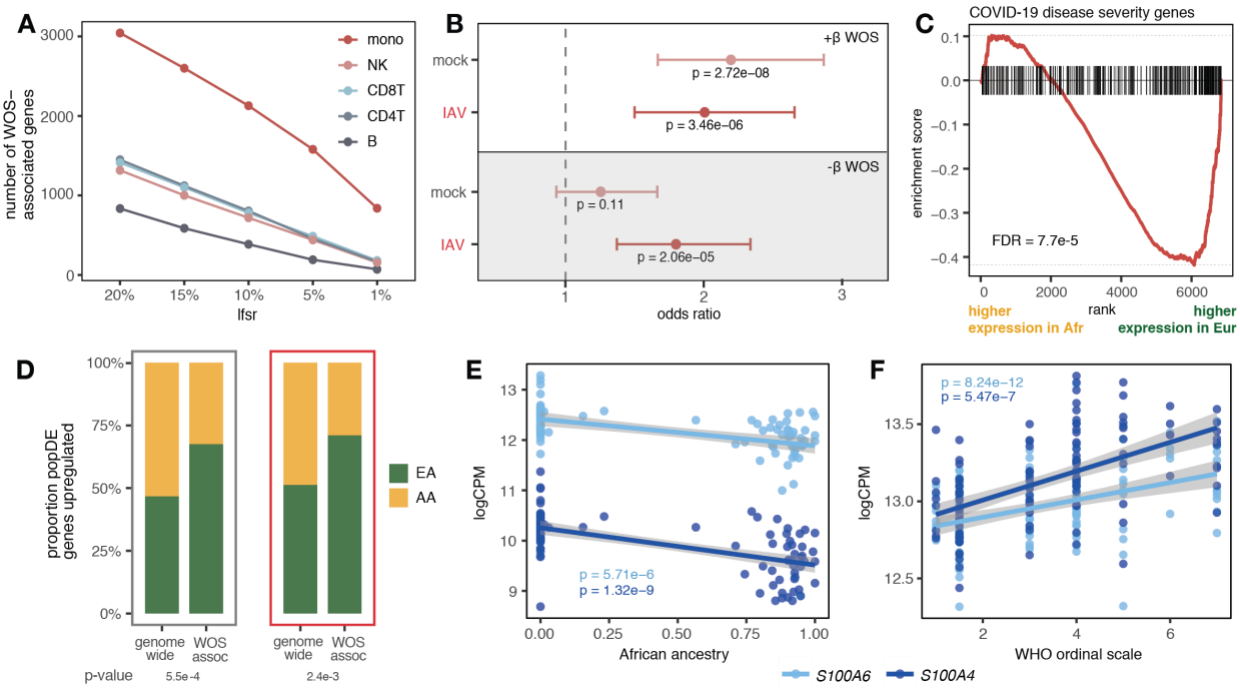


Fig. 3-4. Genes associated with COVID-19 severity display population-associated variation in expression. (A) Number of COVID severity-associated genes by cell type for different significance thresholds (x-axis). (B) Enrichment of popDE genes identified in mock and IAV-infected conditions among genes positively (white) and negatively (gray) associated with severity in monocytes (odds ratio with 95% confidence interval). (C) Enrichment plot for genes positively

Fig. 3-4, continued. associated with COVID severity in monocytes among the IAV-infection popDE effect sizes in monocytes (x-axis). (D) Proportion of genome-wide popDE and severity-associated popDE genes upregulated by individuals with a higher level of European (green) or African (yellow) genetic ancestry in mock (gray) and IAV-infected conditions (red). (E) Correlation between African genetic ancestry proportion and *S100A4*/*S100A6* expression in monocytes after IAV infection. (F) Correlation between WOS and *S100A4*/*S100A6* expression in COVID-19 patients. In (E) and (F), p-values and best-fit slopes were obtained from linear regression models.

Genes where higher expression was associated with COVID-19 severity in monocytes (lfsr < 0.01) were 2.0 to 2.2 times more likely to be identified as popDE genes in our single-cell IAV dataset (lfsr < 0.10) compared to genome-wide expectations (Fisher's exact test and permutations, p-values = 2.7×10^{-8} [mock] and 3.5×10^{-6} [IAV], Figs. 3-4B and S3-5A). These genes also tended to be more highly expressed in monocytes from individuals with more European ancestry (FDRs = 9.8×10^{-5} [mock], 7.7×10^{-5} [IAV], Figs. 3-4C and S3-5B). Consequently, an average of 69% of COVID severity-associated genes across conditions in monocytes showed increased expression with greater European ancestry, a significantly higher proportion than the 49% observed among all popDE genes (Chi-square test, p-values = 5.5×10^{-4} [mock] and 2.4×10^{-3} [IAV], Fig. 3-4D). Finally, we identified several *S100* family genes among those most strongly associated with both genetic ancestry (Fig. 3-4E) and COVID-19 disease severity (Fig. 3-4F). Members of this gene family encode proteins that regulate inflammation and can endogenously activate and amplify inflammatory responses in phagocytes (Xia et al. 2018). *S100A4/A6*/*A8* expression has been associated with patient improvement when upregulated early in the course of COVID-19 infection (Su et al. 2020), and *S100A8*/*A9* are systemically upregulated in immune cells, particularly monocytes, in severe, late-stage COVID-19 patients (Ren et al. 2021). In our data, *S100A4*, *S100A6*, and *S100A8* are all significantly more highly expressed with greater European ancestry early after IAV infection (Fig. 3-4E, Table S3-5), consistent with a potential contribution of genetic

ancestry to observed differences in COVID-19 susceptibility between African Americans and European Americans.

Discussion

Together, our results provide a detailed characterization of the genetic determinants that shape inter-individual and genetic ancestry-associated differences during the early response to viral infection in immune cells. Our findings expand on previous work measuring genetic ancestry effects in isolated cell types (M. N. Lee et al. 2014; Quach et al. 2016) by showing that the majority of ancestry effects on the immune response to IAV are cell type-specific. One clear exception to this overall pattern was genetic ancestry-associated differences in the IFN response. Our analysis reveals that, across all cell types, increased European ancestry is associated with a stronger type I IFN response shortly after influenza infection, which in turn predicts reduced viral titers at later time points. Given the central role played by interferons in conferring antiviral activity to host cells (Ciancanelli et al. 2016), our findings have potential clinical implications not only for influenza infection but also for other viruses, including SARS-CoV-2, for which the timing and magnitude of IFN-mediated antiviral responses are associated with disease progression and severity (J. S. Lee and Shin 2020).

Many of the genetic ancestry-associated differences in immune regulation we observe are driven by allele frequency differences at *cis*-regulatory variants. Among popDE genes in which we identify at least one *cis*-eQTL across cell types and conditions, we estimate that, on average, *cis*-eQTLs explain approximately 53% of the variance in the observed ancestry-associated differences. Our results stress the key role played by genetics in shaping population differences in immune responses, including that these differences are overwhelmingly due to variants found

across populations, but segregating at different frequencies (Quach et al. 2016; Nédélec et al. 2016). We note, however, that for about half of popDE genes, we were not able to identify an eQTL, pointing to additional, co-acting drivers of genetic ancestry-correlated gene expression. These may include other genetic effects (either *cis*-acting effects or *trans*-acting effects we are underpowered to map (Lappalainen et al. 2013; Westra et al. 2013)) or unmeasured environmental factors that are stratified by genetic ancestry.

Viruses have been shown to be among the strongest sources of selection pressure in human evolution (Fumagalli et al. 2011; Enard and Petrov 2018). Among the different forms of natural selection in humans, polygenic selection is thought to be the most pervasive (Pritchard, Pickrell, and Coop 2010), but specific examples of polygenic selection in humans remain rare. Our results provide novel evidence for ancestry-associated directional shifts in molecular traits (i.e., gene expression phenotypes related to specific biological pathways) that are under *cis*-regulatory genetic control, highlighting the potential role of polygenic selection in the history of these phenotypes. The best candidate for polygenic selection was observed for RP genes, in which we consistently found that alleles associated with higher expression are also more prevalent in individuals with more European ancestry. This observation represents one of the few instances of polygenic selection in humans that is supported by functional genomic data. The signature of selection at ribosomal protein genes is particularly interesting in the context of viral infections, as ribosomal proteins facilitate translation initiation of viral transcripts (Huang et al. 2012) and directly interact with viral mRNA and proteins to enable viral protein synthesis (S. Li 2019). Further, a subset of ribosomes, known as immunoribosomes, has been hypothesized to preferentially synthesize antigenically-relevant cellular and viral peptides for immunosurveillance by the MHC class I system, which may allow immune cells to more quickly recognize and

eliminate infected cells (Wei and Yewdell 2019). Together, these observations raise the possibility that polygenic selection on ribosomal pathways, acting heterogeneously on different human populations, has contributed to present-day variation in viral control.

Finally, our results show that genes differentially expressed by genetic ancestry are enriched among genes associated with COVID-19 disease severity. Our findings suggest that immune response variation may therefore interact with or exacerbate environmentally-driven health disparities in viral susceptibility and morbidity, which occur for both influenza and COVID-19 (J. Y. Ko et al. 2021; Chandrasekhar et al. 2017). An important goal for future work is to evaluate whether the variation we observe early in the viral response translates to differences in COVID-19 patient outcomes. Indeed, time course studies (Liu et al. 2021; Bernardes et al. 2020) highlight the importance of temporal dynamics in the immune response to infection, which can include time-dependent reversals of effects. For example, the early upregulation of antiviral and proinflammatory genes shortly after initial infection has been associated with protection but their delayed induction is a hallmark of severe illness (S. Zhou et al. 2021, 1). Our results motivate further studies that investigate whether genetic ancestry-linked effects on innate immunity extend to influence the adaptive immune response as well, and, ultimately, viral clearance and disease severity over the course of viral infections *in vivo*.

Acknowledgements

We thank J. Tung, B. Mittleman, G. Harrison, and members of the Barreiro lab for their constructive comments and feedback. We thank P. Carbonetto and M. Stephens for advice regarding the mash analyses. We thank J. Sanz for guidance with statistics and modeling. We thank J. Ayroles for providing us with the Tn5 transposase used to generate the TM3'seq libraries.

Computational resources were provided by the University of Chicago Research Computing Center. We thank the University of Chicago Cytometry Antibody Technologies Facility (RRID: SCR_017760), particularly D. Leclerc and L. Johnston, for their assistance with the Luminex cytokine assays, and the University of Chicago Genomics Facility (RRID: SCR_019196), especially P. Faber, for their assistance with RNA-sequencing. Figure 3-1A was created with BioRender.com. **Funding:** This work was supported by grant R01-GM134376 to L.B.B. H.E.R was supported by a National Science Foundation Graduate Research Fellowship (DGE-1746045).

Author contributions: L.B.B directed the study. H.E.R. and L.B.B designed the experiments. J.K.F and B.K.T. generated the influenza A Cal/04/09 strain used for infections, and J.K.F. and C.M. performed the plaque assays and baseline antibody titer measurements under the supervision of R.A.L. H.E.R. performed all *in vitro* PBMC infection experiments and sample collections. H.E.R. and M.S.C performed RNA-sequencing library preparations. H.E.R. led the computational analyses, with contributions from J.B.B. H.E.R. and L.B.B. wrote the manuscript, with input from all authors. **Competing interests:** Authors have no competing interests to declare. **Data and materials availability:** Fastq and RNA-sequencing count files are available at GEO under accession GSE162632. Genome sequencing data are available at SRA under accession PRJNA736483. Processed data files, scripts, and associated documentation are available at https://doi.org/10.5281/zenodo.4273999.

Materials and Methods

Peripheral blood mononuclear cell (PBMC) collections. This study has been approved by the Institutional Review Board at the University of Chicago (protocol #: IRB19-0432). All samples were obtained from BioIVT. Signed, written consent was obtained from each participant. Blood was collected from 90 male donors between the ages of 21 – 69, who identified as either African American (AA) (n = 45) or European American (EA) (n = 45), from the same collection site in Miami, Florida (United States) utilizing a standard protocol with a sodium heparin anticoagulant. Briefly, PBMCs were extracted from whole blood using a density gradient, washed with HBSS, reconstituted in CryoStor CS10 (Sigma-Aldrich, C2874) to a concentration of 10 million (M) cells/ml, and subsequently cryopreserved. Between 6 - 10M cells per individual were frozen per vial. In addition, paired serum for each individual was collected and frozen. We decided to only focus on males to avoid the potentially confounding effects of sex-specific transcriptional differences in the response to infection. The sample size (n = 90 biological replicates) was chosen based on prior empirical data (Aguet et al. 2017) showing that this number of individuals provides sufficient power to detect many *cis*-eQTL.

Our sample was restricted to individuals that met the health requirements specified by the US Food and Drug Administration (FDA) and the American Association of Blood Banks (AABB) at the time of collection (based on the "Full-Length Donor History Questionnaire" (DHQ), https://www.aabb.org/docs/default-source/default-document-library/resources/dhq-v2-1/pdfs/dhq-v2-1-prep-pep-art.pdf). BioIVT's standard donor screening practices also prevent biospecimen collection from donors who exhibit flu-like symptoms, fail a temperature check, or are exposed to a COVID-19-infected individual two weeks prior to collection. In addition, all individuals that reported taking a medication on the Medication Deferral List

(https://www.aabb.org/docs/default-source/default-document-library/resources/dhq-v2-1/pdfs/dhq-medication-deferral-list-v2-1-prep-pep-art.pdf) were excluded from blood donation. Finally, following blood collection, blood-borne pathogen testing was performed in accordance with FDA regulations. Only donors who were negative for human immunodeficiency virus (HIV)-1/2 antibodies and hepatitis C virus (HCV) antibodies and non-reactive for hepatitis B virus (HBV) antigen, HBV DNA, HIV-1 RNA, HCV RNA, West Nile virus (WNV) RNA, anti-*Trypanosoma cruzi* antibodies, Zika virus RNA, and anti-*Treponema pallidum* antibodies (serological test for syphilis, STS) were retained in the study.

Individuals in this cohort do show a modest correlation between age and genetic ancestry: European American (EA) individuals are slightly older than African American (AA) individuals in the sample (median age EA = 45, AA = 39, t-test p-value = 0.003). However, the age ranges mostly overlap. Additionally, we always take the effect of age on gene expression levels or the response to IAV into account in our statistical models. Notably, if we remove the small set of EA individuals who are over 55 (n = 6), there is no significant difference in age between EA and AA individuals (p = 0.06). Importantly, popDE effect size estimates are highly consistent whether these individuals are included or excluded from the sample (average adj. R^2 across cell types and conditions = 0.987).

Quantification of baseline Cal/04/09-specific serum IgG antibody levels. The amount of circulating human serum antibodies in the media during the *in vitro* IAV infection experiments is expected to be negligible since PBMCs are washed multiple times prior to culture. Regardless, we measured the baseline levels of Cal/04/09-specific serum IgG antibodies for all individuals in our cohort, as the best available proxy for past immunization in the absence of detailed vaccination or

infection history records for our donors. 96 well plates were coated with a 1:25 dilution of UV-killed Cal/04/09 diluted in PBS. All antigen coated plates were blocked with 1% BSA in PBS prior to addition of serum. Serial dilutions of human sera (4-fold serial dilutions, a total of eight dilutions per sample) were added to coated and blocked plates, and bound immunoglobulin (Ig) was detected with HRP-anti-human IgG antibody (Southern Biotech, 2040-05) followed by ABTS Peroxidase Substrate (SeraCare, 5120-0043). OD_{405} was detected by a Synergy H1 plate reader (BioTek). Area under the curve (AUC) was calculated, and linear regressions were performed to determine associations with various variables (Figs. S3-3C-D). All individuals had detectable levels of Cal/04/09-specific serum titers, but no differences in antibody titers were identified between European and African-ancestry individuals (Fig. S3-3C). Baseline titers were not correlated with IFN response (measured via the IFN score, Fig. S3-3D), nor did they influence gene expression levels or the response to IAV ("Effect of baseline serum titers on gene expression", below).

Generation of virus and viral titers. Influenza A virus California/04/2009 (Cal/04/09) virus was rescued in 293T cells (ATCC, CRL-3216) by plasmid-based transfection with IAV Cal/04/09 in the pDZ vector using methods previously described (Fodor et al. 1999; Hoffmann et al. 2000; Hai et al. 2010). 24 hours following transfection, $7.5x10^5$ MDCK cells (ATCC, CCL-34, NBL-2) were added to the culture in Opti-MEM (ThermoFisher Scientific, 31985062) containing TPCK trypsin (1 μg/ml). For the following two days, 500 μl of Opti-MEM containing TPCK trypsin (2 μg/ml) was added to the culture. One day later, the supernatant was harvested, centrifuged to remove cellular debris, and stored at -80°C. Cal/04/09 was amplified on MDCK cells to generate a stock. Uninfected MDCK cells were cultured for 48 - 72 hours and supernatant was harvested to generate the control, mock-conditioned media. Stocks and Cal/04/09-infected PBMC supernatants

(collected as described below in "*In vitro* PBMC infections and sample collections") were plaqued on MDCK cells. PBMC supernatants were treated with TPCK for 30 minutes prior to plaquing. MDCKs were infected in infection media (PBS with 10% Ca/Mg, 1% penicillin/streptomycin, 5% BSA) at 37°C for 1 hour. Infection media was replaced with an agar overlay (2X MEM, 1 μg/ml TPCK trypsin, 1% DEAE-dextran, 5% $NaCo_3$, 2% oxoid agar), and cells were cultured at 37°C for 40 hours then fixed with 4% formaldehyde. Blocking and immunostaining were done for 1 hour at 25°C in 5% milk. Primary stain was mouse anti-Cal/04/09 (1:5000), secondary stain was peroxidase sheep anti-mouse-HRP (1:5000, GE Healthcare, 45001275). The polyclonal mouse anti-Cal/04/09 was generated in-house (mice were infected with 5,000 PFU of Cal/04/09, and serum was harvested 30 days post infection). TrueBlue Peroxidase Substrate (Kirkegard & Perry Laboratories, 50-647-28) was used as directed for detection of virus plaques.

Viral titer measurements were corrected for time course experiment batch (detailed below, "*In vitro* PBMC infections and sample collections") within time point by fitting a model to estimate the effect of experiment batch on titers (titers ~ batch) using the lm function in R, taking the residuals of this model, and adding back the model intercept. Individuals with negative corrected viral titer measurements (n = 1 for 24 hpi, n = 2 for 48 hpi) were manually set to zero. These batch-corrected values are reported in Fig. 3-2G. A nonparametric 2-group Mann-Whitney U test was used to assess significance between group medians because of the relatively small sample size (n = 20 total).

***In vitro* PBMC infections and sample collections.** PBMCs were unfrozen approximately 14 hours prior to infection and cultured overnight in RPMI 1640 supplemented with 10% fetal bovine serum (Corning, MT35015CV), 2 mM L-glutamine (ThermoFisher Scientific, 25-030-081), and

10 ug/ml gentamicin (ThermoFisher Scientific, 15710064). Infection experiments were performed over 15 batches, where each experimental batch was balanced for self-identified ancestry label to avoid introducing a batch effect confounded with genetic ancestry. The morning of the experiment, 1M PBMCs were plated at a concentration of 1M/ml for each condition, and exposed to either mock-conditioned media (negative control) or Cal/04/09 IAV at a multiplicity of infection (MOI) of 0.5. After 30 minutes, the control media or virus was washed from PBMC cultures. Cells were then replated and incubated for 6 hours at 37°C in 5% CO_2 and 20% O_2. Following the 6 hour incubation, approximately 0.5M cells per sample were collected, washed, and prepared for single-cell capture using the 10X workflow, and approximately 0.5M cells per sample were collected for paired bulk RNA-sequencing. All bulk RNA samples were stored in Qiazol (Qiagen, 79306) at -80°C prior to RNA extraction. Immediately prior to single-cell capture, cells from different samples were combined into two pools (6 samples per pool), each balanced for infection status (mock and IAV-infected) and genetic ancestry (Table S3-1). Multiplexed cell pools were used as input for the single-cell captures. For each cell pool, 10,000 cells were targeted for collection using the Chromium Single Cell 3' Reagent (v2 chemistry) kit (10X Genomics, 120234). Post Gel Bead-in-Emulsion (GEM) generation, the reverse transcription (RT) reaction was performed in a thermal cycler as described (53°C for 45 min, 85°C for 5 min), and post-RT products were stored at -20°C until downstream processing (no longer than 4 days post-RT reaction). For DNA processing, 1M PBMCs were collected, and DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen, 69504) following the "Cultured cells" protocol.

For the time course IAV infection experiment, PBMCs from 10 individuals with the highest and 10 individuals with the lowest IFN responses were unfrozen and cultured as described above in two batches (first batch n = 10, second batch n = 10, each batch balanced for equal numbers of

high [n = 5] vs low [n = 5] responders). Low and high responder groups stratified with respect to expected genetic ancestry. The low responder group included mostly majority-African ancestry individuals (n = 8 African, n = 2 European), and the high responder group contained mostly majority-European ancestry individuals (n = 6 European, n = 4 African) (Table S3-1).

For each individual, 0.5M cells were exposed to mock-conditioned media, and between 3 - 4M cells (depending on the initial number of cells recovered) were infected with Cal/04/09 IAV at an MOI of 0.5. After 30 minutes, the control media or virus was washed from PBMC cultures, and 0.5M cells were replated at a concentration of 1M/ml for each of the following time points: mock, 2 hours, 6 hours, 12 hours, 18 hours, 24 hours, 36 hours, and 48 hours. PBMCs from all 20 individuals were plated for all time points except at 18 hours (3 missing) because a sufficient number of PBMCs were not recovered to plate this time point across individuals. Cells were incubated at 37°C in 5% CO_2 and 20% O_2 until they were collected for downstream processing. At each time point, bulk RNA from PBMCs and culture supernatant were collected for all individuals plated. All bulk RNA samples were stored in Qiazol (Qiagen, 79306) at -80°C prior to RNA extraction. All supernatant samples were stored at -80°C prior to cytokine level and viral titer measurements.

Luminex cytokine assays. Secreted cytokine measurements were assessed using the Bio-Plex Pro Human Inflammation Panel, 37-Plex kit (Bio-Rad, 171AL001M) according to the manufacturer's instructions. Undiluted supernatant samples were used as input, and analyte measurements were detected using the Luminex 200 System. Cytokine measurements were corrected for time course experiment batch by fitting a model to estimate the effect of experiment batch on cytokine levels (cytokine ~ batch) with the lm function in R, taking the residuals of this model, and adding back

the model intercept within time point. These batch-corrected values are reported in Fig. 3-2F. A nonparametric 2-group Mann-Whitney U test was used to assess significance between group medians because of the relatively small sample size (n = 10 per group).

Single-cell RNA-sequencing library preparation and sequencing. Post-RT reaction cleanup, cDNA amplification, and sequencing library preparation were performed as described in the Single Cell 3' Reagent Kits v2 User Guide (10X Genomics). Briefly, cDNA was cleaned with DynaBeads MyOne SILANE beads (ThermoFisher Scientific, NC0949127) and amplified in a thermal cycler using the following program: 98°C for 3 min, [98°C for 15 s, 67°C for 20 s, 72°C for 1 min] x 11 cycles, 72°C 1 min. After cleanup with the SPRIselect reagent kit (Beckman Coulter, B23317), the libraries were constructed by performing the following steps: fragmentation, end-repair, A-tailing, SPRIselect cleanup, adaptor ligation, SPRIselect cleanup, sample index PCR (98°C for 45 s, [98°C for 20 s, 54°C for 30 s, 72°C for 20 s] x 14 cycles, 72°C 1 min), and SPRIselect size selection. Batches of four experiments (corresponding to eight multiplexed single-cell captures) were processed at a time. Prior to sequencing, all multiplexed single-cell libraries (n = 30) were quantified using the KAPA Library Quantification Kit for Illumina Platforms (Roche, 501965234) and pooled in an equimolar ratio. Libraries were sequenced 100 base pair paired-end (R1: 30 cycles, I1: 10 cycles, R2: 85 cycles) on an Illumina NovaSeq 6000 to an average depth of 45,612 mean reads per cell across all batches (average median genes detected per cell across batches = 689).

Bulk RNA-sequencing library preparation and sequencing. Total RNA was extracted using the miRNeasy Micro Kit (Qiagen, 217084) according to the manufacturer's instructions. RNA

libraries were generated using TM3'seq, a 3'-enriched library preparation protocol that generates transcriptomes of the same quality as those generated using standard RNA-seq approaches (Pallares, Picard, and Ayroles 2020). RNA (50 ng) was added to 1 µl of 0.83 uM oligo (Tn5ME-B-30T) and incubated for 3 minutes at 65°C. Following the incubation, the mixture was combined with 1 µl SMARTScribe™ RT (Takara, 639538), 1 µl dNTPs 10mM (NEB, N0447S), 2 µl DTT 0.1M (Takara, 639538), 4 µl 5X First-Strand buffer (Takara, 639538), and 1 µl B-tag-sw oligo and incubated for 1 hour at 42°C for first strand cDNA synthesis and 70°C for reverse transcriptase inactivation. A cDNA amplification mix was prepared by combining 7.5 µl of first strand cDNA product with 7.5 µl of OneTaq HS Quick-load 2X (NEB, M0486L). This reaction was amplified in a thermal cycler set to the following program: 68°C 3 min, 95°C 30 s, [95°C 10 s, 55°C 30 s, 68°C 3 min] x 3 cycles, 68°C 5 min.

Tn5 transposase (gifted from Julien Ayroles' lab, Princeton University) was combined with pre-annealed adaptor A (10 µM), a forward oligo, at a ratio of 11 to 1 and incubated at 37°C for 30 minutes. The Tn5 mix was then diluted at a ratio of 1 to 5 in a solution of 50% glycerol and 50% reassociation buffer (10 mM Tris pH 8.0, 50 mM NaCl, 1 mM EDTA). This mixture was then diluted 1 to 4 in 5X TAPS buffer (50 mM TAPS, 25 mM MgCl2, 50% v/v DMF). The cDNA was tagmented by combining 4 uL of the diluted Tn5 in TAPS buffer with 11 µl of the second strand cDNA and incubated at 55°C for 7 minutes. Tn5 was dissociated from the cDNA by adding 3.5 µL of 0.2% SDS followed by a 7 minute incubation at 55°C. The cDNA was cleaned using Agencourt AMPure XP beads (Beckman Coulter, A63881), and 8 µl of cDNA was combined with 1 µl of i5 primer at 1 µM, 1 µl of i7 primer at 1 µM, and 10 µl of OneTaq HS Quick-load 2X (NEB, M0486L) prior to amplification. Libraries were amplified in a thermal cycler set to the following program: 68°C 3 min, 95°C 30 s, [95°C 10 s, 55°C 30 s, 68°C 3 min] x 18 cycles, 68°C

5 min. PCR products were pooled and cleaned using Agencourt AMPure XP beads (Beckman Coulter, A63881). Library eluate was stored at -20°C until sequencing. All bulk RNA-sequencing libraries were pooled in an equimolar ratio, and this library pool was sequenced 100 base pair single-end on an Illumina NovaSeq 6000 to an average depth of 7.5 million reads per sample.

Low-pass whole genome DNA sequencing and VCF processing. Out of the 90 individuals in the cohort, 89 were successfully genotyped using DNBseq low-pass whole genome sequencing (BGI) at 4X coverage. Variants were called across individuals using the human reference genome (GRCh37), yielding a merged VCF. The ImputeSeq low-pass imputation pipeline (Gencove) was used to perform VCF imputation. The imputed merged VCF was lifted over to GRCh38 with CrossMap (v0.3.9) (Zhao et al. 2014) using the GRCh37 to GRCh38 Ensembl chain file downloaded at ftp://ftp.ensembl.org/pub/assembly_mapping/homo_sapiens/ and the GRCh38 FASTA file from ftp://ftp.ensembl.org/pub/release-92/fasta/homo_sapiens/dna/. For each individual, low-quality variants were filtered by retaining those with a maximum genotype probability (GP in FORMAT field) > 0.90 using QCTOOL (v2.0.7, https://www.well.ox.ac.uk/~gav/qctool_v2/). If the max(GP) for a variant was < 0.90, the variant call was automatically set to missing. Only autosomal, biallelic SNPs were kept for downstream analysis using the SelectVariants function (--select-type-to-include SNP) from the Genome Analysis Toolkit (GATK) (v3.7) (DePristo et al. 2011).

Whole exome sequencing. DNBseq whole exome sequencing (BGI, average 50X coverage) was performed on the 20 individuals included in the high versus low responder time course experiment. To construct the exome sequencing libraries, genomic DNA was randomly fragmented by

sonication (average fragment size between 150 - 250 bp). DNA fragments were end repaired, A-tailed, and ligated with adapters prior to amplification. Size-selected DNA fragments were amplified by ligation-mediated PCR, purified, and hybridized to the exome array for enrichment. Non-hybridized fragments were washed out. Captured products were circularized, and rolling circle amplification was performed to produce DNA nanoballs, which were sequenced on the DNBseq platform.

Low-quality raw reads were removed if: 1) reads contained a sequence adaptor, 2) reads contained a low-quality base ratio (base quality <= 5) greater than 50%, and 3) reads contained an unknown base ("N") ratio greater than 10%. Filtered, high-quality sequencing reads were mapped to the human reference genome (GRCh37) using Burrows-Wheeler Aligner (BWA-MEM v0.7.17) (H. Li and Durbin 2009). Picard tools (v2.5.0) (http://broadinstitute.github.io/picard/) was used to sort the SAM files by coordinate and convert SAM files to BAM files. Duplicate reads were marked using the MarkDuplicates function in Picard tools (v2.5.0). To obtain more accurate base qualities, Base Quality Score Recalibration (BQSR) was performed using the BaseRecalibrator and ApplyBQSR functions from GATK (v4.1.4) (DePristo et al. 2011). Variants were called using GATK's Best Practices for variant analysis pipeline. Briefly, SNPs were detected using the HaplotypeCaller function (GATK, v4.1.4) and selected using the SelectVariants function (GATK, v4.1.4) (DePristo et al. 2011). To obtain a high-confidence call set, hard-filtering was applied via the VariantFiltration function (--filter-expression "QD<2.0 || FS>60 || MQ<40 || MQRankSum<-12.5 || ReadPosRankSum<-8.0") (GATK, v4.1.4) (DePristo et al. 2011). Finally, SnpEff (Cingolani et al. 2012) was used to perform various annotations, including gene-based annotation (e.g., to identify whether SNPs cause protein coding changes and determine which amino acids

are affected) and the addition of pathogenicity prediction scores (e.g., SIFT, Polyphen-2, and CADD).

To evaluate the contribution of rare genetic variants in the type I IFN pathway, we focused our attention on coding variants within genes previously associated with defects in the type I interferon pathway (Zhang 2020; Zhang et al. 2020), including four genes reported to be mutated in patients with life-threatening influenza (*GATA2*, *IRF7*, *IRF9*, *TLR3*), six genes in the TLR3-dependent type I IFN induction pathway (*TICAM1*, *IKBKG*, *UNC93B1*, *TRAF3*, *TBK1*, *IRF3*), and four genes in the IRF7- and IRF9-dependent type I IFN amplification pathway (*IFNAR1*, *IFNAR2*, *STAT1*, *STAT2*).

Of these 14 genes, 13 harbored variants that fall within coding regions (n = 51 coding variants total, all of which have previously been reported in dbSNP). Of these 51 variants, 22 were synonymous mutations that are likely non-pathogenic, 27 were missense variants, 1 was a splice donor variant (*IRF3*), and 1 was a nonsense mutation (*TICAM1*). The *IRF3* splice donor variant is found at high frequency across populations (1000 Genomes Project [1000GP] global allele frequency = 0.37), suggesting that it is not deleterious. The stop-gain mutation in *TICAM1* (c.421C>T, rs387907307) is present in one individual in our cohort (a "high responder") in the heterozygous state. This mutation has previously been associated with herpes simplex encephalitis in children with TRIF deficiency (Sancho-Shimizu et al. 2011); however, the specific allele harbored by the individual in our cohort is the variant responsible for the autosomal recessive form of TRIF deficiency (Sancho-Shimizu et al. 2011), consistent with observed, normal TRIF-dependent signaling in this individual.

Among the missense variants we detected, 25 of 27 are predicted to be benign/tolerated by CADD, SIFT, and/or PolyPhen-2, and/or are common variants in 1000GP (global allele frequency

> 0.10). Two rare variants (rs145480303 in *IRF9* [ExAC allele frequency < 0.001] and rs369756552 in *TICAM1* [1000GP global allele frequency < 0.01]) are predicted to be pathogenic by both PolyPhen-2 and SIFT. However, rs145480303 is reported as a benign variant in ClinVar, and rs369756552 is present in the heterozygous state in one "high responder," suggesting that it does not lead to loss of function. Together, the exome sequencing results suggest that rare, loss-of-function mutations do not explain the large differences in IFN response and viral control observed between high versus low responders.

HLA typing. HLA alleles were imputed from whole-genome sequencing data for each individual using the multi-ethnic HLA imputation panel available through the Michigan Imputation Server. A series of linear models were used to evaluate whether common HLA variants (minor allele frequency $[MAF]_{\text{both populations}} > 0.05$, n = 88) were associated with either the IFN response or IAV transcript levels across individuals for each cell type (e.g., IFN response ~ HLA genotype dosage). Models were fit using the big_univLinReg function from the R package bigsnpr (Privé et al. 2018). P-values were corrected for multiple testing using the Benjamini-Hochberg procedure (p.adjust function in R, method = "BH"), and genotypes were considered significantly associated with response variables if they had an FDR < 0.05. One variant, HLA-DPB1*03:01, showed a significant association (FDR = 0.004) with IAV transcript levels in $CD4^+$T cells, and no variants were associated with IFN response. The potential consequences of this association in our viral infection setting are not clear, especially given that $CD4^+$T cells do not perform MHC class II presentation. However, the HLA-DPB1*03:01 genotype explains only 15.1% of the variance in IAV transcript levels in $CD4^+$T cells, and the association between genetic ancestry and IAV transcript levels in PBMCs remains significant when this genotype is included in the model (p =

0.007), suggesting that this allele does not substantially contribute to genetic ancestry-associated variation in our data.

Estimation of genome-wide admixture levels. Prior to estimation of genome-wide admixture proportions, samples were merged with CEU (n = 99, Utah Residents [CEPH] with Northern and Western European Ancestry) and YRI (n = 108, Yoruba in Ibadan, Nigeria) samples from the 1000 Genomes Project (1000GP) Phase 3 dataset (Auton et al. 2015) (downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/). The proportion of European and African genetic ancestry for each individual was estimated using the supervised clustering algorithm in ADMIXTURE (v1.3.0) (Alexander and Lange 2011). A total of 13,518,147 unlinked SNPs (r^2 between all pairs < 0.1) were used for genetic ancestry assignments, assuming k = 2 ancestral clusters. Self-identified African American (AA, n = 45) individuals had a modest, although highly variable, percentage of European ancestry (mean = 11%, range = 0 - 43%), while self-identified European American (EA, n = 44) individuals displayed more limited levels of African ancestry (mean = 1%, range = 0 - 23%) (Fig. S3-1A, Table S3-1). These estimated quantitative genetic ancestry proportions were used to assess differences in immune responses between populations.

Bulk RNA-sequencing data processing. Adapter sequences and low-quality score bases were trimmed from reads using Trim Galore (v0.6.2, Cutadapt v2.2) (M. Martin 2011) in single-end mode (-q 20 --paired --phred33). Trimmed reads were pseudoaligned to a custom transcriptome containing both the *Homo sapiens* reference transcriptome (GRCh38) and the Cal/04/09 transcriptome (downloaded from Ensembl) using the quant function in kallisto (v0.43) (Bray et al.

2016) (average depth of 4.3 million pseudoaligned reads per sample for the true bulk RNA-seq samples). Gene-level expression estimates were calculated using the R (v3.6.3) package tximport (v1.14.2) (Soneson, Love, and Robinson 2016). Expression data was filtered for protein-coding genes that were sufficiently expressed across all samples (median logCPM > 1). After removing non-coding and lowly-expressed genes, normalization factors to scale the raw library sizes were calculated using calcNormFactors in edgeR (v3.26.8) (Robinson, McCarthy, and Smyth 2010). The voom function in limma (v3.40.6) (Ritchie et al. 2015) was used to apply these size factors, estimate the mean-variance relationship, and convert counts to logCPM values. Technical effects (e.g., library preparation batch) were regressed using the ComBat function in sva (v3.32.1) (https://bioconductor.org/packages/sva/). A model evaluating the technical effect of experiment batch (~ 0 + batch, where batch corresponds to a factor variable representing the 15 experimental batches) on gene expression was fit using the lmFit and eBayes functions, and model residuals were obtained using the residuals.MArrayLM function in limma (Ritchie et al. 2015). The average experiment batch effect was then computed by taking the mean of the capture coefficients across all 15 batches per gene, and this average effect was added back to the residuals.

Single-cell RNA-sequencing mapping, demultiplexing, and initial cell filtering. FASTQ files from each multiplexed capture library were mapped to a custom reference containing GRCh38 and the Cal/04/09 IAV reference genome (downloaded from NCBI, created using cellranger mkref) using the cellranger (v3.0.2) (10X Genomics) count function (G. X. Y. Zheng et al. 2017). souporcell (v2.0, Singularity v3.4.0) (Heaton et al. 2020) in --skip_remap mode (-k 6) was used to demultiplex cells into samples based on genotypes from a common variants file (1000GP samples filtered to SNPs with >= 2% allele frequency in the population, downloaded from

https://github.com/wheaton5/souporcell). Briefly, souporcell clusters cells based on cell allele counts in common variants, assigning all cells with similar allele counts to a single cluster corresponding to one individual, while also estimating singlet/doublet/negative status for that cell. For each batch, hierarchical clustering of the true genotypes known for each individual (obtained from low-pass whole-genome-sequencing) and the cluster genotypes estimated from souporcell was used to assign individual IDs to souporcell cell clusters. All 89 individuals were successfully assigned to a single cluster.

After demultiplexing cells into samples, Seurat (v3.1.5, R v3.6.3) (Stuart et al. 2019) was used to perform quality control filtering of cells. In total, we captured 255,731 cells prior to filtering (range of cells recovered per capture: min. = 5,534, max. = 10,805). Cells were considered "high-quality" and retained for downstream analysis if they had: 1) a "singlet" status called by souporcell, 2) between 200 – 2500 genes detected (nFeature_RNA), and 3) a mitochondrial reads percentage < 10%, leaving 236,993 cells (n = 19,248 genes).

Clustering, cell type assignment, and UMAP analysis. We performed two versions of clustering analysis and cell type assignment: 1) in which IAV genes were kept in the raw count matrix (used as input for pseudobulk calculations), and 2) in which IAV genes were subset out of the raw count matrix (for visualization of the UMAP in Fig. 3-1B). All other steps of the clustering workflow (implemented in Seurat v3.1.5) remained the same. Pseudobulk expression estimates (see below) between clustering versions for cell type-matched clusters were extremely similar (adj. $R^2 > 0.999$ for comparisons between versions). For both clustering iterations, we split the cells by infection status (mock or IAV) and ran SCTransform (Hafemeister and Satija 2019) to normalize and scale the UMI counts within condition. In this step, we simultaneously regressed out variables

corresponding to experiment batch and percent mitochondrial reads per cell. The data was then integrated on infection status using the SelectIntegrationFeatures, PrepSCTIntegration, FindIntegrationAnchors, and IntegrateData workflow (Stuart et al. 2019). Following integration, dimensionality reduction was performed via UMAP (RunUMAP function, dims = 1:30) and PCA (RunPCA function, npcs = 30). A Shared Nearest Neighbor (SNN) Graph was constructed using the FindNeighbors function (dims = 1:20, all other parameters set to default), and clusters were subsequently called using the FindClusters algorithm (resolution = 0.5, all other parameters set to default) (Stuart et al. 2019).

Clusters were annotated based on the expression of canonical immune cell marker genes (CD4$^+$ T: *CD3D*$^+$, *CD3E*$^+$, *CD8A*$^-$; CD8$^+$ T: *CD3D*$^+$, *CD8A*$^+$; NK cells: *CD3D*$^-$, *NKG7*$^+$, *GNLY*$^+$; monocytes: *CD14*$^+$, *LYZ*$^+$; B: *MS4A1*$^+$; granulocytes: *PRSS57*$^+$; dendritic cells (DCs): *HLA-DRA*$^+$, *HLA-DRB1*$^+$, *CCR7*$^+$, *CST3*$^+$, *CD83*$^+$). A small group of cells, which were identified as B cells, clustered with CD4$^+$ T cells in the UMAP (Fig. 3-1B), and we investigated this further to see whether this subset represented a distinct, rare cell type. Further analysis revealed that these cells express markers typical of NKT cells, including *CD3D*, *NKG7*, *IL2*, *TNF*, and *IFNG*, and thus, these cells were manually annotated as NKT cells. In the UMAP constructed from input data containing IAV genes, we excluded 1,832 cells for which we could not confidently assign a cell type, as they clustered on the basis of high IAV transcript expression, leaving us with 235,161 cells across all individuals and conditions for downstream analysis (n CD4$^+$ T cells = 138,801, CD8$^+$ T cells = 32,446, monocytes = 27,020, B cells = 22,877, NK cells = 13,220, DCs = 374, granulocytes = 301, NKT cells = 122).

Our main analyses focus on the five most common cell types found in PBMCs, including CD4$^+$T cells, CD8$^+$T cells, B cells, monocytes, and NK cells. Because we focus on these five major

cell types, this does not allow us to make conclusions about other cell types that are important in the antiviral response, such as plasmacytoid dendritic cells (pDCs), which are key producers of type I IFNs, and neutrophils. We excluded pDCs and neutrophils from our analyses because, across all individuals and conditions, we only identified 374 pDCs and 301 granulocytes. Because these cell types are found at very low numbers within each individual (range = 1 - 17, median = 4 pDCs; range = 1 - 18, median = 4 granulocytes), we do not have the power to detect infection or genetic ancestry effects in these cell types.

Calculation of pseudobulk estimates. Cluster-specific pseudobulk estimates were used to summarize single-cell expression values into bulk-like expression estimates within samples (where, here, a sample is an individual/infection status pair, n = 180). This was performed for all five major cell types ($CD4^+$ T cells, $CD8^+$ T cells, B cells, monocytes, NK cells) and PBMCs, where all high-quality cells from all cell types identified (n = 235,161) were treated as a single aggregate cluster. Within each cluster for each sample, raw UMI counts were summed across all cells assigned to that sample for each gene using the sparse_Sums function in textTinyR (v1.1.3) (https://cran.r-project.org/web/packages/textTinyR/textTinyR.pdf), yielding an n x m expression matrix, where n is the number of samples included in the study (n = 180) and m is the number of genes detected in the single-cell analysis (m = 19,248) for each of the 6 clusters.

Calculation of capture-corrected expression for downstream modeling. From this point forward, pseudobulk estimates were treated as de facto bulk expression data for each cell type considered. As such, calculations of residuals and downstream modeling of infection and genetic ancestry effects (see below) were performed for each cluster independently. For each cell type,

lowly-expressed genes were filtered using cell-type specific cutoffs (removed genes with a median logCPM < 1.5 in $CD4^+$ T cells, monocytes, and PBMCs, < 2.5 in B cells and $CD8^+$ T cells, and < 4.0 in NK cells), leaving the following number of genes per cell type: $CD4^+$ T cells = 9,960, $CD8^+$ T cells = 9,335, B cells = 9,291, monocytes = 10,424, NK cells = 7,109, and PBMCs = 10,430. Cell type-specific logCPM cutoffs were used because of the inherent differences in median logCPM distributions for all genes detected in the single-cell data (n = 19,248) across cell types, which likely reflect variation in the total number of transcripts captured per gene per cell type as well as the number of cells used to calculate the pseudobulk sum estimates per cell type.

After removing lowly-expressed genes, normalization factors to scale the raw library sizes were calculated using calcNormFactors in edgeR (v 3.26.8) (Robinson, McCarthy, and Smyth 2010). The voom function in limma (v3.40.6) (Ritchie et al. 2015) was used to apply these size factors, estimate the mean-variance relationship, and convert raw pseudocounts to logCPM values. A model evaluating the technical effect of capture (~ 0 + capture, where capture corresponds to a factor variable representing the 30 experimental capture batches) on gene expression was fit using the lmFit and eBayes functions, and model residuals were obtained using the residuals.MArrayLM function in limma (Ritchie et al. 2015). The average capture effect was then computed by taking the mean of the capture coefficients across all 30 capture batches per gene, and this average capture effect was added back to the residuals across samples to generate the capture-corrected expression estimates. We corrected for capture because we viewed it as a potentially important source of batch effects in our data. Indeed, although we were careful to balance each capture for infection status (mock and IAV-infected) and genetic ancestry, we could not process all samples in the same day. Therefore, our correction for capture is much like conventional correction for sequencing flow cell, lane, sampling effort, or extraction batch effects in other types of functional genomic data.

The inverse variance weights calculated by voom were obtained and included in the respective lmFit call for all downstream models unless otherwise noted (Ritchie et al. 2015).

While performing quality control checks on our data, we noticed that the density distributions of the capture-corrected expression estimates were bimodal for some samples in certain cell types. We believe this to be a technical artifact associated with the structure of the single-cell data itself, such that the fewer the number of cells used to create the pseudobulk estimate, the greater the bimodal proportion tends to be (Fig. S3-6A). This correlation is, however, not linear. Once we reach a large enough number of cells summed to estimate pseudobulk expression values, most distributions become unimodal. Using a linear model including both bimodal proportion and cell counts as covariates, we found that, on average across cell types and conditions, the bimodal proportion explains 5.03% of the total variance in gene expression estimates across genes, as compared to only 1.19% for cell counts, suggesting that the bimodal proportion more accurately captures this technical artifact than cell counts themselves. Importantly, our popDE effect size estimates are highly consistent whether we use bimodal proportion or cell count in the model (adj. R^2 range across cell types and conditions = 0.877 – 0.991). Since this technical effect could contribute to noise or bias in our estimates, we decided to correct for it by including the appropriate cell-type specific bimodal proportion vector across samples as a quantitative technical covariate in our downstream models. Of note, our popDE effect sizes are highly consistent whether we include or exclude bimodal proportion from the model (adj. R^2 range = 0.845 – 0.999).

The bimodal proportion in each cell type for each sample was calculated by: i) estimating the local minimum of the density distribution, ii) subsetting the x-axis on a restricted range that was specific to each cell type, iii) using the x value where y equals the estimated local minimum

as the bimodal threshold, and iv) calculating the proportion of genes less than this threshold. Kernel density estimation was performed using the density() function from the base R package stats with default bandwidth and kernel parameters (bw = "nrd0", kernel = "gaussian"). Per-sample bandwidths across cell types are reported in Table S3-1. Following estimation of the bimodal proportion via the density function, all expression density distributions were visually inspected to ensure that, if a distribution displayed bimodality, a non-zero bimodal proportion was returned and that it was not substantially over or underestimated. If an expression density distribution was indeed unimodal, then the bimodal proportion was set to zero.

Modeling global infection effects. To obtain estimates of the global infection effects, capture-corrected expression levels of samples corresponding to the same individual were compared in a paired design by introducing individuals as additional covariates to the following differential infection effect model that was run per cell type:

$$\text{M}_1\text{: } E(i,j) \sim \begin{cases} \beta_0(i,j) + \beta_{pB}(i) \cdot pB^{mock}(j) + \varepsilon^{mock}(i,j) & \text{if Condition} = mock \\ \beta_0(i,j) + \beta_{pB}(i) \cdot pB^{IAV}(j) + (\beta_{IAV}(i) + \beta_{age}^{IAV}(i) \cdot age(j)) + \varepsilon^{IAV}(i,j) & \text{if Condition} = IAV \end{cases}$$

Here, $E(i,j)$ represents the capture-corrected expression estimate of gene *i* for individual *j* and $\beta_0(i,j)$ represents the intercept corresponding to gene *i* and individual *j* (i.e., the expectation of gene *i*'s expression level in the mock-infected sample for individual *j*). When evaluated, this model gives the global estimate of the IAV infection effect per gene, $\beta_{IAV}(i)$, approximated using the within-individual variation in gene expression across conditions, controlled for the effects of age on the response to IAV (captured by the $\beta_{age}^{IAV}(i)$ term, where age represents the mean-centered, scaled [mean = 0, sd = 1] age per individual). Further, pB^{cdt} represents the bimodal proportion estimated per sample for the respective cell type being modeled (where *cdt* represents either the

mock or IAV), with $\beta_{pB}(i)$ being the corresponding effect on gene expression. Finally, ε^{cdt} represents the residuals for each respective condition (mock or IAV) for each gene *i*, individual *j* pair.

We did not perform quantile normalization (per gene, across individuals) on the gene expression estimates prior to fitting M_1 (unlike our approach for models M_2/M_3, "Modeling genetic ancestry effects and integration with mashr", below) because, for the large shifts in expression expected in response to infection, quantile normalization compromises the interpretability of the calculated fold changes (Fig. S3-6B). After IAV infection, we expect that many genes will have completely non-overlapping gene expression distributions between conditions but will also show variation in the strength of upregulation/downregulation across genes (such as in the cases of *IFIT1* and *IRF1*, Fig. S3-6B, left). In support of the large effects of IAV, PC1 of the PBMC expression data, which is strongly correlated with IAV infection status, explains 43% of the variance in the dataset, as shown in Fig. S3-1B. If we quantile normalize data in this scenario, the dynamic range of responses to infection will be abbreviated (Fig. S3-6B, right). We believe this is biologically misleading and would particularly affect our specificity score analyses, which rely on variation in the strength of responses to IAV across cell types.

Of note, when modeling the expression estimates only for the pseudobulk "PBMC" data (e.g., all cell types combined, results shown in gray bar in Fig. 3-1C), two additional covariates were added to the model, corresponding to the first two principal components of a PCA performed on an n x m cell type proportion matrix (where n = number of samples = 180, m = number of cell types = 10, with the matrix populated by the cell type proportions for each sample [calculated by the number of cells per cell type cluster for a sample divided by the total number of cells assigned

to that sample]) to account for the majority of the variance introduced by underlying cell type composition (PC1 percent variance explained (PVE) = 53.8%, PC2 PVE = 23.2%, total = 77.0%).

These models were fit using the lmFit and eBayes functions in limma (Ritchie et al. 2015), and the estimates of the global infection effect $\beta_{IAV}(i)$ (i.e., the differential expression effects due to IAV infection) were extracted across all genes along with their corresponding p-values. We controlled for false discovery rates (FDR) using an approach analogous to that of Storey and Tibshirani (Nédélec et al. 2016; Storey and Tibshirani 2003), which makes no explicit assumptions regarding the distribution of the null model but instead derives it empirically. To obtain a null, we performed 10 permutations, where infection status label (mock/IAV) was permuted within individual. We considered genes significantly differentially-expressed upon infection if they had a β_{IAV} $|\log_2 FC| > 0.5$ and an FDR < 0.05.

Calculation of specificity score. This metric was obtained for each gene in two steps: first, we calculated the mean $\log_2$ fold-change response to IAV infection per cell type; second, we computed the coefficient of variation of those mean fold changes across cell types. This analysis was limited to genes significantly differentially expressed following IAV infection ($|\log_2$ fold-change$| > 0.5$ and FDR < 0.05) in at least one cell type. High values indicate highly cell type-specific responses to IAV, while low values indicate shared responses to IAV. While we report specificity scores calculated across all individuals in the main text, these scores are also generalizable to European Americans (EA) and African Americans (AA) separately. If calculated within population, specificity scores are highly correlated between the two genetic ancestry groups (Pearson's r = 0.82). Specificity scores are also robust to variation in average expression levels across genes: specificity scores adjusted for mean gene expression levels within cell type are highly correlated

with unadjusted scores (adj. $R^2 > 0.99$ across all cell type-condition pairs), suggesting that this calculation is largely unaffected by mean-variance trend relationships.

Calculation of IFN score. To construct the IFN score metric, we summarized the expression patterns of genes involved in the type I/II IFN response as a whole, where, within condition, we i) subset on genes belonging to the hallmark IFN gamma and alpha response pathways (Liberzon et al. 2015), ii) mean-centered and scaled the expression values for each gene across individuals, and iii) computed the average scaled expression across genes per individual. We defined the IFN response score as the difference in IFN score between the IAV-infected and mock conditions per individual.

Correlation between genetic ancestry and IFN score and permutations. To estimate the correlation between genetic ancestry and IFN score in the mock and IAV-infected conditions, we performed linear regressions to obtain Pearson correlation coefficients and p-values within each cell type. To evaluate if the observed Pearson correlation coefficient was more extreme than expected by chance, we performed permutation testing. To perform permutations within each cell type, we i) randomly permuted the genetic ancestry label across individuals, ii) re-calculated the Pearson correlation coefficient using these permuted genetic ancestry values, and iii) asked whether the null correlation coefficient was equal to or less than the observed, negative correlation coefficient (n permutations = 1,000). For four out of the five cell types tested (all except $CD8^+$T cells: $p = 0.095$), the observed Pearson correlation coefficient is significantly lower than most values obtained from permutation (B p-value = 0.016, $CD4^+$T = 0.005, NK = 0.002, monocytes =

0.002). Together, this analysis suggests that the association between genetic ancestry and IFN score in the IAV-infected condition is robust, although less so for $CD8^+$T cells.

Effect of baseline serum titers on gene expression. To determine whether variation in baseline IAV serum titers had a significant impact on gene expression levels, we evaluated the effect of baseline Cal/04/09-specific serum IgG antibody titers: (1) on gene expression levels at both baseline and following IAV infection, and (2) on gene expression responses following IAV infection (i.e., the interaction between serum titer and transcriptional response to IAV infection) using linear models taking into account age and serum batch. The models were fit using the lmFit function in limma (Ritchie et al. 2015). P-values were extracted and corrected for multiple testing using the Benjamini-Hochberg procedure (p.adjust function in R, method = "BH"). No genes for which expression levels (model 1) or response magnitude (model 2) significantly (FDR < 0.05) correlated with baseline titers were detected in any of the five cell types. Therefore, the levels of Cal/04/09-specific antibody titers do not significantly influence the early transcriptional response to IAV infection, at least in PBMCs.

Modeling genetic ancestry effects and integration with mashr. Prior to modeling genetic ancestry effects, capture-corrected expression estimates were quantile normalized (QN) within condition using qqnorm in R to minimize the risk of identifying spurious associations due to outlier effects. We note, however, that our key findings are robust to our decision to quantile normalize the data: our popDE and popDR ancestry estimates are highly concordant whether we use QN or non-QN gene expression estimates (adj. R^2 range = 0.66 – 0.92 across all cell types and conditions, mean = 0.82). Moreover, with or without quantile normalization, we observe that higher levels of

European ancestry are associated with increased type I/II interferon pathway activity across cell types following IAV infection (*without* QN: mean Pearson's r = -0.26, Fisher's meta-p = 2.9×10^{-6}; *with* QN: mean Pearson's r = -0.25, Fisher's meta-p = 4.8×10^{-6})**,** and that genetic ancestry-associated genes are enriched among genes positively correlated with COVID-19 disease severity (*without* QN: 2.0 to 2.1 fold-enrichment, p-values = 2.5×10^{-5} [mock] and 9.2×10^{-6} [IAV]; *with* QN: 2.0 to 2.2 fold-enrichment, p-values = 2.7×10^{-8} [mock] and 3.5×10^{-6} [IAV]).

The following nested linear model was used to identify genes for which expression levels correlated with the proportion of African ancestry across individuals successfully genotyped (n = 89/90 individuals) within condition (i.e., popDE genes):

$$\text{M}_2\text{: } E(i,j) \sim \begin{cases} \beta_0(i) + \beta_{AA}^{mock}(i) \cdot AA(j) + \beta_{pB}(i) \cdot pB^{mock}(j) + \beta_{age}(i) \cdot age(j) + \varepsilon^{mock}(i,j) & \text{if Condition} = mock \\ \beta_0(i) + \beta_{IAV}(i) + \beta_{AA}^{IAV}(i) \cdot AA(j) + \beta_{pB}(i) \cdot pB^{IAV}(j) + \beta_{age}(i) \cdot age(j) + \varepsilon^{IAV}(i,j) & \text{if Condition} = IAV \end{cases}$$

Here, $E(i,j)$ represents the capture-corrected expression estimate of gene *i* for individual *j*, $\beta_0(i)$ is the global intercept accounting for the expected expression of gene *i* in a 100% European-ancestry mock-infected individual, $\beta_{AA}^{mock}(i)$ and $\beta_{AA}^{IAV}(i)$ indicate the effects of African admixture (mean-centered, scaled African ancestry proportion, $AA(j)$) on gene *i* within each condition, and $\beta_{IAV}(i)$ represents the intrinsic infection effect of IAV infection. Age represents the mean-centered, scaled (mean = 0, sd = 1) age per individual, with $\beta_{age}(i)$ being the effect of age on expression levels. All other terms in the model are analogous to that described in M_1. Again, the model was fit using limma (Ritchie et al. 2015), and the estimates $\beta_{AA}^{mock}(i)$ and $\beta_{AA}^{IAV}(i)$ of the genetic ancestry effects were extracted across all genes, along with their corresponding p-values. Each of these estimates represents the genetic ancestry-related differential expression effects within each condition.

Genes for which the response to IAV infection correlate with the proportion of African ancestry (i.e., popDR genes) were detected using the following model:

$$\text{M}_3\text{: } E(i,j) \sim \begin{cases} \beta_0(i,j) + \beta_{pB}(i) \cdot pB^{mock}(j) + \varepsilon^{mock}(i,j) \text{ if Condition = } mock \\ \beta_0(i,j) + \beta_{pB}(i) \cdot pB^{IAV}(j) + (\beta_{IAV}(i) + \beta_{age}^{IAV}(i) \cdot age(j) + \beta_{AA}^{IAV}(i) \cdot AA(j)) + \varepsilon^{IAV}(i,j) \text{ if Condition = } IAV \end{cases}$$

This model is similar to M_1 (effect of IAV infection), in that it allows us to obtain estimates based on within-individual variability (i.e., estimation of individual effects), with the difference that the IAV infection effect is no longer built in a genetic ancestry-independent manner, as in model M_1, since it is now dependent on genetic ancestry as follows: $\beta_{IAV}(i) + \beta_{age}^{IAV}(i) \cdot age(j) + \beta_{AA}^{IAV}(i) \cdot AA(j)$. In this model, we explicitly correct for age effects on the response to IAV itself (captured by the $\beta_{age}^{IAV}(i)$ term). In this context, $\beta_{AA}^{IAV}(i)$ denotes the genetic ancestry-infection interaction effect induced by IAV infection (i.e., the effect of genetic ancestry on the response to IAV) corrected for age effects, which represents variation in the response to infection that is correlated with the proportion of African ancestry. The key difference between M_2 and M_3 is that M_2 does not control for the effects of individual identity. Including individual-wise intercepts in M_3 allows us to better take into account the paired nature of the data to assess the effect of admixture on within-individual responses to IAV.

To assess sharing of genetic ancestry effects across cell types and to increase our power to detect these effects, we applied Multivariate Adaptive Shrinkage in R (mashr v0.2.28) (Urbut et al. 2019) to the outputs of our popDE and popDR cell type-by-cell type models. mashr is able to learn prior patterns of effect size sharing across data sets (here, cell types and infection conditions) using an empirical Bayes approach, combining information across genes to fit flexible prior models. mashr was applied independently to both the popDE and popDR priors, so all of the following methods were performed twice, once for the popDE effects and then again for the popDR effects. Effect size priors were obtained directly from limma (Ritchie et al. 2015) and merged into matrices including all effect sizes across cell types, only keeping those genes detected in all cell

types (i.e., n x m matrices, where for popDE effects: n = 6,847 genes, m = 10 conditions [mock- and IAV-infected popDE effects for each of the 5 main cell types], and for popDR effects: n = 6,847 genes, m = 5 conditions [popDR effects for each of the 5 main cell types]). Standard errors of the effect size priors were calculated per gene by multiplying the square root of the posterior variance (s2.post) of each gene by the unscaled standard deviation for the effect size of interest for that gene (stdev.unscaled) estimated by limma, and these values were similarly formatted into matrices as described above. To account for correlations among measurements across conditions in our data, we used the estimate_null_correlation_simple function implemented in mashr (Urbut et al. 2019) to specify a correlation matrix prior to fitting the mash model. We included both the canonical covariance matrices provided by default in mashr and data-driven covariance matrices (defined as the top 5 PCs from a PCA performed on the significant (lfsr < 0.05) signals detected in the condition-by-condition model results) learned from our data in the mash model fit. For both popDE and popDR effects, the mash model was fit to all tests using the mash function. Posterior summaries of the effect sizes, standard deviations, and measures of significance were extracted. We used the estimated local false sign rate (lfsr) to assess significance of our posterior popDE and popDR effects and considered genes significantly population differentially-expressed or differentially-responsive if the lfsr of the posterior mean was < 0.10. We chose to implement mashr to gain power to identify genetic ancestry effects and to obtain improved effect size estimates by leveraging shared information across different cell types. We used mashr for this purpose because it is both computationally tractable and highly flexible in learning patterns of correlation across conditions.

Comparison with Quach et al. data. Raw count files from Quach et al. (Quach et al. 2016) were downloaded from the European Genome-Phenome Archive (https://ega-archive.org/studies/EGAS00001001895). To determine the extent to which our popDE effects in the monocytes coincided with those identified in the Quach et al. data, we modeled genetic ancestry effects in the Quach et al. dataset (197 European-ancestry and African-Ancestry individuals residing in Belgium) utilizing a similar approach to that used for our data ("Modeling genetic ancestry effects and integration with mashr", Methods). Prior effect size and standard error estimates were obtained using a model similar to M_2 but accounting for dataset-specific covariates (e.g., percent GC content, 5' to 3' bias ratio, experiment date, and library date). To assess sharing of genetic ancestry effects across datasets, we integrated our monocyte results and the Quach et al. results using the mashr (Urbut et al. 2019) framework. PopDE effect sizes were largely concordant between the two datasets among genes with significant genetic ancestry effects (lfsr < 0.10) in our single-cell IAV data (Pearson's r = 0.662 for the mock condition, Pearson's r = 0.499 for the IAV condition, $p << 1x10^{-10}$ for both conditions, Fig. S3-2C).

We note that the observed correlation is quite strong, considering that our study is based on individuals living in Miami, Florida, while the cohort in Quach et al. (Quach et al. 2016) includes individuals living in Belgium (likely leading to variation in ancestry effects due to unmeasured environmental and/or social factors that are differentially correlated with genetic ancestry in different countries). Additionally, while Quach et al. evaluated the transcriptomic responses of purified $CD14^+$ monocytes to IAV infection, we performed infections on PBMCs as a whole (allowing us to also capture the effects of paracrine signaling between cell types and direct cell-cell interactions) using a different IAV strain (A/California/04/2009 in this study versus A/USSR/90/1977 in Quach et al).

eQTL mapping and integration with mashr. eQTL mapping was performed using the pseudobulk expression data independently in each cell type against the sets of genes retained after lowly-expressed gene filtering (n genes: $CD4^+$ T cells = 9,960, $CD8^+$ T cells = 9,335, B cells = 9,291, monocytes = 10,424, NK cells = 7,109, PBMCs = 10,430) and was also performed using the paired "true" bulk RNA expression data. A linear regression model was used to examine associations between SNP genotypes and expression levels, in which expression levels were regressed against genotype. Input expression matrices were quantile-normalized within condition prior to running the association. Mock-exposed and IAV-infected eQTL were mapped separately across all cell types. All regressions were performed using the R package MatrixEQTL (v2.3) (Shabalin 2012). Only SNPs with a minor allele frequency > 5% across all individuals were tested, and SNPs with > 10% of missing data or deviating from Hardy-Weinberg equilibrium at $p < 10^{-5}$ were excluded (--maf 0.05 --geno 0.10 --hwe 0.00001 PLINK v1.9 filters, www.cog-genomics.org/plink/1.9/) (Chang et al. 2015). In total, 6,305,923 SNPs passed our quality-control filters. Local associations (i.e., putative *cis*-eQTL) were tested against all SNPs located within the gene body or 100kb upstream and downstream of the transcription start site (TSS) and transcription end site (TES) for each gene tested. We recorded the minimum p-value (i.e., the strongest association) observed for each gene, which we used as statistical evidence for the presence of at least one eQTL for that gene. To estimate an FDR, we permuted the genotype data ten times, re-performed the linear regressions, and recorded the minimum p-values for the gene for each permutation. These sets of minimum p-values were used as an empirical null distribution and FDRs were calculated using the method described in the section "Modeling global infection effects".

Power to detect *cis*-eQTL can be increased by accounting for unmeasured surrogate confounders. To identify these confounders, we first performed PCA on a correlation matrix based

on gene expression for mock and IAV-infected samples. Subsequently, up to 20 principal components (PCs) were regressed out prior to performing the association analysis for each gene. A specific number of PCs to regress in each condition and cell type, corresponding to the number of PCs that empirically led to the detection of the largest number of eQTL in each condition, was then chosen from these results. The exact number of PCs regressed in each of the analyses can be found in Table S3-11. Of note, while PC corrections increase our power to detect eQTL, they do not affect the underlying structure of the expression data.

Mapping was performed combining both European American and African American individuals to increase power. To avoid spurious associations resulting from population structure, the first two eigenvectors obtained from a PCA on the genotype data using SNPRelate (v1.20.1, gdsfmt v1.22.0) (X. Zheng et al. 2012) were included in the Matrix eQTL model. Other covariates included in the linear model were the following: the condition/cell type-specific bimodal proportion and age (mean-centered, scaled), with two additional covariates included when mapping eQTL using the PBMC expression data, corresponding to the first two PCs from the cell type composition PCA described in "Modeling global infection effects".

Our ability to detect eQTL in the pseudobulk expression data was highly dependent on the number of cells identified in each cell type cluster (correlation between the total number of cells recovered per cell type across all individuals/conditions versus the number of significant eQTL [FDR < 0.10] detected: adj. $R^2 = 0.983$, $p = 1x10^{-8}$). To gain power to detect *cis*-eQTL effects using sharing information across cell types, we again implemented mashr (Urbut et al. 2019). Out of necessity of the method, we only considered shared genes that were tested across all cell types (n = 6,573). For each of these genes, we chose a single, top *cis*-SNP, defined as the SNP with the lowest FDR across all cell types (n = 5) and conditions (n = 2), to input into mashr, yielding a total

of 6,573 gene-SNP pairs. We extracted the prior effect sizes (betas) and computed the standard errors (SEs) of these betas (defined as the beta divided by the t-statistic) from the Matrix eQTL outputs for each gene-SNP pair across cell types and conditions. We defined a set of "strong" tests (i.e., the 6,573 top gene-SNP associations) as well as a set of random tests, including both null and non-null tests, which we obtained from randomly sampling 200,000 rows of a matrix containing all gene-SNP pairs tested by Matrix eQTL merged across conditions. Our mashr workflow was as follows: i) the correlation structure among the null tests was learned using the random test subset, ii) the data-driven covariance matrices were learned using the strong test subset, iii) the mash model was fit to the random test subset using canonical and data-driven covariance matrices, with two additional "infection" covariance matrices (i.e., one matrix capturing shared effects in only the mock-exposed samples and another matrix capturing shared effects in only the IAV-infected samples), and iv) the posterior summaries were computed for the strong test subset (Urbut et al. 2019). We used the estimated local false sign rate (lfsr) to assess significance of our posterior eQTL effects and considered a gene-SNP pair to have a significant eQTL effect if the lfsr of the posterior mean was < 0.10, which we defined as an eGene. For the effect size comparison and eGene sharing analyses between the pseudobulk PBMC and true bulk PBMC expression data (Figs. S3-4A and S3-4B), matrix eQTL effect size estimates for the same top gene-SNP associations as defined above were input into mashr (Urbut et al. 2019) to obtain posterior effect size estimates and significance values.

Identification of condition-specific popDE genes and eGenes. Within each cell type, we considered either popDE genes or eGenes as condition-specific (i.e., only showing an effect in either the mock or IAV infection condition) if they had an lfsr < 0.10 in only one condition. Here,

we assume that the risk of identifying a true effect in both mock and IAV-infected cells (i.e., a shared popDE gene/eGene) as falsely condition-specific due to lack of power is low, specifically because we employed the multivariate adaptive shrinkage framework, which draws information across conditions to make better-informed posterior estimates regarding the sharing of effects, so we do not expect to see many posterior effects called as "condition-specific" when, in fact, they are not.

Enrichment of eGenes within popDE genes. We tested for an enrichment of eGenes among the genes identified as popDE genes within each cell type and condition. For each cell type, condition pair, we created two vectors: i) a popDE gene vector, where significant popDE genes (lfsr < 0.10) were coded as a 1 and non-significant popDE genes were coded as a 0, and ii) an eGene vector, where significant eGenes (lfsr < 0.10) were coded as a 1 and non-significant eGenes were coded as a 0. The logistic regression was performed on the popDE gene and eGene vectors using glm in R, where the eGene vector was used as the predictor variable and the popDE gene vector was used as the response variable (popDE[0,1] ~ eGene[0,1]). The odds ratios output by glm were converted to $\log_2$ fold enrichments with a 95% confidence interval (plotted along the x-axis in Fig. 3-3C).

Calculation of predicted and observed population differences in expression. We estimated the predicted *cis*-genetic population differences in gene expression using a method in which we first computed the predicted expression of each gene considering only the posterior effect size of the top *cis*-SNP for that gene and an individual's genotype dosage (a vector of 0, 1, or 2), where, for gene *i*, individual *j*:

$$predicted\ expression_{i,j} = eQTL\ effect\ size_i * genotype_j$$

We then modeled these predicted expression values using a model analogous to that of M_2 (model evaluating the popDE effects, "Modeling genetic ancestry effects and integration with mashr") to obtain the predicted genetic ancestry effects (plotted on the x-axis for genetically-driven popDE genes in Fig. 3-3D). The observed population differences in expression were taken directly from the post-mash effect size estimates of M_2 (plotted on the y-axis for genetically-driven popDE genes in Fig. 3-3D).

Modeling the effect of *cis*-regression on the observed population differences in expression. To assess the impact of *cis*-regression on population-associated expression differences, we used two models evaluating the effect of continuous genetic ancestry (African ancestry proportion) on gene expression: i) a model analogous to M_2 (model evaluating the popDE effects, "Modeling genetic ancestry effects and integration with mashr"), and ii) a model in which, for each gene, the top *cis*-SNP for that gene was regressed by including the genotype dosage for that SNP across individuals as a covariate in the model. The models were fit using limma (Ritchie et al. 2015) and mashr was applied (as described in the section "Modeling genetic ancestry effects and integration with mashr") (Urbut et al. 2019) to the prior effect sizes and standard errors derived from both models independently. The mashr posterior summaries were used to directly obtain the population differences in expression for each gene.

For each significantly enriched GO term (FDR < 0.01, hypergeometric test) identified in Fig. 3-3E (see "Gene set enrichment analyses" section below), we calculated summaries of the observed population difference in expression among the genes that belong to each term that are also popDE genes with evidence of an eQTL in at least one cell type. To do this, for each cell type for each term, we collected the observed population differences among these term-specific

genetically-driven popDE genes and calculated the median and standard error (SE) for these values (plotted on the x-axis in Fig. 3-3G). This was performed for both the observed ("real") model outputs (model i) as well as the *cis*-regressed model outputs (model ii). For each cell type, we obtained a p-value for the real effects using a permutation method. To obtain a null distribution, we performed 1,000 permutations where, for each iteration, we: 1) sampled the same number of observed term-specific, genetically-driven popDE genes for that cell type from a background set of all genetically-driven popDE for that cell type, 2) obtained the population differences in expression among these genes, and 3) calculated the median for these null values. We then computed a one-sided, empirical p-value, where we considered the number of instances more extreme in the median null difference compared to the median observed difference in the real data given the sign of the difference in the real data (i.e., if the observed difference in the real data was < 0, we counted the number of observations in the null distribution equal to or less than the observed value, and if the observed difference in the real data was > 0, we counted the number of observations in the null distribution equal to or greater than the observed value), where p = number of instances more extreme divided by the number of permutations (n = 1,000). Similarly, we obtained a p-value for the *cis*-regressed effects using the same method, except for that in steps 2 and 3, we considered the *cis*-regressed population differences as opposed to those seen in the real data. To calculate the directional p-value for the *cis*-regressed case, we used the magnitude of the median *cis*-regressed population difference but still considered the sign of the median observed population difference.

Processing of COVID-19 patient single-cell RNA-sequencing data. Raw and normalized count files from Su et al. (Su et al. 2020) were downloaded from EMBL-EBI ArrayExpress

(https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9357/). The entire dataset consisted of 549,047 cells across 270 samples (505,616 cells from COVID-19 patients, n = 254 COVID-19 samples). Normalized counts were scaled and the effect of percent mitochondrial reads per cell was regressed using the ScaleData function in Seurat (v3.1.5) (Stuart et al. 2019). After scaling, dimensionality reduction was performed via UMAP (RunUMAP function, dims = 1:30) and PCA (RunPCA function, npcs = 30) (Stuart et al. 2019). A Shared Nearest Neighbor (SNN) Graph was constructed using the FindNeighbors function (dims = 1:20, all other parameters set to default), and clusters were subsequently called using the FindClusters algorithm (resolution = 0.5, all other parameters set to default) (Stuart et al. 2019). Clusters were annotated as described in the section "Clustering, cell type assignment, and UMAP analysis," and any clusters with ambiguous immune cell marker gene expression patterns were excluded. In total, considering the five major PBMC immune cell populations, we identified 493,809 cells ($CD4^+$ T cells = 145,698, $CD8^+$ T cells = 79,626, monocytes = 174,980, B cells = 33,478, and NK cells = 60,027) across samples. Pseudobulk estimates and associated per-sample bimodality proportions were calculated for each cell cluster independently as described in "Calculation of pseudobulk estimates" and "Calculation of capture-corrected expression for downstream modeling". Lowly-expressed genes were filtered using cell-type specific cutoffs (removed genes with a median logCPM < 1.5 in $CD4^+$ T cells and monocytes, < 2.0 in NK cells and $CD8^+$ T cells, and < 3.0 in B cells), leaving the following number of genes per cell type: $CD4^+$ T cells = 10,079, $CD8^+$ T cells = 9,809, B cells = 8,865, monocytes = 10,069, NK cells = 10,301. Batch-corrected expression estimates and inverse variance weights were obtained as described in "Calculation of capture-corrected expression for downstream modeling".

Modeling COVID-19 disease severity-associated genes. Only COVID-19 patient samples from Blood.draw.time.point == “T1” (blood draw performed shortly after the initial clinical diagnosis (Su et al. 2020)) and with non-missing disease severity scores (n = 129) were included in the downstream analyses. Additionally, two individuals were removed due to extremely skewed pseudobulk expression density distributions, leaving n = 127 individuals for downstream analysis. Severity of COVID-19 was assessed using the 9-point World Health Organization (WHO) Ordinal Scale (WOS) for Clinical Improvement (https://www.who.int/publications/i/item/covid-19-therapeutic-trial-synopsis) that includes the following categories: 0 = uninfected - no evidence of infection; 1 = ambulatory - no limitation of activities; 2 = ambulatory - limitation of activities; 3 = hospitalized, mild - no oxygen therapy; 4 = hospitalized, mild - oxygen by mask or nasal prongs; 5 = hospitalized, severe - non-invasive ventilation or high-flow oxygen; 6 = hospitalized, severe - intubation and mechanical ventilation; 7 = hospitalized, severe - ventilation and additional organ support; and 8 = dead – death. If WOS score for an individual was coded as “1 or 2”, this value was manually set to 1.5. If self-identified ethnicity or race was not reported for an individual (n missing ethnicity = 5, n missing race = 18), this missing data was filled with the factor variable most likely to represent the missing value, i.e., for ethnicity, "Unknown / Not Reported" label changed to "Not Hispanic or Latino" label, and for race, "Unknown / Not Reported" label changed to “White” label.

Prior to modeling WOS effects, batch-corrected expression estimates were quantile-normalized using qqnorm in R. The following linear model was used to identify genes for which expression levels were correlated with WOS across individuals (i.e., WOS-associated genes):

$$E(i,j) \sim \beta_0(i) + \beta_{WOS}^{T1}(i) \cdot WOS(j) + \beta_{pB}(i) \cdot pB^{T1}(j) + \beta_{age}(i) \cdot age(j) + \beta_{sex}(i) \cdot sex(j) +$$

$$\beta_{ethnicity}(i) \cdot ethnicity(j) + \beta_{race}(i) \cdot race(j) + \varepsilon^{T1}(i,j)$$

Here, $E(i,j)$ represents the batch-corrected expression estimate of gene *i* for individual *j*, $\beta_0(i)$ is the global intercept accounting for the expected expression of gene *i* in a female, non-Hispanic white individual, and $\beta_{WOS}^{T1}(i)$ indicates the effect of WOS ($WOS(j)$) on gene *i* in blood draw time point T1. Further, pB^{T1} represents the bimodal proportion estimated per sample for the respective cell type being modeled in blood draw T1 samples, age represents the mean-centered, scaled (mean = 0, sd = 1) age per individual, sex represents the self-identified sex for each individual (factor levels = “Male”, “Female”), ethnicity represents the self-identified ethnicity for each individual (factor levels = "Not Hispanic or Latino", "Hispanic or Latino"), and race represents the self-identified race for each individual (factor levels = "White", "Asian", "Black or African American", "Native Hawaiian or Other Pacific Islander", "American Indian/Alaska Native", "More Than One Race"). Finally, ε^{T1} represents the residuals for the T1 time point samples for each gene *i*, individual *j* pair. The model was fit using limma (Ritchie et al. 2015), and the estimate $\beta_{WOS}^{T1}(i)$ was extracted across all genes, along with the corresponding p-values. This estimate represents the WOS-associated differential expression effects within blood draw time point T1.

We applied mashr (v0.2.28) (Urbut et al. 2019) to assess cell type sharing of WOS effects as described in “Modeling genetic ancestry effects and integration with mashr”. Only genes detected across all cell types considered were kept (n = 7,866 genes). Posterior summaries of the effect sizes, standard deviations, and measures of significance were extracted. The estimated local false sign rate (lfsr) was used to assess significance of WOS effects, and we considered genes significantly WOS-associated if the lfsr of the posterior mean was < 0.01. To calculate enrichments of popDE genes among WOS-associated genes (Figs. 3-4B and S3-5A), only WOS-associated

genes in the background set of genes tested for popDE effects across cell types (n = 6,847 genes) were considered.

Gene set enrichment analyses. Gene set enrichment analysis was performed using three independent methods, including fgsea (https://bioconductor.org/packages/release/bioc/html/fgsea.html), GOrilla (Eden et al. 2009), and ClueGO (Bindea et al. 2009), depending on the type of data being evaluated. The enrichment program specifications and the data in which they were used to assess enrichments are described below:

The R package fgsea (v1.10.1) was used to perform gene set enrichment analysis for the global infection effects using the C5 gene ontology (GO) biological processes gene sets (Fig. 3-1D), for the popDE effects using the H hallmark gene sets (Fig. 3-2C) (Subramanian et al. 2005), and for the popDE effects using the WOS-associated gene sets (Figs. 3-4C and S3-5B). For the infection effects, t-statistics were obtained directly from the topTable function in limma (Ritchie et al. 2015), and for the popDE effects, t-statistics were calculated from the posterior mashr outputs, where the t-statistic = posterior effect size divided by the posterior standard error for each gene. For infection effects, the background sets were the sets of genes sufficiently expressed (i.e., passed the lowly-expressed gene filter threshold) for each cell type. For popDE effects, the background set was the set of genes detected in all cell types (i.e., the intersection set of genes that were measurably expressed across all cell types, n = 6,847). This set was chosen as the background because we conducted our enrichment analyses using t-statistics calculated from mash-adjusted posterior effect sizes and standard errors (see "Modeling genetic ancestry effects and integration with mashr"), which could only be calculated for genes detectably expressed in all conditions (i.e.,

all 5 cell types in mock- and IAV-infected conditions). The t-statistics were then ranked, and these pre-ranked t-statistics were used to perform the enrichment using fgsea with the following parameters: minSize = 15, maxSize = 500, nperm = 100000. Enrichments scores (ES) and Benjamini-Hochberg adjusted p-values output by fgsea were collected for each analysis.

We also used fgsea to generate the barcode plots shown in Fig. 3-1G to visualize where the genes in the highlighted pathways are found in the ranked specificity score list among the set of all infection differentially-expressed genes in at least one cell type. To obtain p-values for the ranked list of specificity scores, we used GOrilla (Eden et al. 2009). GOrilla relies on a statistical framework (the minimum hypergeometric score) that allows the calculation of exact p-values for observed enrichments in ranked lists of genes, taking into account multiple testing without needing to perform simulations, unlike fgsea. Because GOrilla only identifies GO terms that are significantly enriched at the top of the ranked gene list, we performed the enrichments in two ways, once with the list ranked from high to low specificity scores and again with the list ranked from low to high specificity scores. The Benjamini-Hochberg adjusted FDR q-values calculated by GOrilla for the "viral gene expression" and "response to type I interferon" terms are reported in Fig. 3-1G.

We performed gene set enrichment analysis for our intersection set of popDE genes and eGenes (Fig. 3-3E) using the ClueGO (v2.5.7) (Bindea et al. 2009) Cytoscape (v3.7.1) (Shannon et al. 2003) module in functional analysis mode, where the target set of genes was the list of popDE eGenes in the mock or IAV condition and the background set was the list of genes tested across all cell types. Specifically, we tested for the enrichment of GO terms related to biological processes (ontology source: GO_BiologicalProcess-EBI-UniProt-GOA_04.09.2018_00h00) using the following parameters: visual style = Groups, default Network Specificity, no GO Term Fusion,

min. GO Tree Interval level = 3, max. GO Tree Interval level = 8, min. number of genes = 3, min. percentage of genes = 4.0, statistical test used = Enrichment/Depletion (two-sided hypergeometric test), p-value correction = Benjamini-Hochberg. For the graphical representation of the enrichment analysis, ClueGO clustering functionality was used (kappa threshold score for considering or rejecting term-to-term links set to 0.4). Only pathways with an FDR < 0.01 were reported.

Supplementary Figures and Tables

Supplementary figures and tables for this chapter are included in Appendix A: Supplementary Figures and Tables.

Chapter IV: Widespread gene-environment interactions in the immune response to SARS-CoV-2 infection

Note:

The following section (*Chapter IV*) represents a currently unpublished version of a manuscript that will be submitted to a journal at a later date.

Authors:

H.E. Randolph, R. Aguirre-Gamboa, V. Locher, E. Ketter, C. Brandolino, A. Dumaine, E. Brunet-Ratnasingham, T. Nakanishi, J.B. Richards, D.E. Kaufmann, and L.B. Barreiro

Abstract

Genome-wide association studies performed in patients with Coronavirus Disease 2019 (COVID-19) have uncovered various loci significantly associated with susceptibility to SARS-CoV-2 infection and COVID-19 disease severity. However, the underlying *cis*-regulatory genetic factors that contribute to heterogeneity in the response to SARS-CoV-2 across different immune cell types and subsets remain largely uncharacterized. Here, we used single-cell RNA-sequencing to quantify genetic contributions to *cis*-regulatory variation in peripheral blood mononuclear cells

in 52 COVID-19 patients with varying degrees of disease severity and 108 healthy controls. Expression quantitative trait loci (eQTL) mapping within each cell type revealed thousands of *cis*-associated variants, with $CD14^+$ monocytes displaying the greatest number of eQTL across all cell types tested. Further, we find hundreds of *cis*-eQTL that are only detected among COVID-19 patients exclusively in $CD14^+$ and $CD16^+$ monocytes or that are significantly associated with functional cell state. Our findings demonstrate that gene-environment interactions are prevalent in the context of SARS-CoV-2 infection *in vivo*, show that continuous cell states can capture dynamic cell state-dependent *cis*-regulatory effects in patient cohorts, and underscore the importance of expanding the study of regulatory variation to relevant cell types and disease contexts.

Summary of Results

Host genetic factors contribute to variation in the immune response and susceptibility to viral infection across individuals. Many immune response expression quantitative trait loci (eQTL) studies have been performed with the goal of linking genetic polymorphisms to transcriptional immune response variation. Previous studies have mainly taken advantage of *in vitro* immune challenge models to map these loci (Fairfax and Knight 2014), and a few have explicitly demonstrated the role of environmental factors in modifying genetic effects, identifying both cell type-specific eQTL and eQTL induced only upon infection (i.e., response eQTL) (M. N. Lee et al. 2014; Nédélec et al. 2016; Quach et al. 2016; Randolph et al. 2021; Aquino et al. 2022). More recently, there has been an effort to consider other types of genetic interaction effects, facilitated by the availability of genotyped population-scale cohorts characterized by single-cell RNA-sequencing (van der Wijst et al. 2020). Continuous cell state-dependent eQTL, or eQTL that

display an interaction with some cellular context defined at single-cell resolution, have been shown to explain more variation in gene expression compared to conventional non-interacting eQTL (Nathan et al. 2022). Further, autoimmune risk variants have been shown to be enriched for these state-dependent loci (Nathan et al. 2022), highlighting the importance of cell state context in better understanding disease-relevant variants.

Substantial immune response variation and disease heterogeneity exists among individuals infected with SARS-CoV-2, the virus that causes COVID-19. While much of this variation can be attributed to environmental and social determinants (F. Zhou et al. 2020), genetic factors have also been documented to play a role. Genome-wide association studies (GWAS) conducted for susceptibility to SARS-CoV-2 infection and severe COVID-19 disease have revealed multiple genome-wide significant loci associated with these traits (Niemi et al. 2021; The Severe Covid-19 GWAS Group 2020). Further, an eQTL mapping study performed in peripheral blood mononuclear cells (PBMCs) collected from healthy individuals exposed to SARS-CoV-2 *in vitro* found that response eQTL were highly cell type-dependent and specific to SARS-CoV-2 when considering the myeloid response (Aquino et al. 2022). However, there have been no studies to date probing how genetic variation impacts immune response diversity in cells derived from COVID-19 patients with active SARS-CoV-2 infection. Here, we explore the influence of genetic interaction effects in the context of SARS-CoV-2 infection *in vivo*, specifically considering cell type-specific, disease state-specific, and cell state-dependent regulatory heterogeneity.

In this study, we used single-cell RNA-sequencing to profile the transcriptomes of PBMCs collected from healthy, non-infected control individuals (n = 108), hospitalized COVID-19 patients across a spectrum of severity during the acute stage of infection (on average, 11 days after symptom onset; n = 52), and a subset of recovered COVID-19 patients resampled at a six-month

follow-up time point (n = 12) (Fig. 4-1A). Additionally, all individuals were genotyped, allowing us to measure *cis*-regulatory genetic effects and gene-environment interaction effects in downstream analyses. Across individuals, we captured 306,324 high-quality single-cell transcriptomes (n = 163,639 controls, n = 142,685 COVID-19 patients). Clustering of these cells followed by cell type label transfer annotation from a multimodal human PBMC reference dataset (Hao et al., detailed in Methods) revealed 30 distinct immune cell types at fine-scale resolution (Fig. 4-1B). In addition to these fine-scale populations, we also defined a set of high-level cell types based on marker gene annotations of clusters, which corresponded to the six major cell types identified, including $CD4^+$ T cells, $CD8^+$ T cells, B cells, natural killer (NK) cells, $CD14^+$ monocytes, and $CD16^+$ monocytes. Within these high-level cell type populations, we then collapsed our single-cell gene expression estimates into pseudobulk estimates per sample, generating six bulk-like gene expression matrices that were used for subsequent modeling of disease state and severity effects.

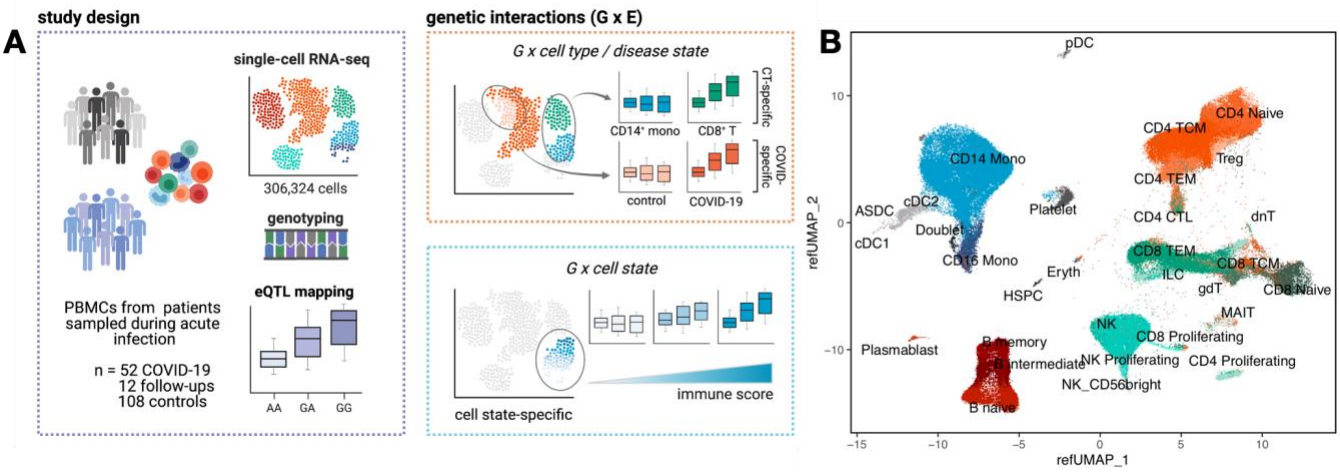


Fig. 4-1. Summary of the study cohort and aims. (A) Study design (left) and examples of various gene-environment interactions, including cell type-, disease state-, and cell state-dependent effects, evaluated in this study (right). (B) UMAP of all cells (n = 306,324) collected across individuals projected onto a reference UMAP derived from Hao et al. (Hao et al. 2021).

We first examined the impact of SARS-CoV-2 infection on gene expression signatures by comparing all COVID-19 patients against healthy controls. We found that $CD14^+$ monocytes showed the most pronounced effects of SARS-CoV-2 infection [n = 1,464 differentially expressed (DE) genes, defined as those with $|\log_2$ fold change$| > 0.5$ and false discovery rate (FDR) < 0.05, corresponding to 14.5% of genes tested], with NK cells (11.8% DE genes) and $CD16^+$ monocytes (11.3% DE genes) displaying slightly weaker infection effects. In contrast, cell types in the adaptive immune compartment were much less responsive (1.5 - 2.7% DE genes). Given the relatively large effect of SARS-CoV-2 infection on expression profiles in certain cell types, we next sought to more precisely dissect how severity of infection influences variation in cell type proportions across individuals and gene expression patterns within COVID-19 patients. Disease severity was assessed using a five-point scale of respiratory support needed at the time of patient sampling, encompassing the following groups: Moderate ("MOD", n = 17), Severe ("SEV", n = 11), 2-Critical ("CRIT2", n = 4), 3-Critical ("CRIT3", n = 17), and 4-Critical ("CRIT4", n = 1). Non-critical patients (i.e., those with moderate and severe disease) required no to minimal oxygen supplementation, whereas critical patients required mechanical ventilation, ranging from non-invasive ventilation (CRIT2) and intubation (CRIT3) to extracorporeal membrane oxygenation (CRIT4).

We found that SARS-CoV-2 infection restructures the underlying cell type composition of various PBMC populations and subsets compared to healthy individuals, with the magnitude of disease severity further modifying this effect. This was most evident in the substantial expansion of $CD14^+$ monocytes in moderate, severe, and critical cases compared to healthy donors ($p < 0.001$ for all comparisons against controls; here, all critical patients [CRIT2 – 4] were considered as a

single group), with this increase being most prominent among severe and critical cases (Fig. 4-2A). Further, we observed that the proportion of $CD14^{+}$ monocytes was strongly associated with disease severity, with more serious cases often displaying a greater proportion of $CD14^{+}$ monocytes 11 days post symptom onset (Fig. 4-2B). We also detected fewer NK cells specifically in critical patients ($p < 0.01$) and reductions of both $CD56^{bright}$ NK cells ($p < 1 \times 10^{-5}$) and plasmacytoid dendritic cells (pDCs) ($p < 0.01$) in all severity groups compared to non-infected individuals (Fig. 4-2A). pDCs are known for their ability to secrete large quantities of type I interferon (IFN) following viral infection (Fitzgerald-Bocarsly, Dai, and Singh 2008), and NK cells are key facilitators of antiviral immunity, with $CD56^{bright}$ NK cells being efficient producers of IFN-γ, TNF-α, and GM-CSF (Poli et al. 2009). Together, this suggests that SARS-CoV-2 infection leads to dysregulation of the immune response marked by poor cytokine production and blunted innate immune responses.

To further tease apart how variation in severity influences the immune response, we explored differences in global gene expression patterns within COVID-19 patients. We formally modeled the effect of severity on gene expression in COVID-19 patients, considering respiratory support severity score as a numeric variable, which allowed us to capture genes with expression levels linearly associated with severity. By far, $CD14^{+}$ monocytes showed the largest number of genes correlated with severity (n = 1,997, 19.7% of the transcriptome; FDR < 0.05), while other cell types had much less prominent effects (0 - 1.9% severity-associated genes). In line with this, principal component analysis (PCA) on the $CD14^{+}$ monocyte pseudobulk expression data revealed that variation in disease severity had a profound impact on the transcriptional response of these cells, reflected in principal component (PC) 1 (12.8% percent variance explained [PVE]) and PC2

(11.2% PVE), which both separated non-critical patients (moderate/severe) from critical patients (Fig. 4-2C).

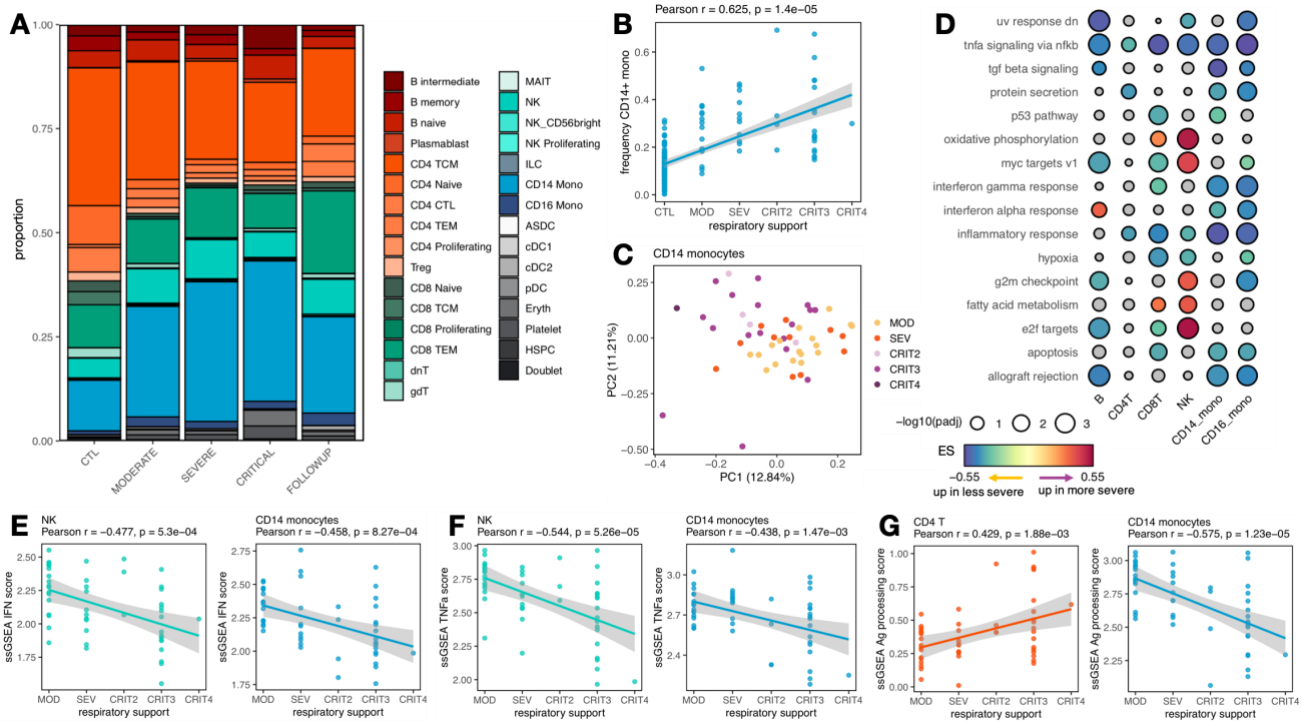


Fig. 4-2. Effects of COVID-19 disease severity at the time of patient sampling. (A) Cell type proportions in healthy controls and COVID-19 patients stratified by severity. (B) Correlation between respiratory support at the time of patient sampling and frequency of $CD14^+$ monocytes. (C) PCA decomposition of the $CD14^+$ monocyte expression data in COVID-19 patients colored by respiratory support score. (D) GO Hallmark enrichments for severity effects in COVID-19 patients across cell types. Colored circles represent pathways with FDR < 0.10; gray circles represent non-significant pathways. (E-G) Correlation between respiratory support score in various cell types and ssGSEA (E) IFN score, (F) TNF-α score, and (G) antigen processing score. In (B) and (E-G), p-values and best-fit slopes were obtained from linear regression models.

We then performed gene set enrichment analysis for the MSigDB Hallmark pathways (Liberzon et al. 2015) to define the functional pathways differentiating the transcriptional signatures of patients (Fig. 4-2D). We identified various immune response pathways significantly associated with severity effects on expression, including the TNF-α signaling via NF-κB pathway in all cell types tested (FDR = 0.58 in $CD4^+$ T cells and FDR < 1×10^{-3} in other cell types) and the

IFN-γ response (FDR < 1 x 10^{-3}), IFN-α response (FDR < 0.02), and inflammatory response (FDR < 1 x 10^{-3}) pathways in $CD14^+$ and $CD16^+$ monocytes. These enrichments were detected among genes more highly expressed in less severe cases, suggesting that patients with milder disease engage a stronger immune response compared to those with more severe disease. To more precisely characterize severity-associated immune response heterogeneity, we computed single-sample gene set enrichment analysis (ssGSEA) scores capturing the activity of functional pathways in each individual across cell types (detailed in Methods). Consistent with our enrichment analyses, respiratory support score was negatively correlated with ssGSEA IFN score (representing the combined IFN-γ and IFN-α response pathways; Pearson r < -0.46, p < 1 x 10^{-3}) (Fig. 4-2E) and TNF-α score (Pearson r < -0.44, p < 0.005) (Fig. 4-2F) in NK cells and $CD14^+$ monocytes. Moreover, we created an antigen (Ag) processing and presentation score based on the corresponding Biological Process gene set (Subramanian et al. 2005), given the previously reported finding that SARS-CoV-2 inhibits the major histocompatibility complex (MHC) class I pathway, a pathway that plays a crucial role in antiviral immunity in lung epithelial cells (Yoo et al. 2021). We found that Ag processing score was positively associated with severity in $CD4^+$ T cells, while the opposite was true for $CD14^+$ monocytes (Pearson r = -0.58, p = 1.2 x 10^{-5}), indicating that pathways associated with antigen presentation are shut down in circulating monocytes specifically (Fig. 4-2G).

To measure the contribution of cell type-specific and disease state-specific genetic variation during the course of an *in vivo* viral infection, we mapped *cis*-eQTL, defined as SNPs located either within or flanking (±100 kilobases) each gene of interest, using the pseudobulk expression estimates for each cell type separately in healthy controls (n = 107, one individual not successfully genotyped) and COVID-19 patients (n = 50, two samples removed as outliers). To

increase our power to detect shared and cell type- or state-specific effects, we utilized a multivariate adaptive shrinkage framework (mash) (Urbut et al. 2019) to leverage correlation structure information within our dataset.

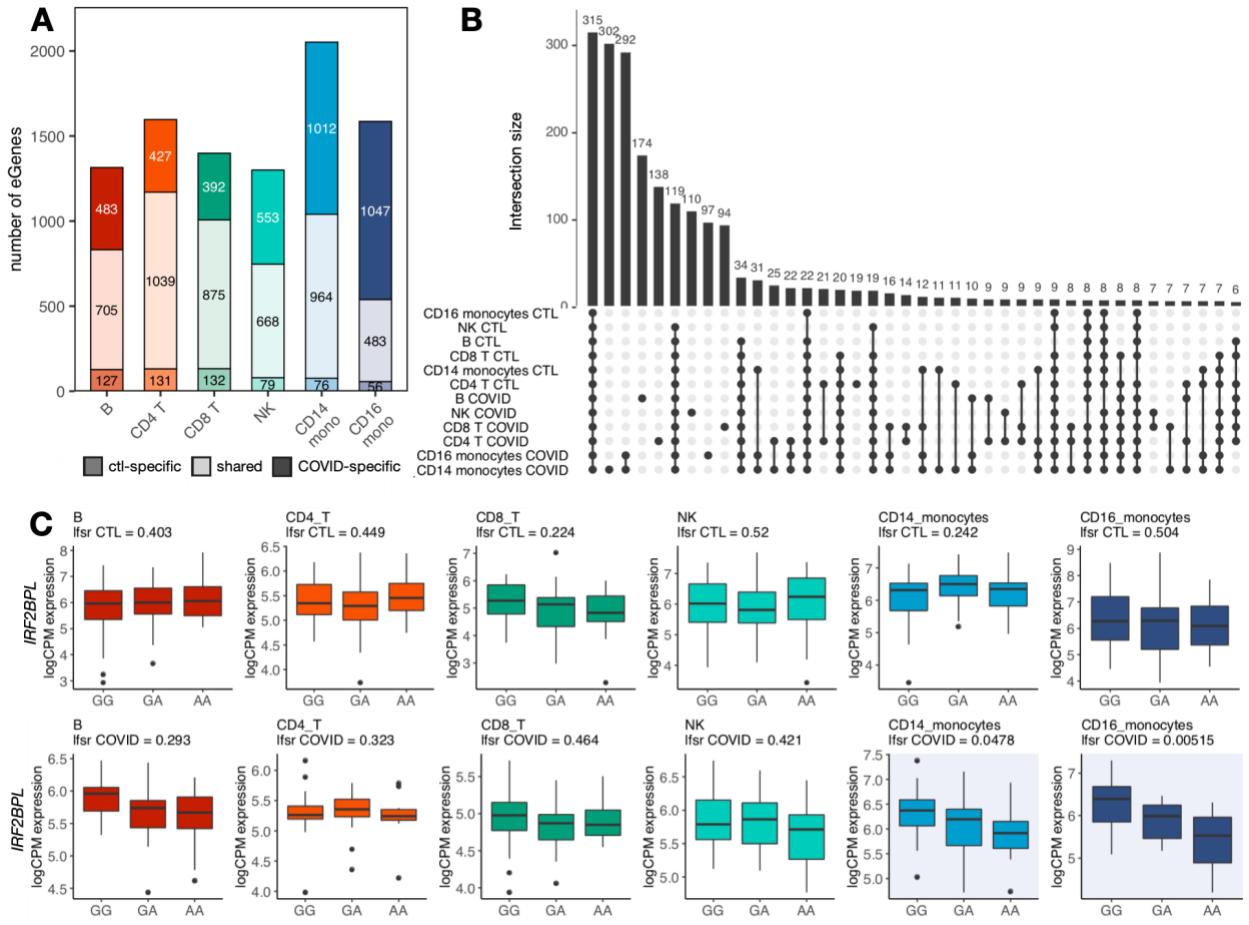


Fig. 4-3. *Cis*-regulatory effects are cell type-specific and disease state-specific. (A) Number of shared and disease state-specific eGenes within cell type. (B) eGene sharing patterns across cell types in control individuals and COVID-19 patients. (C) Example of a patient-specific genetic effect seen only in $CD14^{+}$ and $CD16^{+}$ monocytes in the gene *IRF2BPL* (shaded plots denote lfsr < 0.05).

Across cell types and disease states, we identified 2,928 genes with at least one significant *cis*-eQTL [local false sign rate (lfsr) < 0.10, 39.6% of genes tested; referred to as eGenes] (Fig. 4-3A). NK cells (n eGenes = 1,300) and B cells (n eGenes = 1,315) displayed the fewest number of

genetic effects, while $CD14^+$ monocytes had the greatest (n eGenes = 2,052) (Fig. 4-3A). Generally, most genetic effects were shared between healthy controls and COVID-19 patients within cell type (over 50% in most cell types, Fig. 4-3A); however, in monocytes this pattern did not hold. Instead, many genetic effects were dependent on infection status within $CD16^+$ and $CD14^+$ monocytes (69.5% in $CD16^+$ monocytes, 53.0% in $CD14^+$ monocytes), with the vast majority of these state-specific genetic effects seen only in COVID-19 patients (94.9% in $CD16^+$ monocytes and 93.0% in $CD14^+$ monocytes) (Fig. 4-3A). Indeed, considering sharing patterns across all cell types and disease states (controls vs COVID-19 patients), we detected hundreds of eGenes (n = 691) specific to $CD14^+$ monocytes (n = 302), $CD16^+$ monocytes (n = 97), or both (n = 292) exclusively in COVID-19 patients, suggesting SARS-CoV-2 response eQTL are widespread in monocytes (Fig. 4-3B). One example of a variant falling into this category is the top *cis*-eQTL (rs2363506) for *IRF2BPL*, a gene known to be involved in IFN regulation and shown to be upregulated in cases of severe COVID-19 compared to asymptomatic individuals (Masood et al. 2021), that exhibits a significant genetic effect unique to monocytes of COVID-19 patients (lfsr = 0.048 in $CD14^+$ monocytes of patients, 5.2×10^{-3} in $CD16^+$ monocytes of patients; lfsr > 0.20 in all other cell type-disease state combinations) (Fig. 4-3C).

Finally, given that most of our cell type- and disease state-specific eQTL were discovered in monocytes, we wanted to ask how *cis*-regulatory effects might fluctuate with functional cell states in $CD14^+$ monocytes of COVID-19 patients at the single-cell level. To measure cell state-dependent *cis*-regulatory effects, we used a continuous definition of cell state, which has previously been shown to capture more state-dependent regulatory variation compared to analogous discrete classifications (Nathan et al. 2022). Cell state scores representing various biological processes were derived from the 14 immunological and metabolic Hallmark gene set

pathways (Liberzon et al. 2015). For each pathway, we calculated a numeric score summarizing the activity of that pathway for each single $CD14^+$ monocyte in the patient dataset using ssGSEA (details in Methods). We then implemented a poisson mixed effects interaction model to map single-cell *cis*-eQTL, as this approach has already been used to successfully identify continuous state-dependent eQTL in $CD4^+$ T cells (Nathan et al. 2022). To test for genotype-cell state interactions, we modeled unique molecular identifier (UMI) counts per gene in single cells as a function of genotype at the eQTL variant, controlling for various per-individual and per-cell technical and biological covariates, including age, sex, gene expression PCs, genotype PCs, total UMI count, and percentage of mitochondrial UMIs. We tested only those top gene-SNP pairs with evidence of an eQTL in COVID-19 patients ascertained in the $CD14^+$ monocyte pseudobulk analysis (n = 1,976) for cell state-dependent genotype effects.

Of the 14 immune response and metabolism pathways that we tested, significant cell state-dependent interactions with genotype (likelihood ratio test [LRT] q value < 0.10) were detected in four: IFN-γ response (n interacting eGenes = 33), fatty acid metabolism (n interacting eGenes = 46), oxidative phosphorylation (n interacting eGenes = 94), and TNF-α signaling via NF-κB (n interacting eGenes = 168). The TNF-α signaling via NF-κB pathway stands out as the pathway with the most significant state-dependent eQTL, corresponding to 8.5% of eGenes tested. One of the top TNF-α signaling-dependent eGenes was rs1288848 (LRT q = 0.063), a lead *cis*-eQTL for *C15orf48*, which shows a strong genetic effect in cells with high TNF-α signaling scores (quantiles 5 and 6) but virtually no genetic effect in cells with low scores (quantiles 1 and 2) (Fig. 4-4A).

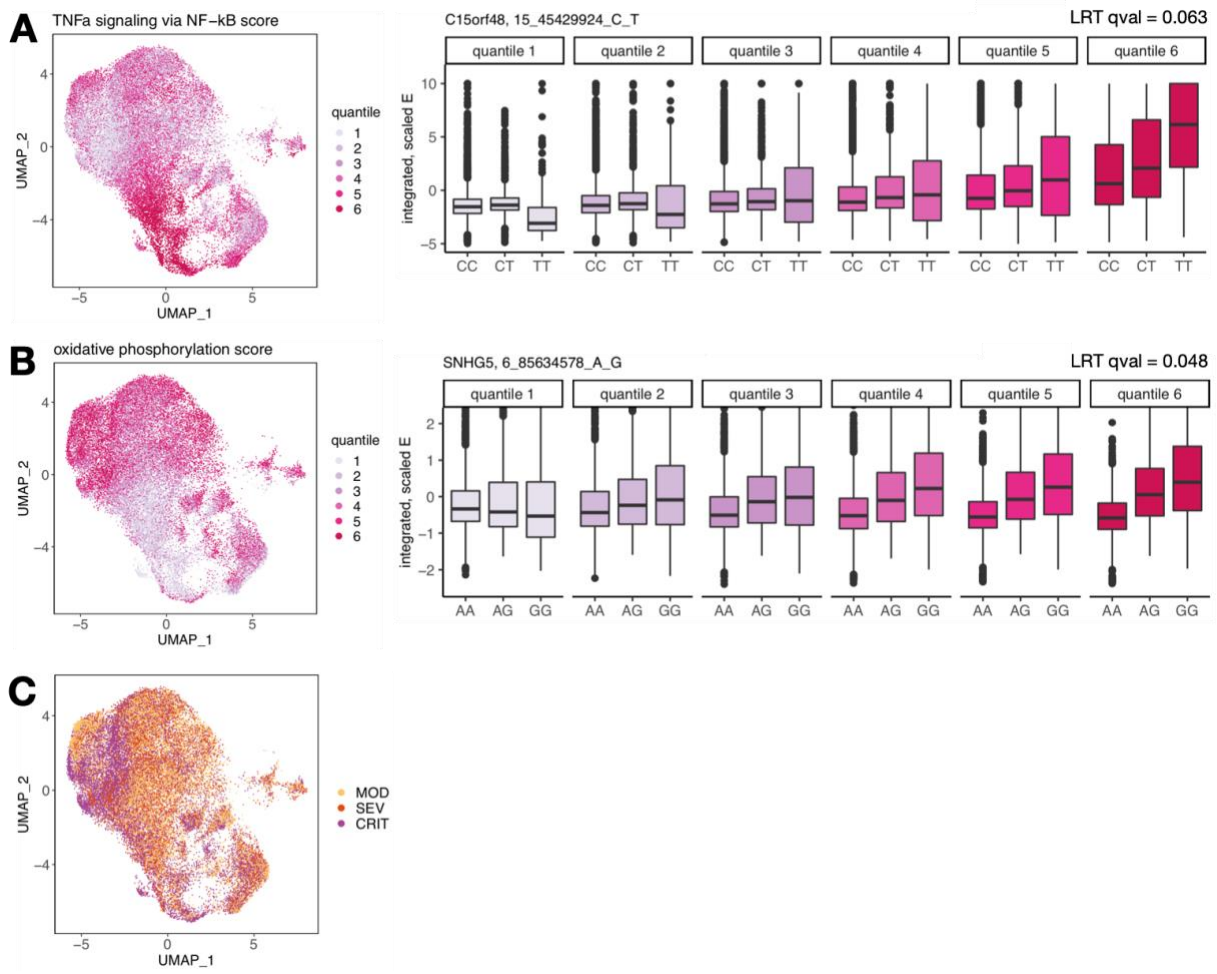


Fig. 4-4. Cell state-dependent eQTL in CD14$^+$ monocytes of COVID-19 patients. (A-B) UMAPs of CD14$^+$ monocytes in COVID-19 patients colored by (A) TNF-α signaling via NF-κB score quantiles and (B) oxidative phosphorylation score quantiles (left), and examples of cell state-dependent eQTL for each of the corresponding functional pathways (right). Single-cell gene expression estimates (y-axis) binned by cell state score quantiles (corresponding to the colors in the UMAP) show that genetic effects are dynamic and vary based on underlying cell state. (C) UMAP of patient CD14$^+$ monocytes colored by disease severity.

Likewise, numerous cell state-interacting eQTL were found for the oxidative phosphorylation pathway, with 6.8% of eGenes tested showing state-dependent genetic variation. The lead eQTL variant (rs2758849) for *SNHG5* displays a significant interaction with oxidative phosphorylation scores (LRT q = 0.048) in the same direction of effect as the TNF-α interacting eGene (i.e., cells with high scores exhibit more prominent genetic effects compared to those with low scores) (Fig.

4-4B). Notably, oxidative phosphorylation scores correlated with patient severity ($p = 7.7 \times 10^{-3}$), with high-scoring cells more often sampled from severe and critical COVID-19 cases (Fig. 4-4C), indicating that functional cell state scores regulated by dynamic *cis* effects capture clinically relevant attributes of patient cohorts to some degree.

Together, this work demonstrates that gene-environment interactions constitute a sizable proportion of the *cis* genetic factors regulating the transcriptional immune response to viral infection *in vivo*. We show that cell type-specific, disease state-specific, and cell state-dependent genetic variation is abundant, affecting 35.3% of all genes tested, particularly in $CD14^{+}$ monocytes. Further, we establish that single cells can harbor distinct eQTL effects that are dependent on their underlying immunological or metabolic functional states and that, in certain cases, these continuous states are associated with clinical features of patients. More broadly, genetic interaction effects likely play a role in dynamically modulating immune responses throughout the course of an infection and may also contribute to differential disease outcomes, especially considering the fact that monocytes, and more generally the myeloid cell compartment, are susceptible to immune dysregulation following SARS-CoV-2 infection (Schulte-Schrepping et al. 2020; Knoll, Schultze, and Schulte-Schrepping 2021). Although we have described how gene-environment interactions shape immune responses in one specific viral infection setting, it is necessary to define how such effects contribute to a wider range of disease states and environmental contexts to better understand the genetic and environmental underpinnings of immune response variation across individuals. As the number of patient cohorts with single-cell phenotyping and genotyping data rise, it will be important to extend this framework to other single-cell eQTL mapping studies to measure the full extent of cell state-dependent regulatory heterogeneity.

Materials and Methods

Participants and samples. We investigated prospectively hospitalized COVID-19 patients between April and September 2020 with symptomatic infection and a positive SARS-CoV-2 nasopharyngeal swab (NSW) polymerase chain reaction (PCR) who were admitted to the Centre Hospitalier de l'Université de Montréal (CHUM) and recruited into the Biobanque Québécoise de la COVID-19 (BQC19) (Tremblay et al. 2021). Patients had no known prior exposure to SARS-CoV-2 (i.e., all infections were primary infections), were not vaccinated at the time of patient sampling, and did not undergo plasma transfer therapy. Blood draws were performed at 11 days post-symptom onset during acute infection ("DSO11" time point, n = 52 samples) and at a later follow-up time point (approximately 6 months after recovery) for a small subset of individuals (n = 12 samples). Additionally, PBMCs collected from healthy control individuals prior to the COVID-19 pandemic living in Montréal (n = 18 samples) and from asymptomatic, SARS-CoV-2 antibody negative uninfected controls (UC) early in the pandemic in Spring 2020 (n = 16) were processed for single-cell data collection in parallel with infected patient samples (detailed below) for a total of 82 samples. The study was approved by the respective IRBs (multicentric protocol: MP-02-2020-8929) and written, informed consent obtained from all participants or, when incapacitated, their legal guardian before enrollment and sample collection. Patients were stratified based on severity of respiratory support at the DSO11 time point: critical patients required mechanical ventilation [noninvasive ventilation, endotracheal intubation, extracorporeal membrane oxygenation (ECMO)], patients with severe disease required oxygen supplementation by nasal cannula, and patients with moderate disease required no supplemental oxygen.

DNA sequencing and imputation. DNA was extracted from whole blood of each sample using the Chemagic™ DNA Blood 400 Kit H96 kit (Perkin Elmer, CMG-1091). SNP genotyping (200 ng input DNA per sample) was conducted using the Axiom™ Precision Medicine Research Array from Applied Biosystems (Applied Biosystems, 902981) per the manufacturer's instructions. The array was processed using the GeneTitan™ Multi-Channel instrument (Applied Biosystems). All samples in the study were grouped together with the Axiom Analysis Suite 5.1.1 software, and the "Best Practice Workflow" analysis was performed using the following high-quality call rate parameters: Axiom_PMRA.r3 library and threshold configuration Human.v5 with minimum call rate of 97.0%. Marker quality control tests were performed on a subset of ancestrally homogeneous participants, who were determined via comparison to 2,504 individuals across 5 super populations from 1000 Genomes Phase 3 data (Auton et al. 2015). We performed batch effect quality control and a replicate discordance check and removed variants that failed either test. Finally, variants with low allele frequencies (MAF < 0.001), low genotyping call rates (< 98%), and that deviated from Hardy-Weinberg equilibrium (HWE) (p-value < $1x10^{-6}$) were removed. Variants retained were considered our high-quality genetic markers. High-quality variants were further filtered and retained if they passed the following additional criteria: MAF > 0.01, marker-wise missingness < 0.01, single nucleotide substitutions with single character allele-codes (A, C, G, or T) (PLINK --snps-only 'just-acgt' option), not in a high linkage disequilibrium (LD) region.

Sample quality filtering was performed considering the set of filtered genotypes described above. Outlier samples with a high genotype missingness rate (overall missing genotype rate > 0.04) or high/low principal component corrected heterozygosity rate on autosomal chromosomes (> ±3SD, respectively, in our cohort) were considered low-quality and removed. We then determined sex chromosome composition by estimating X chromosome marker heterozygosity

using PLINK (--check-sex 0.4 0.7). Since the distribution of X chromosome heterozygosity estimates (F estimates) showed a gap between 0.4 and 0.7, we obtained sex chromosome numbers and compared them to self-reported sex. Individuals with discordant self-reported sex and genetic sex were removed prior to genotype imputation. All other samples passing quality control filters were used for imputation. Genotype phasing and imputation was performed using the Michigan Imputation Server (Das et al. 2016) with the TOPMed reference panel (Taliun et al. 2021).

Whole blood processing. At the time of sampling, whole blood was collected in up to three tubes containing acid citrate dextrose (ACD) and were processed within 6 hours of collection. Blood from the same donor was pooled and centrifuged at 400 g for 10 min at room temperature (RT). After centrifugation, plasma was collected into aliquots of 250 ul, 500 ul, and 1 ml. Aliquots of plasma were kept at -80°C for future use. After plasma collection, the remaining blood was topped up to 30 ml with HBSS medium at RT. Ficoll-Paque separation was then used to isolate PBMCs. PBMCs were washed with R+ (RPMI 1640 + 0.1M HEPES + 20U/ml Penicillin-Streptomycin), resuspended in 5 ml R+ with 10% fetal bovine serum (FBS), and counted at a 1:10 dilution with Trypan blue. Cells were spun down at 400 g for 10 min at 4°C and resuspended in cold FBS at 20 M/ml. A freezing solution of FBS with 20% DMSO was added drop-by-drop to cell suspension while the tube was continuously agitated. Cell suspensions were transferred into cryovials (1 ml/vial), immediately placed into Mr. Frosty Freezing Containers, and stored at -80°C. The following day, PBMCs were transferred to liquid nitrogen for long-term storage.

Sample processing for single-cell RNA-sequencing. PBMCs were unfrozen in batches of three to four samples, rested for 2 hours in RPMI 1640 supplemented with 10% FBS (Corning,

MT35015CV), 2 mM L-glutamine (ThermoFisher Scientific, 25-030-081), and 10 ug/ml gentamicin (ThermoFisher Scientific, 15710064), and immediately processed for single-cell RNA-sequencing. Cells from different samples were pooled per batch for a total of 21 multiplexed batches (n = 82 samples). For each multiplexed cell pool, 12,000 cells were targeted for collection using the Chromium Next GEM Single Cell 3' Reagent (v3.1 Dual Index chemistry) kit (10X Genomics, 1000268). After GEM generation, the reverse transcription (RT) reaction was performed in a thermal cycler as described (53°C for 45 min, 85°C for 5 min), and post-RT products were stored at -20°C for up to one week until downstream processing.

Single-cell RNA-sequencing library preparation and sequencing. Post-RT reaction cleanup, cDNA amplification and sequencing library preparation were performed as described in the Single Cell 3' Reagent Kits v3.1 (Dual Index) User Guide (10X Genomics). Briefly, cDNA was cleaned with DynaBeads MyOne SILANE beads (ThermoFisher Scientific, 37002D) and amplified in a thermal cycler using the following program: 98°C for 3 min, [98°C for 15 s, 63°C for 20 s, 72°C for 1 min] x 11 cycles, 72°C 1 min. After cleanup with the SPRIselect reagent kit (Beckman Coulter, B23317), libraries were constructed by performing the following steps: fragmentation, end-repair, A-tailing, double-sided SPRIselect cleanup, adaptor ligation, SPRIselect cleanup, sample index PCR (98°C for 45 s, [98°C for 20 s, 54°C for 30 s, 72°C for 20 s] x 14 cycles, 72°C 1 min), and double-sided SPRIselect size selection. Prior to sequencing, all multiplexed single-cell libraries (n = 21) were quantified using the KAPA Library Quantification Kit for Illumina Platforms (Roche, 50-196-5234) and pooled in an equimolar ratio. Libraries were sequenced 100 base pair paired-end on an Illumina NovaSeq 6000 to an average depth of 48,613 mean reads per cell across all batches (average median genes detected per cell across batches = 1,627).

Single-cell RNA-sequencing data processing and integration. FASTQ files from each multiplexed capture library were mapped to the pre-built GRCh38 human reference genome (downloaded 10X Genomics) using the cellranger (v6.0.1) count function (G. X. Y. Zheng et al. 2017). souporcell (v2.0, Singularity v3.4.0) (Heaton et al. 2020) in --skip_remap mode (-k 3 or -k 4) was used to demultiplex cells into samples based on genotypes from a common variants file (1000GP samples filtered to SNPs with >= 2% allele frequency in the population, downloaded from https://github.com/wheaton5/souporcell). For each batch, hierarchical clustering of the known genotypes (obtained from DNA-sequencing) and cluster genotypes estimated by souporcell was used to assign individuals to souporcell cell clusters. All samples were successfully assigned to a single set of cells. After demultiplexing, Seurat (v3.1.5, R v3.6.3) (Stuart et al. 2019) was used to perform cell-level quality control filtering. In total, we captured 275,642 cells prior to filtering. High-quality cells were retained for downstream analysis if they had: 1) a "singlet" status called by souporcell, 2) between 500 – 4000 genes detected (nFeature_RNA), 3) a mitochondrial UMI percentage < 20%, and 4) less than 25,000 total molecules (nCount_RNA), leaving 181,348 cells. Gene filtering was performed using CreateSeuratObject min.cells parameter, in which only genes present in at least five cells were kept (n = 28,907 genes).

Due to the large discrepancy between the number of cells assayed in healthy control individuals (n = 38,663) versus COVID-19 patients (n = 142,685) in this dataset, we integrated a publicly available set of high-quality cells derived from control, non-infected individuals (n = 124,976 cells, 90 samples) described in *Chapter III* (Randolph et al. 2021), hereafter referred to as the "non-infected IAV controls". First, we removed IAV-derived transcripts (n = 10 genes) from the raw count matrix of the non-infected IAV controls. Next, we merged both datasets, split the resulting Seurat object by dataset ("COVID" or "IAV controls"), and ran SCTransform

(Hafemeister and Satija 2019) to normalize and scale the UMI counts within dataset. We simultaneously regressed out variables corresponding to experiment batch, percent mitochondrial UMIs per cell, and individual label in both datasets, and additionally, we regressed out infection status in the COVID dataset. We then integrated the data on dataset using the SelectIntegrationFeatures, PrepSCTIntegration, FindIntegrationAnchors, and IntegrateData framework (Stuart et al. 2019). After integration, dimensionality reduction was performed via UMAP (RunUMAP function, dims = 1:30) and PCA (RunPCA function, npcs = 30). A Shared Nearest Neighbor (SNN) Graph was constructed using the FindNeighbors function (dims = 1:20, all other parameters set to default), and clusters were subsequently called using the FindClusters algorithm (resolution = 0.5, all other parameters set to default) (Stuart et al. 2019). In total, our integrated dataset consisted of 306,324 high-quality cells across all samples (n = 181,348 from the COVID dataset, n = 124,976 from the non-infected IAV dataset, n = 172 samples altogether).

Cell type assignment. We performed two versions of cell type annotation: 1) via label transfer to map information about fine-scale cell type populations (e.g., UMAP in Figure 4-1), and 2) via canonical marker gene expression to generate pseudobulk data for the six major cell type clusters ($CD4^+$ T cells, $CD8^+$ T, B cells, NK cells, $CD14^+$ monocytes, $CD16^+$ monocytes). The concordance between label transfer and cell type marker annotations for the major cell type clusters was high (83.1%), in particular for the $CD14^+$ monocytes (97.9% concordance). To perform the label transfer, we downloaded a multimodal human PBMC reference dataset derived from scRNA-seq paired with CITE-seq described in Hao et al. (Hao et al. 2021). We then followed the Seurat v4 Reference Mapping workflow, consisting of the FindTransferAnchors and MapQuery functions, with the Hao et al. reference dataset used as our reference UMAP and the following parameters:

normalization.method = “SCT” and reference.reduction = "spca". For cell type marker annotation, the following marker genes were used to assign clusters to major cell type populations: $CD4^+$ T: *CD3D*$^+$, *CD3E*$^+$, *CD8A*$^–$; $CD8^+$ T: *CD3D*$^+$, *CD8A*$^+$; NK cells: *CD3D*$^–$, *NKG7*$^+$, *GNLY*$^+$; $CD14^+$ monocytes: *CD14*$^+$, *LYZ*$^+$, *FCGR3A*$^–$; $CD16^+$ monocytes: *LYZ*$^+$, *FCGR3A*$^+$; B: *MS4A1*$^+$. In total, we manually annotated 281,525 high-quality cells across all individuals and conditions for downstream analysis (n $CD4^+$ T cells = 111,004, $CD8^+$ T cells = 54,944, $CD14^+$ monocytes = 58,715, $CD16^+$ monocytes = 6,133, B cells = 29,774, NK cells = 20,955).

Calculation of pseudobulk estimates. Pseudobulk estimates were used to summarize single-cell expression values into bulk-like expression estimates within samples. This was performed for all six major cell types ($CD4^+$ T cells, $CD8^+$ T cells, B cells, $CD14^+$ monocytes, $CD16^+$ monocytes, NK cells). Within each cell type cluster for each sample, raw UMI counts were summed across all cells assigned to that sample for each gene using the sparse_Sums function in textTinyR (v1.1.3) (https://cran.r-project.org/web/packages/textTinyR/textTinyR.pdf), yielding an n x m expression matrix, where n is the number of samples included in the study (n = 172) and m is the number of genes detected in the single-cell analysis (m = 28,907) for each of the 6 clusters.

Calculation of residuals for modeling. For each cell type, lowly-expressed genes were filtered using cell type-specific cutoffs (removed genes with a median logCPM < 1.5 in $CD4^+$ T cells and $CD14^+$ monocytes, < 2.0 in B cells and $CD8^+$ T cells, < 3.0 in NK cells, and < 3.5 in $CD16^+$ monocytes), leaving the following number of genes per cell type: $CD4^+$ T cells = 10,206, $CD8^+$ T cells = 9,910, B cells = 10,085, $CD14^+$ monocytes = 10,123, $CD16^+$ monocytes = 9,239, and NK cells = 9,325. Six samples were removed for downstream analysis due to quality control issues,

including one COVID-19 patient sample with a low number of cells (n = 20 cells) and five samples that consistently clustered as outliers on various cell type gene expression PCAs (one COVID-19 patient, one COVID dataset control, and three non-infected IAV controls). After removing lowly-expressed genes, normalization factors to scale the raw library sizes were calculated using calcNormFactors in edgeR (v 3.26.8) (Robinson, McCarthy, and Smyth 2010). The voom function in limma (v3.40.6) (Ritchie et al. 2015) was used to apply these size factors, estimate the mean-variance relationship, and convert raw pseudocounts to logCPM values. The inverse variance weights calculated by voom were obtained and included in the respective lmFit call for all downstream models unless otherwise noted (Ritchie et al. 2015).

Calculation of per-individual ssGSEA scores. To construct the ssGSEA Hallmark pathway scores, we calculated single sample Gene Set Enrichment Analysis (ssGSEA) scores from the pseudobulk COVID-19 patient logCPM gene expression estimates corrected for age, sex, and body mass index (BMI) using the Gene Set Variation Analysis (GSVA, v1.32.0) package in R with default parameters and method = "ssgsea" (Hänzelmann, Castelo, and Guinney 2013). ssGSEA is a method that allows you summarize gene expression patterns for any desired target gene set, and for each sample, it will return a score representative of that gene set. These scores were calculated per cell type, and for each of the pathway-specific ssGSEA scores, the input gene set was derived from either a Hallmark or Gene Ontology (GO) Biological Process gene set (Liberzon et al. 2015) intersected with the set of severity-associated genes for that cell type (described in "Modeling severity effects"). The following gene sets were used to define the per-sample pathway scores: IFN score – combined Hallmark IFN-γ and IFN-α response pathways, TNF-α score – Hallmark

TNF-α signaling via NF-κB pathway, and Ag processing score – GO Biological Process antigen processing and presentation pathway.

Modeling SARS-CoV-2 infection effects. Only healthy controls and COVID-19 patients sampled during the primary infection time point were retained for modeling of infection effects (i.e., follow-up samples were excluded, n = 154). The following linear model was used to identify genes differentially expressed between healthy control individuals and COVID-19 patients:

$$E(i,j) \sim \begin{cases} \beta_0(i) + \beta_{age}(i) \cdot age(j) + \beta_{sex}(i) \cdot sex(j) + \beta_{dataset}(i) \cdot dataset(j) + \varepsilon^{ctl}(i,j) \text{ if condition} = ctl \\ \beta_0(i) + \beta_{COVID}(i) + \beta_{age}(i) \cdot age(j) + \beta_{sex}(i) \cdot sex(j) + \\ \beta_{dataset}(i) \cdot dataset(j) + \varepsilon^{COVID}(i,j) \text{ if condition} = COVID \end{cases}$$

Here, $E(i,j)$ represents the expression estimate of gene *i* for individual *j*, $\beta_0(i)$ is the global intercept accounting for the expected expression of gene *i* in a non-infected female measured in the COVID dataset, and $\beta_{COVID}(i)$ represents the global estimate of the effect of SARS-CoV-2 infection per gene. Age represents the mean-centered, scaled (mean = 0, sd = 1) age per individual, with $\beta_{age}(i)$ being the effect of age on expression levels, sex represents the self-identified sex for each individual (factor levels = “Female”, “Male”), with $\beta_{sex}(i)$ capturing the effect of male sex on expression levels, and dataset represents the dataset in which the sample was obtained (factor levels = "COVID", "IAV controls"), with $\beta_{dataset}(i)$ capturing the effect of a sample originating from the Randolph et al. dataset. Finally, ε^{cdt} represents the residuals for each respective condition (control or COVID) for each gene *i*, individual *j* pair. The model was fit using the lmFit and eBayes functions in limma (Ritchie et al. 2015), and the estimates of the global infection effect $\beta_{COVID}(i)$

(i.e., the differential expression effects due to SARS-CoV-2 infection) were extracted across all genes along with their corresponding p-values. We controlled for false discovery rates (FDR) using an approach analogous to that of Storey and Tibshirani (Nédélec et al. 2016; Storey and Tibshirani 2003), which derives the distribution of the null model empirically. To obtain a null, we performed 10 permutations, where infection status label (i.e., control/COVID) was permuted across individuals. We considered genes significantly differentially expressed upon infection if they had β_{COVID} $|\log_2 FC| > 0.5$ and an FDR < 0.05.

Modeling severity effects. To model the effect of COVID-19 disease severity on gene expression, we restricted our analyses to COVID-19 patients sampled during the primary infection time point for which we had information about disease severity (n = 50). Severity of COVID-19 was assessed using a five-point scale of respiratory support needed at the time of patient sampling that includes the following categories: 0/Moderate = no supplemental oxygen (n = 17); 1/Severe = nasal cannula (n = 11); 2/Critical = non-invasive ventilation (n = 4); 3/Critical = intubation (n = 17); 4/Critical = extracorporeal membrane oxygenation (ECMO) (n = 1). The following model was used to evaluate the effect of severity at the time of patient sampling on expression:

$$E(i,j) \sim \beta_0(i) + \beta_{severity}(i) \cdot severity(j) + \beta_{age}(i) \cdot age(j) + \beta_{sex}(i) \cdot sex(j) + \beta_{BMI}(i) \cdot BMI(j) + \varepsilon(i,j)$$

Here, $E(i,j)$ represents the expression estimate of gene *i* for individual *j*, $\beta_0(i)$ is the global intercept accounting for the expected expression of gene *i* in a female COVID-19 patient, and $\beta_{severity}(i)$ indicates the effect of severity on gene *i* during the primary sampling time point. Severity ($severity(j)$) represents the respiratory support score per individual and was treated as a numeric variable. Body mass index (BMI) represents the mean-centered, scaled (mean = 0, sd = 1) BMI

per individual, with $\beta_{BMI}(i)$ being the effect of BMI on expression levels. If BMI was not reported for an individual (n missing = 17), this missing data was filled with the average BMI across patients. All other terms in the model are analogous to that described in "Modeling SARS-CoV-2 infection effects". The model was fit using the lmFit and eBayes functions in limma (Ritchie et al. 2015), and the estimates of $\beta_{severity}(i)$ were extracted across all genes along with their corresponding p-values. We again controlled for false discovery rates (FDR) by empirically deriving the null distribution. To obtain a null, we performed 10 permutations, where respiratory support score (i.e., 0 - 5) was permuted across patients. We considered genes significantly correlated with disease severity if they had an FDR < 0.05.

Gene set enrichment analyses. The R package fgsea (v1.10.1) (Korotkevich et al. 2021) was used to perform gene set enrichment analysis for the severity effects using the H hallmark gene sets (Subramanian et al. 2005). Ranked t-statistics for each cell type were obtained directly from the topTable function in limma (Ritchie et al. 2015), and the background set for a cell type was the sets of genes sufficiently expressed (i.e., passed the lowly-expressed gene filter threshold) for that cell type. Pre-ranked t-statistics were used to perform the enrichment using fgsea with the following parameters: minSize = 15, maxSize = 500, nperm = 100000. Enrichments scores (ES) and Benjamini-Hochberg adjusted p-values output by fgsea were collected for each analysis.

eQTL mapping and integration with mashr. eQTL mapping was performed in each cell type using the integrated pseudobulk expression data. A linear regression model was used to ascertain associations between SNP genotypes and expression levels. Input expression matrices were quantile-normalized within condition prior to association testing. eQTL were mapped separately

in healthy control individuals and in COVID-19 patients sampled at the first time point (i.e., follow-ups were excluded) using the R package MatrixEQTL (v2.3) (Shabalin 2012). Prior to mapping, SNPs were filtered using the following criteria in the COVID-19 dataset and the Randolph et al. dataset separately: 1) keep those with a minor allele frequency > 5% across all individuals, 2) exclude those with > 10% of missing data, and 3) exclude those that deviate from Hardy-Weinberg equilibrium at $p < 10^{-5}$ (--maf 0.05 --geno 0.10 --hwe 0.00001 PLINK v1.9 filters) (Chang et al. 2015). Only SNPs that passed these filters and were present in both datasets were retained and merged across datasets (n = 4,387,963 SNPs kept). Local associations (i.e., putative *cis*-eQTL) were tested against all SNPs located within the gene body and 100kb upstream and downstream of the transcription start site (TSS) and transcription end site (TES) for each gene tested.

We accounted for unmeasured surrogate confounders by performing PCA on a correlation matrix based on the gene expression data. Subsequently, up to 15 principal components (PCs) were regressed out prior to performing the association analysis for each gene. A specific number of PCs to regress in each cell type-condition pair, corresponding to the number of PCs that empirically led to the detection of the largest number of eQTL in each condition, was then chosen empirically (shown in Table S4-1). To avoid spurious associations resulting from population structure, the first two eigenvectors obtained from a PCA on the genotype data using SNPRelate (v1.20.1, gdsfmt v1.22.0) (X. Zheng et al. 2012) were included in the linear model. Other covariates included were age (mean-centered, scaled), sex, and number of cells detected per sample, and additionally, dataset when mapping eQTL using the healthy control expression values.

To gain power to detect *cis*-eQTL effects using sharing information across cell types, we implemented mashr (Urbut et al. 2019). We considered shared genes that were tested across all

cell types (n = 7,403). For each of these genes, we chose a single, top *cis*-SNP, defined as the SNP with the lowest FDR across all cell types (n = 6) and conditions (n = 2), to input into mashr. We extracted the effect sizes and computed the standard errors of these betas from the Matrix eQTL outputs for each gene-SNP pair across cell types and conditions. We defined a set of strong tests (i.e., the 7,403 top gene-SNP associations) as well as a set of random tests, which we obtained from randomly sampling 200,000 rows of a matrix containing all gene-SNP pairs tested merged across conditions. The mashr workflow was as follows: i) the correlation structure among the null tests was learned using the random test subset, ii) the data-driven covariance matrices were learned using the strong test subset (from 5 PCs), iii) the mash model was fit to the random test subset using canonical and data-driven covariance matrices, and iv) the posterior summaries were computed for the strong test subset. We used the local false sign rate (lfsr) to assess significance of our posterior eQTL effects and considered a gene-SNP pair to have a significant eQTL effect if the lfsr was < 0.10.

Calculation of functional cell state scores per cell. To obtain the cell state scores used for modeling cell state-dependent single-cell eQTL, first, the raw single-cell UMI counts across all samples were obtained per cell type. All subsequent processing steps were performed for each cell type independently. Raw per cell counts in the form of a Seurat object were split by dataset, and SCTransform was used to normalize and scale the UMI counts within dataset, regressing the effects of experiment batch, percent mitochondrial UMIs per cell, and age in both datasets, and additionally, sex in the COVID-19 dataset. The SelectIntegrationFeatures, PrepSCTIntegration, FindIntegrationAnchors, and IntegrateData pipeline was then used to integrate cells, returning all features following integration (features.to.integrate = all_features) (G. X. Y. Zheng et al. 2017).

The scaled data matrix (@scale.data slot) of the integrated data, which holds the residuals of the corrected log-normalized integrated counts, was obtained, and these values were used to calculate ssGSEA scores (using the same parameters described above in "Calculation of per-individual ssGSEA scores") per cell for pathways of interest. Unlike the ssGSEA scores described above, which were calculated for each pseudobulk sample, here, we applied ssGSEA to the scaled SCTransform gene x cell matrix, allowing us to generate cell state scores for each single cell in the dataset. The pathways of interest included the immune-related and metabolism-related pathways in the MSigDB Hallmark gene sets (n = 14) (Liberzon et al. 2015).

Modeling cell state-genotype interaction effects. We used a Poisson mixed effects (PME) model to test for cell state-dependent eQTL because this model has previously been demonstrated to detect significant cell state-genotype interaction effects in single-cell data (Nathan et al. 2022). Only COVID-19 patients sampled during the primary infection time point were included in these analyses (n = 50), and only gene-SNP pairs in which we had evidence of a significant eQTL (lfsr < 0.10) in the $CD14^+$ monocyte pseudobulk analysis in COVID-19 patients were tested (n = 1,976). Further, single-cell eQTL modeling was restricted to the $CD14^+$ monocyte single-cell expression estimates. To control for genetic background and latent confounders, we included both genotype and expression PCs in our single-cell cell state eQTL models. We computed genotype PCs using the same approach as above in "eQTL mapping and integration with mashr". Expression PCs were calculated from non-batch corrected integrated and scaled counts using the same method as described in "Calculation of functional state scores per cell," but omitting the batch correction step (i.e., no variables were regressed in the SCTransform call). PCA was run on the cell x gene matrix

of non-corrected integrated and scaled counts subset on the top 3,000 variable features using the prcomp_irlba function in the R package irlba (v2.3.5.1) (Baglama, Reichel, and Lewis 2022).

To test for interactions with cell state, we used the following Poisson mixed effects interaction model, where each gene's UMI counts were modeled as a function of genotype and additional donor-level and cell-level covariates. For each gene:

$$\log(E_i) \sim \beta_0 + \beta_G X_{d,G} + \beta_{age} X_{d,age} + \beta_{sex} X_{d,sex} + \beta_{nUMI} \log(X_{i,nUMI}) + \beta_{MT} X_{i,MT} + \sum_{k=1}^{3} \beta_{gPC_k} X_{d,gPC_k} + \sum_{k=1}^{5} \beta_{ePC_k} X_{i,ePC_k} + \beta_{cell\ state} X_{i,cell\ state} + \beta_{G\ x\ cell\ state} X_{d,G} X_{i,cell\ state} + (\phi_d \mid d) + (\kappa_b \mid b) + \varepsilon$$

Here, E is the expression of the gene in cell i, β_0 is an intercept, and ε represents the residuals. All other βs represent fixed effects for various covariates in cell i, donor d, or experimental batch b as follows: G = genotype at the eQTL variant, *age* = scaled age of donor, *sex* = sex of donor, *nUMI* = number of UMI per cell (accounts for sequencing depth), *MT* = percent of mitochondrial UMIs per cell, *gPC* = genotype PCs, *ePC* = single-cell expression PCs prior to batch correction, and *cell state* = functional cell state score per cell (described above). Donor was modeled as a random individual effect ($\phi_d \mid d$) to account for the fact that multiple cells were sampled per individual, and experimental batch was also modeled as a random effect ($\kappa_b \mid b$). Finally, $\beta_{G\ x\ cell\ state} X_{d,G} X_{i,cell\ state}$ represents the cell state x genotype interaction term of interest.

Single-cell poisson mixed interaction models were fit using the glmer function in the lme4 R package (v 1.1-29) with the following parameters: family = "poisson", nAGQ = 0, and control = glmerControl(optimizer = "nloptwrap") (Bates et al. 2015). To determine significance, we used a likelihood ratio test (LRT) comparing two models, one with and one without the cell state interaction term and calculated a p-value for the test statistic against the Chi-squared distribution

with one degree of freedom. To correct for multiple hypothesis testing, we performed one permutation in which cell state scores were permuted across all cells per pathway tested, and we obtained a null LRT p-value distribution using the same framework as above with our permuted data. We then calculated q-values for the cell state-genotype interaction term using the empirical p-value distribution across all tested eQTL using the empPvals and qvalue functions from the qvalue package (v2.16.0) (Storey et al. 2023) in R.

Supplementary Tables

Supplementary tables for this chapter are included in Appendix A: Supplementary Figures and Tables.

Chapter V: Discussion

Together, these projects represent an extensive exploration into human immune response variation after pathogenic infection. This work examines the gene regulatory patterns that underlie response heterogeneity and establishes the *cis*-genetic factors that shape between-individual and between-population variation in the early response to viral infection across various immune cell types both *in vitro* and *in vivo*. We used a unique combination of single-cell RNA-sequencing paired with immunological assays to profile the transcriptomes of thousands of cells across hundreds of individuals and linked transcriptional variation with pertinent functional immune response heterogeneity. Specifically in the context of influenza A exposure, we showed that: 1) population-associated variation in gene expression is common and often cell type-specific, 2) individuals better able to mount an interferon response early in infection were also better able to restrict viral replication later in infection, and 3) a signature of polygenic selection for ribosomal protein genes exists and is supported by functional genomic data, all previously undiscovered findings. Further, we also demonstrated that *cis*-regulatory variants explain a substantial proportion of population heterogeneity in the transcriptional immune response to SARS-CoV-2 and that a large fraction of *cis*-eQTL operate in cell type-specific, condition-specific, and cell state-specific manners in hospitalized COVID-19 patients. While these contributions substantially advance the field, this work is not exhaustive and raises many follow-up questions.

Interferon response diversity

Although we detect population-associated variation in the interferon response following influenza A infection (*Chapter III*), further investigation is needed to define the broader basic and clinical implications of this observation. The interferon response is a critical pathway induced upon the detection of viral pattern recognition receptors in vertebrates, and defects in this response due to monogenic mutations in interferon-related genes lead to dysregulated immune responses and severe viral infections (Rodero and Crow 2016; Dupuis et al. 2003). It has been shown repeatedly across nature that the interferon response is one of the most diverged pathways between human populations (*Chapter III*) and between species (Iwama and Moran 2023; Snyder-Mackler et al. 2016). Yet, this divergence seems contradictory, as deleterious mutations in the type I interferon signaling pathway are very rare and often lead to severe susceptibility to viral disease (Dupuis et al. 2003). This paradox invites one to speculate many questions worthy of further study: why does such a crucial pathway show significant variation across people, and what are the upper and lower extremes tolerated in healthy individuals? Is there an advantage to having plasticity in the interferon response, and, if so, why and what causes this? How does the relative importance of timing versus strength of the interferon response vary across the course of viral infection and between individuals?

We find that a stronger, early induced interferon response is associated with better viral clearance at later time points *in vitro* (*Chapter III*); however, how or whether this translates to the clinic must be formally tested. Various studies suggest that a delayed and suppressed interferon response early in infection is associated with severe disease progression (J. S. Lee and Shin 2020; Galbraith et al. 2022). Other conflicting studies suggest the existence of an "interferon paradox," in which robust, sustained interferon signaling is associated with poor outcome and the

establishment of persistent viral infection (J. S. Lee and Shin 2020; Galbraith et al. 2022; Papatriantafyllou 2013). Direct follow-up studies to address how interferon response strength and kinetics vary across individuals and across time are warranted, particularly in the context of how these phenotypes correlate with viral disease severity and outcome. We do not observe a strong *cis*-genetic signal driving the observed population variation in the interferon response early after infection, suggesting that *trans*-genetic or environmental effects may play a role. It is of great interest to ascertain the causal factors leading to variation in this antiviral response, as doing so may result in clinically actionable findings.

Variation in adaptive immunity

The work described in *Chapters III* and *IV* and much of the immune response eQTL literature have focused primarily on the response of innate immune cells to experimental challenges. One advantage of the study designs described in *Chapters III* and *IV* is that they inherently captured responses of mixed populations of immune cells, specifically PBMCs, meaning that the signals of both innate (monocytes, NK cells) and adaptive (T cells, B cells) immune cells were measured and reported. Although we included adaptive immune cells in our analyses, we used relatively early time points post-infection (6 hours following IAV infection *in vitro*; approximately 11 days after symptom onset in COVID-19 patients with no past history of exposure) to assay transcriptional responses in both studies. Because of this, we were limited in our conclusions considering variation in the adaptive immune response. Experiments tailored to measure population-associated variation in T and B cell responses are needed to supplement these findings. Specifically, studies that seek to characterize transcriptional variation as well as T-cell

and B-cell receptor repertoire diversity across individuals and populations will allow us to discern the extent of adaptive response heterogeneity. Further, studies that exploit the interconnected nature of the immune system will allow us to more comprehensively assess how variation in innate immunity impacts adaptive immunity, how variation in cell-cell interactions and paracrine signaling dictate response differences, and how both arms of the immune system work together to give rise to an overall phenotype.

Dynamic immune response variation and the utility of iPSCs

A natural extension of surveying variation in the immune response at longer time scales is investigating the dynamics of gene regulation throughout an immune response, or measuring how gene regulatory effects change over time following infection. Most immune response eQTL studies to date have ignored the dynamic nature of the immune response, as they have only probed gene regulatory patterns at a single time point. To study the dynamics of gene regulation, experiments in which time series gene expression data are generated at detailed temporal resolution are needed. The objective of these studies would be to map quantitative trait loci that show an interaction with time (e.g., genetic effects that only appear early or late in the response) or identify other quantitative trait loci associated with immune response trajectory (e.g., overall magnitude of the immune response or the global maximum immune response).

Dense time course experiments are difficult to perform with primary human cells and tissues due to irregularities in patient sampling and availability, potentially leading to the introduction of major batch effects. Historically, cell lines have been used to mitigate these issues; however, cell lines do not always faithfully recapitulate features of their primary cell counterparts,

especially considering karyotype and cell marker expression (Geraghty et al. 2014), and they cannot match the genetic diversity represented across individuals because they often stem from a single donor. More recently, researchers have turned to induced pluripotent stem cells, or iPSCs, as they represent an alternative source of biological material that overcomes these problems. iPSCs are cells that have been reprogrammed from adult somatic cells (e.g., fibroblasts) into an induced state of pluripotency and self-renewal (Yu et al. 2007; Okita et al. 2011). These cells are advantageous because they can self-renew indefinitely, can feasibly be generated from any individual, and, theoretically, have the capacity to undergo directed differentiation into any cell type present in the three primary germ layers of the human body.

Because of these properties, iPSCs are an attractive model to study tissue types that are difficult to obtain as primary samples, such as tissue-derived immune cells. Further, they allow for the design of more complex studies (e.g., a multiple time point time course in which hundreds of thousands of cells would be needed for each time point) because cell cultures can be scaled up (Strober et al. 2019), which is difficult or impossible to do with primary immune cells. They are also an excellent model to study the genetic basis of complex human traits as they preserve phenotypic differences between individuals. Indeed, 5-46% of the variation in iPSC phenotypes arises from differences between individuals, and many of these phenotypes can be mapped to specific loci, which demonstrates their utility as a powerful model for immune response eQTL studies (Kilpinen et al. 2017; Banovich et al. 2018). Notably, immune cells derived from iPSCs are a stable, renewable source. Because of this, large-scale iPSC banks could be established to study immune responses and their genetic determinants reliably and reproducibly. Therefore, studies that seek to examine immune response dynamics across individuals would benefit from relying on iPSCs as a resource.

CRISPR screens to investigate polygenic selection

The advent of CRISPR/Cas9 as a genome editing tool, along with its many variants that enable the precise editing of regulatory and epigenomic landscapes in a high-throughput manner, has revolutionized the genomics field. CRISPR screens have been widely used to introduce genetically-encoded perturbations in pools of target cells, which are then challenged with an experimental pressure (e.g., drug treatment), to reveal mutations that confer resistance or susceptibility to the challenge (Bock et al. 2022). While it is possible to perform CRISPR editing in primary cells, as has been described in primary human T cells (Freimer et al. 2022), its efficacy is often limited in other primary immune cells due numerous factors, including the reliance of editing on homology-directed repair, a process that is inefficient in nondividing cells (Nami et al. 2018), the difficulties of transducing primary cells (Burke 2003), and the availability of a sufficient number of primary cells to target. Recently, the success of the CRISPR/Cas9 system has been extended to human iPSCs (De Masi et al. 2020), opening up the range of cell types that can be targeted. Pooled CRISPR screens in CRISPR-edited iPSC-derived immune cells followed by pathogen challenge unlocks the possibility of probing coding and regulatory elements in disease-relevant cell types to identify and causally test whether perturbation results in significant changes to expression pattens or relevant downstream phenotypes, such as sensitivity to infection.

Pooled CRISPR screens can also be used to investigate and define *trans*-regulatory networks (Freimer et al. 2022). The work described in *Chapters III* and *IV* largely ignores the existence of any *trans*-effects on gene expression patterns, although it is an important consideration. Identifying *trans*-eQTL has proven to be difficult because very large sample sizes are required to map these loci, partly due to their much smaller effect sizes compared to *cis*-eQTL (Võsa et al. 2021). An alternative method to detecting *trans*-regulated genes involves the use of

CRISPR editing to experimentally edit genes of interest and subsequently measure changes in expression of other genes. This approach relies on the "omnigenic model" hypothesis, which proposes that gene regulatory networks are highly interconnected and that most genes expressed in disease-relevant cells affect disease risk via network effects (Boyle, Li, and Pritchard 2017). Through CRISPR perturbation, both upstream and downstream regulators of disease genes can be defined, and thus, *cis*- and *trans*-regulatory relationships uncovered in a targeted way.

In *Chapter II*, we reviewed how natural selection has acted on immune response phenotypes, with a particular focus on signatures of positive selection at immune-related loci in the genome. Traditional tests of positive selection, such as the integrated haplotype score (iHS) or F_{ST} (Voight et al. 2006; P. C. Sabeti et al. 2006), rely on outlier approaches to pinpoint selected regions in the genome. Such selection scans fail to detect signatures of polygenic selection, which occurs when natural selection acts on polygenic phenotypes and drives subtle shifts in allele frequencies across many loci (Berg and Coop 2014). Current tests for polygenic selection are subject to confounding factors (e.g., population structure and/or environmental effects), suffer from effect size misestimation (Berg and Coop 2014), or rely on *a priori* information about gene sets involved in particular processes (*Chapter III*). In *Chapter III*, we identify a putative signature of polygenic selection linked to expression levels of ribosomal protein genes using pre-defined gene ontology gene sets, which is not ideal. However, CRISPR-mediated *trans*-regulatory network discovery bypasses the need to rely on gene set information or GWAS data to test for polygenic selection. Once a *trans* network is defined, within network genes, we can test whether trait-increasing or trait-decreasing alleles are overrepresented among individuals with high or low trait values, respectively. In theory, this approach would allow one to identify signatures of polygenic selection among many molecular quantitative traits in a less biased way.

Expanding the global diversity of immunogenomics studies

One of the most critical gaps in biomedical research today is the lack of non-European ancestry individuals in genomics studies, specifically among cohorts designed to characterize immune variation among healthy individuals in the general population. Indeed, most genomics studies to date – approximately 86% – have been conducted solely in individuals of European descent (Fatumo et al. 2022). In *Chapter III*, we focus on characterizing population-associated variation in gene expression between individuals of African and European genetic ancestry living in the United States. A natural extension of this work is surveying a more extensive set of populations throughout various environments globally. Already, some studies have sought to expand the catalog of diversity in human genomics research, such as the Human Genome Diversity Panel and the 1000 Genomes Project (Bergström et al. 2020; Auton et al. 2015). Through these efforts, previously uncharacterized common genetic variants unique to regions outside of Europe have been discovered, highlighting the value of diversifying the global populations included in such studies. In addition, conducting GWAS in ancestrally diverse populations has increased the accuracy and applicability of genomics research, enabling us to better define linkage disequilibrium patterns, characterize differences in allele frequencies between populations, calculate polygenic risk scores, and perform fine mapping (Peterson et al. 2019; Hindorff et al. 2018; Asimit et al. 2016; A. R. Martin et al. 2019; Duncan et al. 2019).

Increasing representation of global populations in genomics studies that address immune response variation will bolster our awareness of the functional heterogeneity that exists today. Specifically, extending population studies of immune responses to a larger array of genetic backgrounds may reveal differences in the frequencies of disease-relevant alleles and population-associated variation in immune gene regulation. Investing in the sampling of diverse cohorts will

also likely advance our understanding of disease etiology, with the goal of making genomics research more broadly applicable to ensure an equitable distribution of the purported benefits promised by personalized medicine. Of note, precautions must be taken to safeguard individuals in historically underrepresented populations in scientific research from exploitation. Research projects that: 1) clearly define how donor samples will be collected, used, and/or banked through proper informed consent, 2) engage with communities through the sharing of research results and data, and 3) involve local scientists and institutes in the research efforts should be the standard to increase inclusivity and protect against the misuse of data and samples.

The influence of the environment

Although the work described herein primarily focuses on identifying the genetic drivers of between-individual variation in immune responses, another question that arises is how environmental factors play a role in shaping this variation. In *Chapter III*, we found that lead *cis* variants explain approximately 53% of the variance in the observed population differences in gene expression that we detect, suggesting that a considerable *cis* genetic component drives these differences. While a large proportion of this variance can be accounted for, the other half remains unexplained and is likely controlled by either *trans* genetic or environmental determinants. Because we define genes differentially expressed between populations as those with expression levels significantly correlated with global genetic ancestry estimates, it stands to reason that a significant fraction of the signal we identify is driven by environmental confounders that are correlated with quantitative genetic ancestry estimates, not genetic associations themselves. In the same vein, it has been shown that combining genomic data with self-reported ancestry labels and

electronic health records is valuable to measure fine-scale population structure that, in some cases, is linked to shared culture and environment (Belbin et al. 2021). Using this approach, the prediction of complex disease risk can be improved within fine-scale groups (Belbin et al. 2021). Therefore, a better understanding of both genetics and environment is needed to push the field of genomic medicine forward, and defining the relative contribution of genetics versus environment to immune response variation is of great interest.

Environmental factors such as socioeconomic status, diet, stress levels, access to healthcare, quality of treatment in the healthcare system, past and recurring exposures to pathogens, and vaccination history, among many others, are known to significantly influence acute and chronic disease severity and outcome (J. Y. Ko et al. 2021; Chandrasekhar et al. 2017; Raifman and Raifman 2020; Eisenberg et al. 2007; Sears and Genuis 2012). Many of these non-genetic aspects have been ignored in genomic studies of common diseases for various reasons, for example because it may not have been the predominant goal of the study to measure these effects or because more complex study designs are required to collect this type of meta data alongside genetic data, e.g., surveys designed to capture self-reported responses as accurately and completely as possible. This is an oversimplification, as these environmental factors likely represent unmeasured sources of variation that are clinically relevant and interesting from a biological standpoint in most human immune response studies. While some environmental correlates of immune responses are easier to measure accurately (e.g., antibody titers) than others (e.g., socioeconomic status), most are still incomplete approximations of an individual's overall environmental context. Therefore, immunogenomics studies designed to better integrate environmental data and estimate environmental effects are necessary.

In *Chapter III*, we attempted to partially address this limitation by considering prior exposure to a specific strain of influenza A virus. We relied on molecular assay data, specifically H1N1 Cal/04/09 IgG antibody titers measured from serum, as a best available proxy for past immunization and exposure in the absence of detailed vaccination or infection history records for our donors. Although we found that antibody titers were not correlated with genetic ancestry or interferon response among individuals in our study, suggesting that variation in past exposure to that specific influenza A strain did not influence our main conclusions, other studies designed to address this are nevertheless worthwhile. To more completely account for prior infection and vaccination history, methods such as VirScan (Shrock, Shrock, and Elledge 2022) can be applied. This technique enables high-throughput antibody profiling and relies on screening sera samples against a library of peptides encompassing the human viral proteome to identify viral peptides recognized by an individual's antibodies (Shrock, Shrock, and Elledge 2022). Using this information paired with other genomic tools, we can ask questions such as how past viral exposures shape variation in the response to future pathogen challenges and whether increased immune cell repertoire diversity correlates with pathogen burden.

To capture the effects of more complex environmental signals on immune response variation, we can collect primary samples from multiple populations (e.g., African-ancestry and European-ancestry individuals) living in different locations (e.g., the United States, Europe, and Africa), aiming to keep genetic ancestry relatively constant while varying environment. Even with such study designs, pinpointing the exact environmental factors driving immune response variation will still be challenging due to confounders. Going forward, it will be crucial to prioritize immunogenomics studies that not only cover a greater range of populations globally but that also

thoroughly assess clinically relevant environmental variables. Only then can we begin to actualize the full potential of genomic medicine and precision healthcare.

References

Abadie, Valérie, Ludvig M. Sollid, Luis B. Barreiro, and Bana Jabri. 2011. "Integration of Genetic and Immunological Insights into a Model of Celiac Disease Pathogenesis." *Annual Review of Immunology* 29 (1): 493–525. https://doi.org/10.1146/annurev-immunol-040210-092915.

Aguet, François, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, et al. 2017. "Genetic Effects on Gene Expression across Human Tissues." *Nature* 550 (7675): 204–13. https://doi.org/10.1038/nature24277.

Alasoo, Kaur, Julia Rodrigues, Subhankar Mukhopadhyay, Andrew J. Knights, Alice L. Mann, Kousik Kundu, Christine Hale, Gordon Dougan, and Daniel J. Gaffney. 2018. "Shared Genetic Effects on Chromatin and Gene Expression Indicate a Role for Enhancer Priming in Immune Response." *Nature Genetics* 50 (3): 424–31. https://doi.org/10.1038/s41588-018-0046-7.

Alexander, David H., and Kenneth Lange. 2011. "Enhancements to the ADMIXTURE Algorithm for Individual Ancestry Estimation." *BMC Bioinformatics* 12 (1): 246. https://doi.org/10.1186/1471-2105-12-246.

Alvarez, Monica I., Luke C. Glover, Peter Luo, Liuyang Wang, Elizabeth Theusch, Stefan H. Oehlers, Eric M. Walton, et al. 2017. "Human Genetic Variation in VAC14 Regulates Salmonella Invasion and Typhoid Fever through Modulation of Cholesterol." *Proceedings of the National Academy of Sciences*, August, 201706070. https://doi.org/10.1073/pnas.1706070114.

Aquino, Yann, Aurélie Bisiaux, Zhi Li, Mary O'Neill, Javier Mendoza-Revilla, Sarah Hélène Merkling, Gaspard Kerner, et al. 2022. "Environmental and Genetic Drivers of Population Differences in SARS-CoV-2 Immune Responses." bioRxiv. https://doi.org/10.1101/2022.11.22.517073.

Ashley, Euan A. 2016. "Towards Precision Medicine." *Nature Reviews Genetics* 17 (9): 507–22. https://doi.org/10.1038/nrg.2016.86.

Asimit, Jennifer L., Konstantinos Hatzikotoulas, Mark McCarthy, Andrew P. Morris, and Eleftheria Zeggini. 2016. "Trans-Ethnic Study Design Approaches for Fine-Mapping." *European Journal of Human Genetics* 24 (9): 1330–36. https://doi.org/10.1038/ejhg.2016.1.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Baglama, Jim, Lothar Reichel, and B. W. Lewis. 2022. “Irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices.” https://CRAN.R-project.org/package=irlba.

Bakker, Olivier B., Raul Aguirre-Gamboa, Serena Sanna, Marije Oosting, Sanne P. Smeekens, Martin Jaeger, Maria Zorro, et al. 2018. “Integration of Multi-Omics Data and Deep Phenotyping Enables Prediction of Cytokine Responses.” *Nature Immunology* 19 (7): 776–86. https://doi.org/10.1038/s41590-018-0121-3.

Banovich, Nicholas E., Yang I. Li, Anil Raj, Michelle C. Ward, Peyton Greenside, Diego Calderon, Po Yuan Tung, et al. 2018. “Impact of Regulatory Variation across Human IPSCs and Differentiated Cells.” *Genome Research* 28 (1): 122–31. https://doi.org/10.1101/gr.224436.117.

Barreiro, Luis B., Meriem Ben-Ali, Hélène Quach, Guillaume Laval, Etienne Patin, Joseph K. Pickrell, Christiane Bouchier, et al. 2009. “Evolutionary Dynamics of Human Toll-Like Receptors and Their Different Contributions to Host Defense.” *PLOS Genet* 5 (7): e1000562. https://doi.org/10.1371/journal.pgen.1000562.

Barreiro, Luis B., and Lluís Quintana-Murci. 2010. “From Evolutionary Genetics to Human Immunology: How Selection Shapes Host Defence Genes.” *Nature Reviews Genetics* 11 (1): 17–30. https://doi.org/10.1038/nrg2698.

Barreiro, Luis B., Ludovic Tailleux, Athma A. Pai, Brigitte Gicquel, John C. Marioni, and Yoav Gilad. 2012. “Deciphering the Genetic Architecture of Variation in the Immune Response to Mycobacterium Tuberculosis Infection.” *Proceedings of the National Academy of Sciences* 109 (4): 1204–9. https://doi.org/10.1073/pnas.1115761109.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using Lme4.” *Journal of Statistical Software* 67 (October): 1–48. https://doi.org/10.18637/jss.v067.i01.

Belbin, Gillian M., Sinead Cullina, Stephane Wenric, Emily R. Soper, Benjamin S. Glicksberg, Denis Torre, Arden Moscati, et al. 2021. “Toward a Fine-Scale Population Health Monitoring System.” *Cell* 184 (8): 2068-2083.e11. https://doi.org/10.1016/j.cell.2021.03.034.

Berg, Jeremy J., and Graham Coop. 2014. “A Population Genetic Signal of Polygenic Adaptation.” *PLOS Genetics* 10 (8): e1004412. https://doi.org/10.1371/journal.pgen.1004412.

Bergström, Anders, Shane A. McCarthy, Ruoyun Hui, Mohamed A. Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, et al. 2020. “Insights into Human Genetic Variation and Population History from 929 Diverse Genomes.” *Science* 367 (6484). https://doi.org/10.1126/science.aay5012.

Bernardes, Joana P., Neha Mishra, Florian Tran, Thomas Bahmer, Lena Best, Johanna I. Blase, Dora Bordoni, et al. 2020. "Longitudinal Multi-Omics Analyses Identify Responses of Megakaryocytes, Erythroid Cells, and Plasmablasts as Hallmarks of Severe COVID-19." *Immunity* 53 (6): 1296-1314.e9. https://doi.org/10.1016/j.immuni.2020.11.017.

Bigham, Abigail W., Kati J. Buckingham, Sofia Husain, Mary J. Emond, Kathryn M. Bofferding, Heidi Gildersleeve, Ann Rutherford, et al. 2011. "Host Genetic Risk Factors for West Nile Virus Infection and Disease Progression." *PloS One* 6 (9): e24745. https://doi.org/10.1371/journal.pone.0024745.

Bindea, Gabriela, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf-Herman Fridman, Franck Pagès, Zlatko Trajanoski, and Jérôme Galon. 2009. "ClueGO: A Cytoscape Plug-in to Decipher Functionally Grouped Gene Ontology and Pathway Annotation Networks." *Bioinformatics* 25 (8): 1091–93. https://doi.org/10.1093/bioinformatics/btp101.

Bock, Christoph, Paul Datlinger, Florence Chardon, Matthew A. Coelho, Matthew B. Dong, Keith A. Lawson, Tian Lu, et al. 2022. "High-Content CRISPR Screening." *Nature Reviews Methods Primers* 2 (1): 1–23. https://doi.org/10.1038/s43586-021-00093-4.

Booker, Tom R., Benjamin C. Jackson, and Peter D. Keightley. 2017. "Detecting Positive Selection in the Genome." *BMC Biology* 15 (1): 98. https://doi.org/10.1186/s12915-017-0434-y.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. "An Expanded View of Complex Traits: From Polygenic to Omnigenic." *Cell* 169 (7): 1177–86. https://doi.org/10.1016/j.cell.2017.05.038.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (5): 525–27. https://doi.org/10.1038/nbt.3519.

Brinkworth, Jessica F, and Luis B Barreiro. 2014. "The Contribution of Natural Selection to Present-Day Susceptibility to Chronic Inflammatory and Autoimmune Disease." *Current Opinion in Immunology* 31 (December): 66–78. https://doi.org/10.1016/j.coi.2014.09.008.

Brodin, Petter, and Mark M. Davis. 2017. "Human Immune System Variation." *Nature Reviews Immunology* 17 (1): 21–29. https://doi.org/10.1038/nri.2016.125.

Brodin, Petter, Vladimir Jojic, Tianxiang Gao, Sanchita Bhattacharya, Cesar J Lopez Angel, David Furman, Shai Shen-Orr, et al. 2015. "Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences." *Cell* 160 (0): 37–47. https://doi.org/10.1016/j.cell.2014.12.020.

Broxmeyer, Hal E., Scott Cooper, Giao Hangoc, Ji-Liang Gao, and Philip M. Murphy. 1999. "Dominant Myelopoietic Effector Functions Mediated by Chemokine Receptor CCR1."

Journal of Experimental Medicine 189 (12): 1987–92. https://doi.org/10.1084/jem.189.12.1987.

Burke, Bernard. 2003. "Macrophages as Novel Cellular Vehicles for Gene Therapy." *Expert Opinion on Biological Therapy* 3 (6): 919–24. https://doi.org/10.1517/14712598.3.6.919.

Çalışkan, Minal, Samuel W. Baker, Yoav Gilad, and Carole Ober. 2015. "Host Genetic Variation Influences Gene Expression Response to Rhinovirus Infection." *PLOS Genetics* 11 (4): e1005111. https://doi.org/10.1371/journal.pgen.1005111.

Chandrasekhar, Rameela, Chantel Sloan, Edward Mitchel, Danielle Ndi, Nisha Alden, Ann Thomas, Nancy M. Bennett, et al. 2017. "Social Determinants of Influenza Hospitalization in the United States." *Influenza and Other Respiratory Viruses* 11 (6): 479–88. https://doi.org/10.1111/irv.12483.

Chang, Christopher C., Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4: 7. https://doi.org/10.1186/s13742-015-0047-8.

Cheung, Vivian G., Richard S. Spielman, Kathryn G. Ewens, Teresa M. Weber, Michael Morley, and Joshua T. Burdick. 2005. "Mapping Determinants of Human Gene Expression by Regional and Genome-Wide Association." *Nature* 437 (7063): 1365–69. https://doi.org/10.1038/nature04244.

Ciancanelli, Michael J, Laurent Abel, Shen-Ying Zhang, and Jean-Laurent Casanova. 2016. "Host Genetics of Severe Influenza: From Mouse Mx1 to Human IRF7." *Current Opinion in Immunology*, Innate immunity, 38 (February): 109–20. https://doi.org/10.1016/j.coi.2015.12.002.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3." *Fly* 6 (2): 80–92. https://doi.org/10.4161/fly.19695.

Cooper, Max D., and Matthew N. Alder. 2006. "The Evolution of Adaptive Immune Systems." *Cell* 124 (4): 815–22. https://doi.org/10.1016/j.cell.2006.02.001.

Copeland, Shannon, H. Shaw Warren, Stephen F. Lowry, Steve E. Calvano, and Daniel Remick. 2005. "Acute Inflammatory Response to Endotoxin in Mice and Humans." *Clinical and Vaccine Immunology* 12 (1): 60–67. https://doi.org/10.1128/CDLI.12.1.60-67.2005.

Cuomo, Anna S E, Tobias Heinen, Danai Vagiaki, Danilo Horta, John C Marioni, and Oliver Stegle. 2022. "CellRegMap: A Statistical Framework for Mapping Context-Specific Regulatory Variants Using ScRNA-Seq." *Molecular Systems Biology* 18 (8): e10663. https://doi.org/10.15252/msb.202110663.

Das, Sayantan, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E. Locke, Alan Kwong, Scott I. Vrieze, et al. 2016. "Next-Generation Genotype Imputation Service and Methods." *Nature Genetics* 48 (10): 1284–87. https://doi.org/10.1038/ng.3656.

De Masi, Claudia, Paola Spitalieri, Michela Murdocca, Giuseppe Novelli, and Federica Sangiuolo. 2020. "Application of CRISPR/Cas9 to Human-Induced Pluripotent Stem Cells: From Gene Editing to Drug Discovery." *Human Genomics* 14 (1): 25. https://doi.org/10.1186/s40246-020-00276-2.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5): 491–98. https://doi.org/10.1038/ng.806.

Deschamps, Matthieu, Guillaume Laval, Maud Fagny, Yuval Itan, Laurent Abel, Jean-Laurent Casanova, Etienne Patin, and Lluis Quintana-Murci. 2016. "Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes." *The American Journal of Human Genetics* 98 (1): 5–21. https://doi.org/10.1016/j.ajhg.2015.11.014.

Doeing, Diana C., Jessica L. Borowicz, and Elahé T. Crockett. 2003. "Gender Dimorphism in Differential Peripheral Blood Leukocyte Counts in Mice Using Cardiac, Tail, Foot, and Saphenous Vein Puncture Methods." *BMC Clinical Pathology* 3 (1): 3. https://doi.org/10.1186/1472-6890-3-3.

Duffy, Darragh, Vincent Rouilly, Valentina Libri, Milena Hasan, Benoit Beitz, Mikael David, Alejandra Urrutia, et al. 2014. "Functional Analysis via Standardized Whole-Blood Stimulation Systems Defines the Boundaries of a Healthy Immune Response to Complex Stimuli." *Immunity* 40 (3): 436–50. https://doi.org/10.1016/j.immuni.2014.03.002.

Duncan, L., H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson, and B. Domingue. 2019. "Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations." *Nature Communications* 10 (1): 3328. https://doi.org/10.1038/s41467-019-11112-0.

Dupuis, Stéphanie, Emmanuelle Jouanguy, Sami Al-Hajjar, Claire Fieschi, Ibrahim Zaid Al-Mohsen, Suliman Al-Jumaah, Kun Yang, et al. 2003. "Impaired Response to Interferon-Alpha/Beta and Lethal Viral Disease in Human STAT1 Deficiency." *Nature Genetics* 33 (3): 388–91. https://doi.org/10.1038/ng1097.

Eames, Hayley L., Alastair L. Corbin, and Irina A. Udalova. 2016. "Interferon Regulatory Factor 5 in Human Autoimmunity and Murine Models of Autoimmune Disease." *Translational Research* 167 (1): 167–82. https://doi.org/10.1016/j.trsl.2015.06.018.

Eden, Eran, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. 2009. "GOrilla: A Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists." *BMC Bioinformatics* 10 (February): 48. https://doi.org/10.1186/1471-2105-10-48.

Edwards, Stacey L., Jonathan Beesley, Juliet D. French, and Alison M. Dunning. 2013. “Beyond GWASs: Illuminating the Dark Road from Association to Function.” *American Journal of Human Genetics* 93 (5): 779–97. https://doi.org/10.1016/j.ajhg.2013.10.012.

Eisenberg, Joseph N.S., Manish A. Desai, Karen Levy, Sarah J. Bates, Song Liang, Kyra Naumoff, and James C. Scott. 2007. “Environmental Determinants of Infectious Disease: A Framework for Tracking Causal Links and Guiding Public Health Research.” *Environmental Health Perspectives* 115 (8): 1216–23. https://doi.org/10.1289/ehp.9806.

El Awady, Mostafa K., Mohamed A. Anany, Gamal Esmat, Naglaa Zayed, Ashraf A. Tabll, Amr Helmy, Abdel Rahman El Zayady, et al. 2011. “Single Nucleotide Polymorphism at Exon 7 Splice Acceptor Site of OAS1 Gene Determines Response of Hepatitis C Virus Patients to Interferon Therapy.” *Journal of Gastroenterology and Hepatology* 26 (5): 843–50. https://doi.org/10.1111/j.1440-1746.2010.06605.x.

Enard, David, and Dmitri A. Petrov. 2018. “Evidence That RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans.” *Cell* 175 (2): 360-371.e13. https://doi.org/10.1016/j.cell.2018.08.034.

———. 2020. “Ancient RNA Virus Epidemics through the Lens of Recent Adaptation in Human Genomes.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 375 (1812): 20190575. https://doi.org/10.1098/rstb.2019.0575.

Fairfax, Benjamin P., Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, et al. 2014. “Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression.” *Science* 343 (6175): 1246949. https://doi.org/10.1126/science.1246949.

Fairfax, Benjamin P, and Julian C Knight. 2014. “Genetics of Gene Expression in Immunity to Infection.” *Current Opinion in Immunology*, Immunogenetics and transplantation * Special section: Effects of endogenous immune stimulants, 30 (October): 63–71. https://doi.org/10.1016/j.coi.2014.07.001.

Fatumo, Segun, Tinashe Chikowore, Ananyo Choudhury, Muhammad Ayub, Alicia R. Martin, and Karoline Kuchenbaecker. 2022. “A Roadmap to Increase Diversity in Genomic Studies.” *Nature Medicine* 28 (2): 243–50. https://doi.org/10.1038/s41591-021-01672-4.

Ferraro, Alessandra, Anna Morena D’Alise, Towfique Raj, Natasha Asinovski, Roxanne Phillips, Ayla Ergun, Joseph M. Replogle, et al. 2014. “Interindividual Variation in Human T Regulatory Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (12): E1111-1120. https://doi.org/10.1073/pnas.1401343111.

Fitzgerald-Bocarsly, Patricia, Jihong Dai, and Sukhwinder Singh. 2008. “Plasmacytoid Dendritic Cells and Type I IFN: 50 Years of Convergent History.” *Cytokine & Growth Factor Reviews* 19 (1): 3–19. https://doi.org/10.1016/j.cytogfr.2007.10.006.

Fodor, Ervin, Louise Devenish, Othmar G. Engelhardt, Peter Palese, George G. Brownlee, and Adolfo García-Sastre. 1999. “Rescue of Influenza A Virus from Recombinant DNA.” *Journal of Virology* 73 (11): 9679–82.

Fraser, Hunter B. 2013. “Gene Expression Drives Local Adaptation in Humans.” *Genome Research* 23 (7): 1089–96. https://doi.org/10.1101/gr.152710.112.

Freimer, Jacob W., Oren Shaked, Sahin Naqvi, Nasa Sinnott-Armstrong, Arwa Kathiria, Christian M. Garrido, Amy F. Chen, et al. 2022. “Systematic Discovery and Perturbation of Regulatory Genes in Human T Cells Reveals the Architecture of Immune Networks.” *Nature Genetics* 54 (8): 1133–44. https://doi.org/10.1038/s41588-022-01106-y.

Fumagalli, Matteo, and Manuela Sironi. 2014. “Human Genome Variability, Natural Selection and Infectious Diseases.” *Current Opinion in Immunology*, Immunogenetics and transplantation * Special section: Effects of endogenous immune stimulants, 30 (October): 9–16. https://doi.org/10.1016/j.coi.2014.05.001.

Fumagalli, Matteo, Manuela Sironi, Uberto Pozzoli, Anna Ferrer-Admettla, Linda Pattini, and Rasmus Nielsen. 2011. “Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution.” *PLOS Genetics* 7 (11): e1002355. https://doi.org/10.1371/journal.pgen.1002355.

Galbraith, Matthew D., Kohl T. Kinning, Kelly D. Sullivan, Paula Araya, Keith P. Smith, Ross E. Granrath, Jessica R. Shaw, et al. 2022. “Specialized Interferon Action in COVID-19.” *Proceedings of the National Academy of Sciences* 119 (11): e2116730119. https://doi.org/10.1073/pnas.2116730119.

Geraghty, R. J., A. Capes-Davis, J. M. Davis, J. Downward, R. I. Freshney, I. Knezevic, R. Lovell-Badge, et al. 2014. “Guidelines for the Use of Cell Lines in Biomedical Research.” *British Journal of Cancer* 111 (6): 1021–46. https://doi.org/10.1038/bjc.2014.166.

Gibson, Greg, and Bruce Weir. 2005. “The Quantitative Genetics of Transcription.” *Trends in Genetics: TIG* 21 (11): 616–23. https://doi.org/10.1016/j.tig.2005.08.010.

Hafemeister, Christoph, and Rahul Satija. 2019. “Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression.” *Genome Biology* 20 (1): 296. https://doi.org/10.1186/s13059-019-1874-1.

Hai, Rong, Mirco Schmolke, Zsuzsanna T. Varga, Balaji Manicassamy, Taia T. Wang, Jessica A. Belser, Melissa B. Pearce, Adolfo García-Sastre, Terrence M. Tumpey, and Peter Palese. 2010. “PB1-F2 Expression by the 2009 Pandemic H1N1 Influenza Virus Has Minimal Impact on Virulence in Animal Models.” *Journal of Virology* 84 (9): 4442–50. https://doi.org/10.1128/JVI.02717-09.

Hänzelmann, Sonja, Robert Castelo, and Justin Guinney. 2013. “GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data.” *BMC Bioinformatics* 14 (1): 7. https://doi.org/10.1186/1471-2105-14-7.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2021. “Integrated Analysis of Multimodal Single-Cell Data.” *Cell* 184 (13): 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

Heaton, Haynes, Arthur M. Talman, Andrew Knights, Maria Imaz, Daniel J. Gaffney, Richard Durbin, Martin Hemberg, and Mara K. N. Lawniczak. 2020. “Souporcell: Robust Clustering of Single-Cell RNA-Seq Data by Genotype without Reference Genotypes.” *Nature Methods* 17 (6): 615–20. https://doi.org/10.1038/s41592-020-0820-1.

Hindorff, Lucia A., Vence L. Bonham, Lawrence C. Brody, Margaret E. C. Ginoza, Carolyn M. Hutter, Teri A. Manolio, and Eric D. Green. 2018. “Prioritizing Diversity in Human Genomics Research.” *Nature Reviews Genetics* 19 (3): 175–85. https://doi.org/10.1038/nrg.2017.89.

Hoffmann, Erich, Gabriele Neumann, Yoshihiro Kawaoka, Gerd Hobom, and Robert G. Webster. 2000. “A DNA Transfection System for Generation of Influenza A Virus from Eight Plasmids.” *Proceedings of the National Academy of Sciences* 97 (11): 6108–13. https://doi.org/10.1073/pnas.100133697.

Hou, Wanqiu, James S. Gibbs, Xiuju Lu, Christopher B. Brooke, Devika Roy, Robert L. Modlin, Jack R. Bennink, and Jonathan W. Yewdell. 2012. “Viral Infection Triggers Rapid Differentiation of Human Blood Monocytes into Dendritic Cells.” *Blood* 119 (13): 3128–31. https://doi.org/10.1182/blood-2011-09-379479.

Huang, Jing-Ying, Wen-Chi Su, King-Song Jeng, Tien-Hsien Chang, and Michael M. C. Lai. 2012. “Attenuation of 40S Ribosomal Subunit Abundance Differentially Affects Host and HCV Translation and Suppresses HCV Replication.” *PLoS Pathogens* 8 (6). https://doi.org/10.1371/journal.ppat.1002766.

Iwama, Rafael E., and Yehu Moran. 2023. “Origins and Diversification of Animal Innate Immune Responses against Viral Infections.” *Nature Ecology & Evolution* 7 (2): 182–93. https://doi.org/10.1038/s41559-022-01951-4.

Iwasaki, Akiko, and Ruslan Medzhitov. 2015. “Control of Adaptive Immunity by the Innate Immune System.” *Nature Immunology* 16 (4): 343–53. https://doi.org/10.1038/ni.3123.

Jensen, Søren, and Allan Randrup Thomsen. 2012. “Sensing of RNA Viruses: A Review of Innate Immune Receptors Involved in Recognizing RNA Virus Invasion.” *Journal of Virology* 86 (6): 2900–2910. https://doi.org/10.1128/JVI.05738-11.

Karlsson, Elinor K., Dominic P. Kwiatkowski, and Pardis C. Sabeti. 2014. “Natural Selection and Infectious Disease in Human Populations.” *Nature Reviews Genetics* 15 (6): 379–93. https://doi.org/10.1038/nrg3734.

Kelley-Hedgepeth, Alyson, Donald M. Lloyd-Jones, Alicia Colvin, Karen A. Matthews, Janet Johnston, MaryFran R. Sowers, Barbara Sternfeld, Richard C. Pasternak, and Claudia U. Chae. 2008. "Ethnic Differences in C-Reactive Protein Concentrations." *Clinical Chemistry* 54 (6): 1027–37. https://doi.org/10.1373/clinchem.2007.098996.

Kelso, Janet, and Kay Prüfer. 2014. "Ancient Humans and the Origin of Modern Humans." *Current Opinion in Genetics & Development*, Genetics of human evolution, 29 (December): 133–38. https://doi.org/10.1016/j.gde.2014.09.004.

Kenney, Adam D., James A. Dowdle, Leonia Bozzacco, Temet M. McMichael, Corine St Gelais, Amanda R. Panfil, Yan Sun, et al. 2017. "Human Genetic Determinants of Viral Diseases." *Annual Review of Genetics* 51: 241–63. https://doi.org/10.1146/annurev-genet-120116-023425.

Kilpinen, Helena, Angela Goncalves, Andreas Leha, Vackar Afzal, Kaur Alasoo, Sofie Ashford, Sendu Bala, et al. 2017. "Common Genetic Variation Drives Molecular Heterogeneity in Human IPSCs." *Nature* 546 (7658): 370–75. https://doi.org/10.1038/nature22403.

Kim-Hellmuth, Sarah, Matthias Bechheim, Benno Pütz, Pejman Mohammadi, Yohann Nédélec, Nicholas Giangreco, Jessica Becker, et al. 2017. "Genetic Regulatory Effects Modified by Immune Activation Contribute to Autoimmune Disease Associations." *Nature Communications* 8 (1): 266. https://doi.org/10.1038/s41467-017-00366-1.

Klein, Sabra L., and Katie L. Flanagan. 2016. "Sex Differences in Immune Responses." *Nature Reviews. Immunology* 16 (10): 626–38. https://doi.org/10.1038/nri.2016.90.

Knoll, Rainer, Joachim L. Schultze, and Jonas Schulte-Schrepping. 2021. "Monocytes and Macrophages in COVID-19." *Frontiers in Immunology* 12. https://www.frontiersin.org/articles/10.3389/fimmu.2021.720109.

Ko, Dennis C., Eric R. Gamazon, Kajal P. Shukla, Richard A. Pfuetzner, Dale Whittington, Tarah D. Holden, Mitchell J. Brittnacher, et al. 2012. "Functional Genetic Screen of Human Diversity Reveals That a Methionine Salvage Enzyme Regulates Inflammatory Cell Death." *Proceedings of the National Academy of Sciences* 109 (35): E2343–52. https://doi.org/10.1073/pnas.1206701109.

Ko, Jean Y, Melissa L Danielson, Machell Town, Gordana Derado, Kurt J Greenlund, Pam Daily Kirley, Nisha B Alden, et al. 2021. "Risk Factors for Coronavirus Disease 2019 (COVID-19)–Associated Hospitalization: COVID-19–Associated Hospitalization Surveillance Network and Behavioral Risk Factor Surveillance System." *Clinical Infectious Diseases* 72 (11): e695–703. https://doi.org/10.1093/cid/ciaa1419.

Ko, Wen-Ya, Kristin A. Kaercher, Emanuela Giombini, Paolo Marcatili, Alain Froment, Muntaser Ibrahim, Godfrey Lema, et al. 2011. "Effects of Natural Selection and Gene Conversion on the Evolution of Human Glycophorins Coding for MNS Blood Polymorphisms in Malaria-Endemic African Populations." *The American Journal of Human Genetics* 88 (6): 741–54. https://doi.org/10.1016/j.ajhg.2011.05.005.

Ko, Wen-Ya, Prianka Rajan, Felicia Gomez, Laura Scheinfeldt, Ping An, Cheryl A. Winkler, Alain Froment, et al. 2013. “Identifying Darwinian Selection Acting on Different Human APOL1 Variants among Diverse African Populations.” *The American Journal of Human Genetics* 93 (1): 54–66. https://doi.org/10.1016/j.ajhg.2013.05.014.

Korotkevich, Gennady, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N. Artyomov, and Alexey Sergushichev. 2021. “Fast Gene Set Enrichment Analysis.” bioRxiv. https://doi.org/10.1101/060012.

Kwiatkowski, Dominic P. 2005. “How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria.” *The American Journal of Human Genetics* 77 (2): 171–92. https://doi.org/10.1086/432519.

Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. ‘t Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, et al. 2013. “Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans.” *Nature* 501 (7468): 506–11. https://doi.org/10.1038/nature12531.

Lee, Jeong Seok, and Eui-Cheol Shin. 2020. “The Type I Interferon Response in COVID-19: Implications for Treatment.” *Nature Reviews Immunology* 20 (10): 585–86. https://doi.org/10.1038/s41577-020-00429-3.

Lee, Mark N., Chun Ye, Alexandra-Chloé Villani, Towfique Raj, Weibo Li, Thomas M. Eisenhaure, Selina H. Imboywa, et al. 2014. “Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells.” *Science* 343 (6175): 1246980. https://doi.org/10.1126/science.1246980.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, Shuo. 2019. “Regulation of Ribosomal Proteins on Viral Infection.” *Cells* 8 (5). https://doi.org/10.3390/cells8050508.

Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. “The Molecular Signatures Database Hallmark Gene Set Collection.” *Cell Systems* 1 (6): 417–25. https://doi.org/10.1016/j.cels.2015.12.004.

Liu, Can, Andrew J. Martins, William W. Lau, Nicholas Rachmaninoff, Jinguo Chen, Luisa Imberti, Darius Mostaghimi, et al. 2021. “Time-Resolved Systems Immunology Reveals a Late Juncture Linked to Fatal COVID-19.” *Cell* 184 (7): 1836-1857.e22. https://doi.org/10.1016/j.cell.2021.02.018.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” *Cell* 161 (5): 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Manry, Jérémy, Yohann Nédélec, Vinicius M. Fava, Aurélie Cobat, Marianna Orlova, Nguyen Van Thuc, Vu Hong Thai, Guillaume Laval, Luis B. Barreiro, and Erwin Schurr. 2017. "Deciphering the Genetic Control of Gene Expression Following Mycobacterium Leprae Antigen Stimulation." *PLOS Genetics* 13 (8): e1006952. https://doi.org/10.1371/journal.pgen.1006952.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. "Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities." *Nature Genetics* 51 (4): 584–91. https://doi.org/10.1038/s41588-019-0379-x.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.Journal* 17 (1): 10–12. https://doi.org/10.14806/ej.17.1.200.

Masood, Kiran Iqbal, Maliha Yameen, Javeria Ashraf, Saba Shahid, Syed Faisal Mahmood, Asghar Nasir, Nosheen Nasir, et al. 2021. "Upregulated Type I Interferon Responses in Asymptomatic COVID-19 Infection Are Associated with Improved Clinical Outcome." *Scientific Reports* 11 (1): 22958. https://doi.org/10.1038/s41598-021-02489-4.

Mendez, Fernando L., Joseph C. Watkins, and Michael F. Hammer. 2013. "Neandertal Origin of Genetic Variation at the Cluster of OAS Immunity Genes." *Molecular Biology and Evolution* 30 (4): 798–801. https://doi.org/10.1093/molbev/mst004.

Mestas, Javier, and Christopher C. W. Hughes. 2004. "Of Mice and Not Men: Differences between Mouse and Human Immunology." *The Journal of Immunology* 172 (5): 2731–38. https://doi.org/10.4049/jimmunol.172.5.2731.

Montecino-Rodriguez, Encarnacion, Beata Berent-Maoz, and Kenneth Dorshkind. 2013. "Causes, Consequences, and Reversal of Immune System Aging." *The Journal of Clinical Investigation* 123 (3): 958–65. https://doi.org/10.1172/JCI64096.

Morley, Michael, Cliona M. Molony, Teresa M. Weber, James L. Devlin, Kathryn G. Ewens, Richard S. Spielman, and Vivian G. Cheung. 2004. "Genetic Analysis of Genome-Wide Variation in Human Gene Expression." *Nature* 430 (7001): 743–47. https://doi.org/10.1038/nature02797.

Mozzi, Alessandra, Chiara Pontremoli, and Manuela Sironi. 2018. "Genetic Susceptibility to Infectious Diseases: Current Status and Future Perspectives from Genome-Wide Approaches." *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 66 (December): 286–307. https://doi.org/10.1016/j.meegid.2017.09.028.

Nami, Fatemeharefeh, Mohsen Basiri, Leila Satarian, Cameron Curtiss, Hossein Baharvand, and Catherine Verfaillie. 2018. "Strategies for In Vivo Genome Editing in Nondividing Cells." *Trends in Biotechnology* 36 (8): 770–86. https://doi.org/10.1016/j.tibtech.2018.03.004.

Nathan, Aparna, Samira Asgari, Kazuyoshi Ishigaki, Cristian Valencia, Tiffany Amariuta, Yang Luo, Jessica I. Beynor, et al. 2022. “Single-Cell EQTL Models Reveal Dynamic T Cell State Dependence of Disease Loci.” *Nature*, May, 1–9. https://doi.org/10.1038/s41586-022-04713-1.

Nédélec, Yohann, Joaquín Sanz, Golshid Baharian, Zachary A. Szpiech, Alain Pacis, Anne Dumaine, Jean-Christophe Grenier, et al. 2016. “Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens.” *Cell* 167 (3): 657-669.e21. https://doi.org/10.1016/j.cell.2016.09.025.

Ness, Roberta B., Catherine L. Haggerty, Gail Harger, and Robert Ferrell. 2004. “Differential Distribution of Allelic Variants in Cytokine Genes among African Americans and White Americans.” *American Journal of Epidemiology* 160 (11): 1033–38. https://doi.org/10.1093/aje/kwh325.

Netea, Mihai G., Jorge Domínguez-Andrés, Luis B. Barreiro, Triantafyllos Chavakis, Maziar Divangahi, Elaine Fuchs, Leo A. B. Joosten, et al. 2020. “Defining Trained Immunity and Its Role in Health and Disease.” *Nature Reviews Immunology* 20 (6): 375–88. https://doi.org/10.1038/s41577-020-0285-6.

Nielsen, Rasmus, Joshua M. Akey, Mattias Jakobsson, Jonathan K. Pritchard, Sarah Tishkoff, and Eske Willerslev. 2017. “Tracing the Peopling of the World through Genomics.” *Nature* 541 (7637): 302–10. https://doi.org/10.1038/nature21347.

Nielsen, Rasmus, Carlos Bustamante, Andrew G. Clark, Stephen Glanowski, Timothy B. Sackton, Melissa J. Hubisz, Adi Fledel-Alon, et al. 2005. “A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees.” *PLoS Biology* 3 (6): e170. https://doi.org/10.1371/journal.pbio.0030170.

Niemi, Mari E. K., Juha Karjalainen, Rachel G. Liao, Benjamin M. Neale, Mark Daly, Andrea Ganna, Gita A. Pathak, et al. 2021. “Mapping the Human Genetic Architecture of COVID-19.” *Nature* 600 (7889): 472–77. https://doi.org/10.1038/s41586-021-03767-x.

Okin, Daniel, and Ruslan Medzhitov. 2012. “Evolution of Inflammatory Diseases.” *Current Biology* 22 (17): R733–40. https://doi.org/10.1016/j.cub.2012.07.029.

Okita, Keisuke, Yasuko Matsumura, Yoshiko Sato, Aki Okada, Asuka Morizane, Satoshi Okamoto, Hyenjong Hong, et al. 2011. “A More Efficient Method to Generate Integration-Free Human IPS Cells.” *Nature Methods* 8 (5): 409–12. https://doi.org/10.1038/nmeth.1591.

Oliva, Meritxell, Manuel Muñoz-Aguirre, Sarah Kim-Hellmuth, Valentin Wucher, Ariel D. H. Gewirtz, Daniel J. Cotter, Princy Parsana, et al. 2020. “The Impact of Sex on Gene Expression across Human Tissues.” *Science* 369 (6509). https://doi.org/10.1126/science.aba3066.

Pallares, Luisa F., Serge Picard, and Julien F. Ayroles. 2020. “TM3’seq: A Tagmentation-Mediated 3’ Sequencing Approach for Improving Scalability of RNAseq Experiments.” *G3: Genes, Genomes, Genetics* 10 (1): 143–50. https://doi.org/10.1534/g3.119.400821.

Papatriantafyllou, Maria. 2013. “The Interferon Paradox.” *Nature Reviews Immunology* 13 (6): 392–93. https://doi.org/10.1038/nri3461.

Pennington, Renee, Chandler Gatenbee, Brett Kennedy, Henry Harpending, and Gregory Cochran. 2009. “Group Differences in Proneness to Inflammation.” *Infection, Genetics and Evolution*, Includes papers from the Special Issue “Parasitology in Mexico,” 9 (6): 1371–80. https://doi.org/10.1016/j.meegid.2009.09.017.

Peterson, Roseann E., Karoline Kuchenbaecker, Raymond K. Walters, Chia-Yen Chen, Alice B. Popejoy, Sathish Periyasamy, Max Lam, et al. 2019. “Genome-Wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations.” *Cell* 179 (3): 589–603. https://doi.org/10.1016/j.cell.2019.08.051.

Piasecka, Barbara, Darragh Duffy, Alejandra Urrutia, Hélène Quach, Etienne Patin, Céline Posseme, Jacob Bergstedt, et al. 2018. “Distinctive Roles of Age, Sex, and Genetics in Shaping Transcriptional Variation of Human Immune Responses to Microbial Challenges.” *Proceedings of the National Academy of Sciences* 115 (3): E488–97. https://doi.org/10.1073/pnas.1714765115.

Pickrell, Joseph K., John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. 2010. “Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing.” *Nature* 464 (7289): 768–72. https://doi.org/10.1038/nature08872.

Poli, Aurélie, Tatiana Michel, Maud Thérésine, Emmanuel Andrès, François Hentges, and Jacques Zimmer. 2009. “CD56bright Natural Killer (NK) Cells: An Important NK Cell Subset.” *Immunology* 126 (4): 458–65. https://doi.org/10.1111/j.1365-2567.2008.03027.x.

Pritchard, Jonathan K., Joseph K. Pickrell, and Graham Coop. 2010. “The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation.” *Current Biology: CB* 20 (4): R208-215. https://doi.org/10.1016/j.cub.2009.11.055.

Privé, Florian, Hugues Aschard, Andrey Ziyatdinov, and Michael G B Blum. 2018. “Efficient Analysis of Large-Scale Genome-Wide Data with Two R Packages: Bigstatsr and Bigsnpr.” *Bioinformatics* 34 (16): 2781–87. https://doi.org/10.1093/bioinformatics/bty185.

Quach, Hélène, and Lluis Quintana-Murci. 2017. “Living in an Adaptive World: Genomic Dissection of the Genus Homo and Its Immune Response.” *Journal of Experimental Medicine* 214 (4): 877–94. https://doi.org/10.1084/jem.20161942.

Quach, Hélène, Maxime Rotival, Julien Pothlichet, Yong-Hwee Eddie Loh, Michael Dannemann, Nora Zidane, Guillaume Laval, et al. 2016. “Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations.” *Cell* 167 (3): 643-656.e17. https://doi.org/10.1016/j.cell.2016.09.024.

Quintana-Murci, Lluís, and Andrew G. Clark. 2013. “Population Genetic Tools for Dissecting Innate Immunity in Humans.” *Nature Reviews Immunology* 13 (4): 280–93. https://doi.org/10.1038/nri3421.

Raifman, Matthew A., and Julia R. Raifman. 2020. “Disparities in the Population at Risk of Severe Illness From COVID-19 by Race/Ethnicity and Income.” *American Journal of Preventive Medicine* 59 (1): 137–39. https://doi.org/10.1016/j.amepre.2020.04.003.

Randolph, Haley E., Jessica K. Fiege, Beth K. Thielen, Clayton K. Mickelson, Mari Shiratori, João Barroso-Batista, Ryan A. Langlois, and Luis B. Barreiro. 2021. “Genetic Ancestry Effects on the Response to Viral Infection Are Pervasive but Cell Type Specific.” *Science* 374 (6571): 1127–33. https://doi.org/10.1126/science.abg0928.

Ren, Xianwen, Wen Wen, Xiaoying Fan, Wenhong Hou, Bin Su, Pengfei Cai, Jiesheng Li, et al. 2021. “COVID-19 Immune Features Revealed by a Large-Scale Single-Cell Transcriptome Atlas.” *Cell* 184 (7): 1895-1913.e19. https://doi.org/10.1016/j.cell.2021.01.053.

Richardus, Jan H., and Anton E. Kunst. 2001. “Black–White Differences in Infectious Disease Mortality in the United States.” *American Journal of Public Health* 91 (8): 1251–53.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47–e47. https://doi.org/10.1093/nar/gkv007.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Rodero, Mathieu P., and Yanick J. Crow. 2016. “Type I Interferon–Mediated Monogenic Autoinflammation: The Type I Interferonopathies, a Conceptual Overview.” *Journal of Experimental Medicine* 213 (12): 2527–38. https://doi.org/10.1084/jem.20161596.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. “Positive Natural Selection in the Human Lineage.” *Science (New York, N.Y.)* 312 (5780): 1614–20. https://doi.org/10.1126/science.1124309.

Sabeti, Pardis C., Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, et al. 2007. “Genome-Wide Detection and Characterization of

Positive Selection in Human Populations." *Nature* 449 (7164): 913–18. https://doi.org/10.1038/nature06250.

Sams, Aaron J., Anne Dumaine, Yohann Nédélec, Vania Yotova, Carolina Alfieri, Jerome E. Tanner, Philipp W. Messer, and Luis B. Barreiro. 2016. "Adaptively Introgressed Neandertal Haplotype at the OAS Locus Functionally Impacts Innate Immune Responses in Humans." *Genome Biology* 17 (November): 246. https://doi.org/10.1186/s13059-016-1098-6.

Sancho-Shimizu, Vanessa, Rebeca Pérez de Diego, Lazaro Lorenzo, Rabih Halwani, Abdullah Alangari, Elisabeth Israelsson, Sylvie Fabrega, et al. 2011. "Herpes Simplex Encephalitis in Children with Autosomal Recessive and Dominant TRIF Deficiency." *The Journal of Clinical Investigation* 121 (12): 4889–4902. https://doi.org/10.1172/JCI59259.

Schoenborn, Jamie R., and Christopher B. Wilson. 2007. "Regulation of Interferon-γ During Innate and Adaptive Immune Responses." In *Advances in Immunology*, 96:41–101. Academic Press. https://doi.org/10.1016/S0065-2776(07)96002-2.

Schulte-Schrepping, Jonas, Nico Reusch, Daniela Paclik, Kevin Baßler, Stephan Schlickeiser, Bowen Zhang, Benjamin Krämer, et al. 2020. "Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment." *Cell* 182 (6): 1419-1440.e23. https://doi.org/10.1016/j.cell.2020.08.001.

Sears, Margaret E., and Stephen J. Genuis. 2012. "Environmental Determinants of Chronic Disease and Medical Approaches: Recognition, Avoidance, Supportive Therapy, and Detoxification." *Journal of Environmental and Public Health* 2012: 356798. https://doi.org/10.1155/2012/356798.

Ségurel, Laure, and Lluis Quintana-Murci. 2014. "Preserving Immune Diversity through Ancient Inheritance and Admixture." *Current Opinion in Immunology*, Immunogenetics and transplantation * Special section: Effects of endogenous immune stimulants, 30 (October): 79–84. https://doi.org/10.1016/j.coi.2014.08.002.

Shabalin, Andrey A. 2012. "Matrix EQTL: Ultra Fast EQTL Analysis via Large Matrix Operations." *Bioinformatics* 28 (10): 1353–58. https://doi.org/10.1093/bioinformatics/bts163.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504. https://doi.org/10.1101/gr.1239303.

Shrock, Ellen L., Christine L. Shrock, and Stephen J. Elledge. 2022. "VirScan: High-Throughput Profiling of Antiviral Antibody Epitopes." *Bio-Protocol* 12 (13): e4464. https://doi.org/10.21769/BioProtoc.4464.

Siddle, Katherine J., and Lluis Quintana-Murci. 2014. “The Red Queen’s Long Race: Human Adaptation to Pathogen Pressure.” *Current Opinion in Genetics & Development* 29 (December): 31–38. https://doi.org/10.1016/j.gde.2014.07.004.

Smale, Stephen T. 2010. “Selective Transcription in Response to an Inflammatory Stimulus.” *Cell* 140 (6): 833–44. https://doi.org/10.1016/j.cell.2010.01.037.

Snyder-Mackler, Noah, Joaquín Sanz, Jordan N. Kohn, Jessica F. Brinkworth, Shauna Morrow, Amanda O. Shaver, Jean-Christophe Grenier, et al. 2016. “Social Status Alters Immune Regulation and Response to Infection in Macaques.” *Science (New York, N.Y.)* 354 (6315): 1041–45. https://doi.org/10.1126/science.aah3580.

Soneson, Charlotte, Michael I. Love, and Mark D. Robinson. 2016. “Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences.” *F1000Research* 4 (February): 1521. https://doi.org/10.12688/f1000research.7563.2.

Stephens Patrick R., Altizer Sonia, Smith Katherine F., Alonso Aguirre A., Brown James H., Budischak Sarah A., Byers James E., et al. 2016. “The Macroecology of Infectious Diseases: A New Perspective on Global-scale Drivers of Pathogen Distributions and Impacts.” *Ecology Letters* 19 (9): 1159–71. https://doi.org/10.1111/ele.12644.

Steuerman, Yael, Merav Cohen, Naama Peshes-Yaloz, Liran Valadarsky, Ofir Cohn, Eyal David, Amit Frishberg, et al. 2018. “Dissection of Influenza Infection In Vivo by Single-Cell RNA Sequencing.” *Cell Systems* 6 (6): 679-691.e4. https://doi.org/10.1016/j.cels.2018.05.008.

Storey, John D., Andrew J. Bass, Alan Dabney, David Robinson, and Gregory Warnes. 2023. “Qvalue: Q-Value Estimation for False Discovery Rate Control.” Bioconductor version: Release (3.16). https://doi.org/10.18129/B9.bioc.qvalue.

Storey, John D., and Robert Tibshirani. 2003. “Statistical Significance for Genomewide Studies.” *Proceedings of the National Academy of Sciences* 100 (16): 9440–45. https://doi.org/10.1073/pnas.1530509100.

Strober, B. J., R. Elorbany, K. Rhodes, N. Krishnan, K. Tayeb, A. Battle, and Y. Gilad. 2019. “Dynamic Genetic Regulation of Gene Expression during Cellular Differentiation.” *Science* 364 (6447): 1287–90. https://doi.org/10.1126/science.aaw0040.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. “Comprehensive Integration of Single-Cell Data.” *Cell* 177 (7): 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

Su, Yapeng, Daniel Chen, Dan Yuan, Christopher Lausted, Jongchan Choi, Chengzhen L. Dai, Valentin Voillet, et al. 2020. “Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19.” *Cell* 183 (6): 1479-1495.e20. https://doi.org/10.1016/j.cell.2020.10.037.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.

Taliun, Daniel, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, et al. 2021. “Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program.” *Nature* 590 (7845): 290–99. https://doi.org/10.1038/s41586-021-03205-y.

The Severe Covid-19 GWAS Group. 2020. “Genomewide Association Study of Severe Covid-19 with Respiratory Failure.” *New England Journal of Medicine* 383 (16): 1522–34. https://doi.org/10.1056/NEJMoa2020283.

Thomsen, Michelle M., Sofie E. Jørgensen, Merete Storgaard, Lasse S. Kristensen, Jakob Gjedsted, Mette Christiansen, Hans Henrik Gad, Rune Hartmann, and Trine H. Mogensen. 2019. “Identification of an IRF3 Variant and Defective Antiviral Interferon Responses in a Patient with Severe Influenza.” *European Journal of Immunology* 49 (11): 2111–14. https://doi.org/10.1002/eji.201848083.

Traherne, J. A. 2008. “Human MHC Architecture and Evolution: Implications for Disease Association Studies.” *International Journal of Immunogenetics* 35 (3): 179–92. https://doi.org/10.1111/j.1744-313X.2008.00765.x.

Trapnell, Cole. 2015. “Defining Cell Types and States with Single-Cell Genomics.” *Genome Research* 25 (10): 1491–98. https://doi.org/10.1101/gr.190595.115.

Tremblay, Karine, Simon Rousseau, Ma’n H. Zawati, Daniel Auld, Michaël Chassé, Daniel Coderre, Emilia Liana Falcone, et al. 2021. “The Biobanque Québécoise de La COVID-19 (BQC19)—A Cohort to Prospectively Study the Clinical and Biological Determinants of COVID-19 Clinical Trajectories.” *PLOS ONE* 16 (5): e0245031. https://doi.org/10.1371/journal.pone.0245031.

Urbut, Sarah M., Gao Wang, Peter Carbonetto, and Matthew Stephens. 2019. “Flexible Statistical Methods for Estimating and Testing Effects in Genomic Studies with Multiple Conditions.” *Nature Genetics* 51 (1): 187–95. https://doi.org/10.1038/s41588-018-0268-8.

Virolainen, Samuel J., Andrew VonHandorf, Kenyatta C. M. F. Viel, Matthew T. Weirauch, and Leah C. Kottyan. 2022. “Gene–Environment Interactions and Their Impact on Human Health.” *Genes & Immunity*, December, 1–11. https://doi.org/10.1038/s41435-022-00192-6.

Voight, Benjamin F., Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K. Pritchard. 2006. “A Map of Recent Positive Selection in the Human Genome.” *PLOS Biology* 4 (3): e72. https://doi.org/10.1371/journal.pbio.0040072.

Võsa, Urmo, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, et al. 2021. "Large-Scale Cis- and Trans-EQTL Analyses Identify Thousands of Genetic Loci and Polygenic Scores That Regulate Blood Gene Expression." *Nature Genetics* 53 (9): 1300–1310. https://doi.org/10.1038/s41588-021-00913-z.

Wang, Liuyang, Kelly J. Pittman, Jeffrey R. Barker, Raul E. Salinas, Ian B. Stanaway, Graham D. Williams, Robert J. Carroll, et al. 2018. "An Atlas of Genetic Variation Linking Pathogen-Induced Cellular Traits to Human Disease." *Cell Host & Microbe* 24 (2): 308-323.e6. https://doi.org/10.1016/j.chom.2018.07.007.

Wei, Jiajie, and Jonathan W. Yewdell. 2019. "Immunoribosomes: Where's There's Fire, There's Fire." *Molecular Immunology*, EMBO Workshop on Antigen Processing and Presentation, Salamanca, Spain, 2017, 113 (September): 38–42. https://doi.org/10.1016/j.molimm.2017.12.026.

Westra, Harm-Jan, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, et al. 2013. "Systematic Identification of Trans EQTLs as Putative Drivers of Known Disease Associations." *Nature Genetics* 45 (10): 1238–43. https://doi.org/10.1038/ng.2756.

Wijst, MGP van der, DH de Vries, HE Groot, G Trynka, CC Hon, MJ Bonder, O Stegle, et al. 2020. "The Single-Cell EQTLGen Consortium." Edited by Helena Pérez Valle, Peter Rodgers, Stephen B Montgomery, and Maud Fagny. *ELife* 9 (March): e52155. https://doi.org/10.7554/eLife.52155.

Wolfe, Nathan D., Claire Panosian Dunavan, and Jared Diamond. 2007. "Origins of Major Human Infectious Diseases." *Nature* 447 (7142): 279–83. https://doi.org/10.1038/nature05775.

Xia, Chang, Zachary Braunstein, Amelia C. Toomey, Jixin Zhong, and Xiaoquan Rao. 2018. "S100 Proteins As an Important Regulator of Macrophage Inflammation." *Frontiers in Immunology* 8. https://doi.org/10.3389/fimmu.2017.01908.

Ye, Chun Jimmie, Ting Feng, Ho-Keun Kwon, Towfique Raj, Michael T. Wilson, Natasha Asinovski, Cristin McCabe, et al. 2014. "Intersection of Population Variation and Autoimmunity Genetics in Human T Cell Activation." *Science (New York, N.Y.)* 345 (6202): 1254665. https://doi.org/10.1126/science.1254665.

Yoo, Ji-Seung, Michihito Sasaki, Steven X. Cho, Yusuke Kasuga, Baohui Zhu, Ryota Ouda, Yasuko Orba, Paul de Figueiredo, Hirofumi Sawa, and Koichi S. Kobayashi. 2021. "SARS-CoV-2 Inhibits Induction of the MHC Class I Pathway by Targeting the STAT1-IRF1-NLRC5 Axis." *Nature Communications* 12 (1): 6602. https://doi.org/10.1038/s41467-021-26910-8.

Yu, Junying, Maxim A. Vodyanik, Kim Smuga-Otto, Jessica Antosiewicz-Bourget, Jennifer L. Frane, Shulan Tian, Jeff Nie, et al. 2007. "Induced Pluripotent Stem Cell Lines Derived

from Human Somatic Cells." *Science* 318 (5858): 1917–20. https://doi.org/10.1126/science.1151526.

Zhang, Qian. 2020. "Human Genetics of Life-Threatening Influenza Pneumonitis." *Human Genetics* 139 (6): 941–48. https://doi.org/10.1007/s00439-019-02108-3.

Zhang, Qian, Paul Bastard, Zhiyong Liu, Jérémie Le Pen, Marcela Moncada-Velez, Jie Chen, Masato Ogishi, et al. 2020. "Inborn Errors of Type I IFN Immunity in Patients with Life-Threatening COVID-19." *Science* 370 (6515). https://doi.org/10.1126/science.abd4570.

Zhao, Hao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. 2014. "CrossMap: A Versatile Tool for Coordinate Conversion between Genome Assemblies." *Bioinformatics* 30 (7): 1006–7. https://doi.org/10.1093/bioinformatics/btt730.

Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. "Massively Parallel Digital Transcriptional Profiling of Single Cells." *Nature Communications* 8 (1): 14049. https://doi.org/10.1038/ncomms14049.

Zheng, Xiuwen, David Levine, Jess Shen, Stephanie M. Gogarten, Cathy Laurie, and Bruce S. Weir. 2012. "A High-Performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data." *Bioinformatics (Oxford, England)* 28 (24): 3326–28. https://doi.org/10.1093/bioinformatics/bts606.

Zhou, Fei, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, et al. 2020. "Clinical Course and Risk Factors for Mortality of Adult Inpatients with COVID-19 in Wuhan, China: A Retrospective Cohort Study." *Lancet (London, England)* 395 (10229): 1054–62. https://doi.org/10.1016/S0140-6736(20)30566-3.

Zhou, Sirui, Guillaume Butler-Laporte, Tomoko Nakanishi, David R. Morrison, Jonathan Afilalo, Marc Afilalo, Laetitia Laurent, et al. 2021. "A Neanderthal OAS1 Isoform Protects Individuals of European Ancestry against COVID-19 Susceptibility and Severity." *Nature Medicine*, February, 1–9. https://doi.org/10.1038/s41591-021-01281-1.

Appendix A: Supplementary Figures and Tables

Supplementary Figures for Chapter III

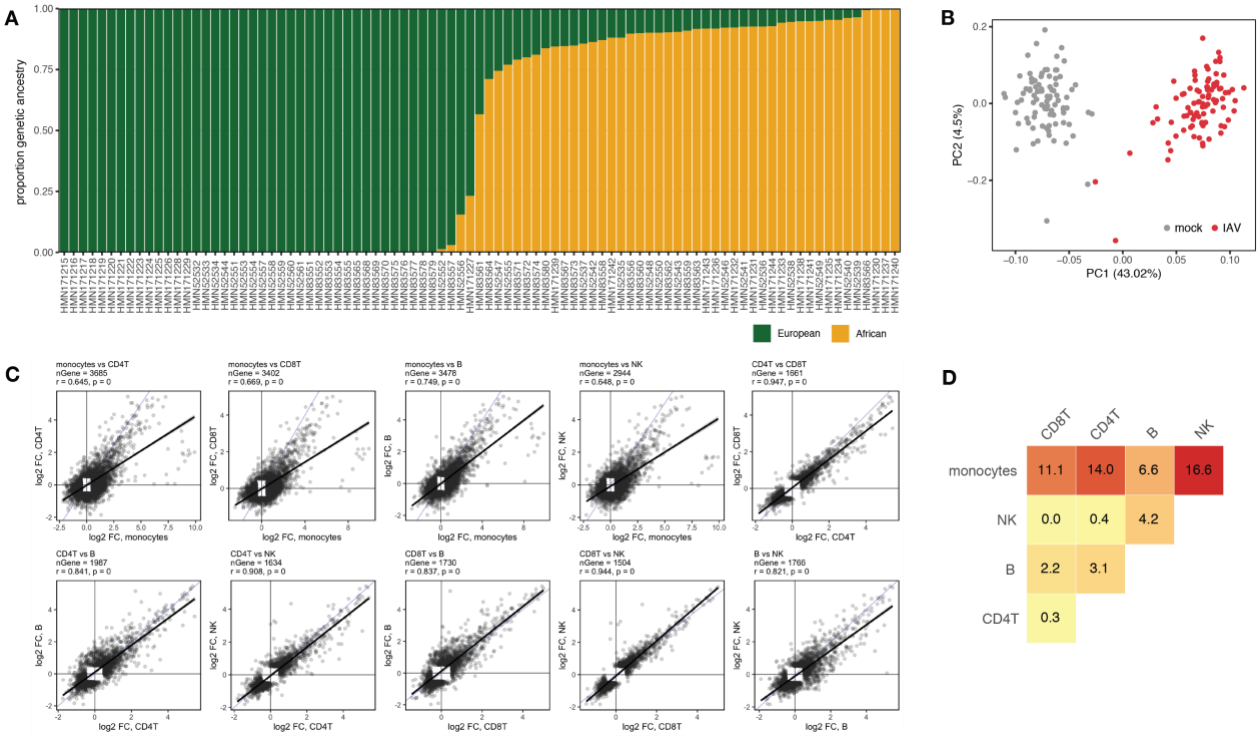


Fig. S3-1. Overview of samples and global infection effects. (A) Quantitative genetic ancestry proportions partitioned into European (green) and African (yellow) components for each individual. (B) PCA decomposition of the pseudobulk PBMC expression data in mock-exposed (grey) and IAV-infected (red) samples. PC1 (percent variance explained = 43.02%) separates samples by infection status. (C) Pairwise effect size correlations across cell types among genes that are DE ($|\log_2 FC| > 0.5$, FDR < 0.05) upon IAV infection in either of the cell types being compared. Black line shows the best-fit line from a linear model, blue line shows the x = y line. P-values were obtained from linear regression models (D) Pairwise comparisons of the percentage of DE ($|\log_2 FC| > 0.5$, FDR < 0.05) genes in both cell types being compared that show discordant effect sizes.

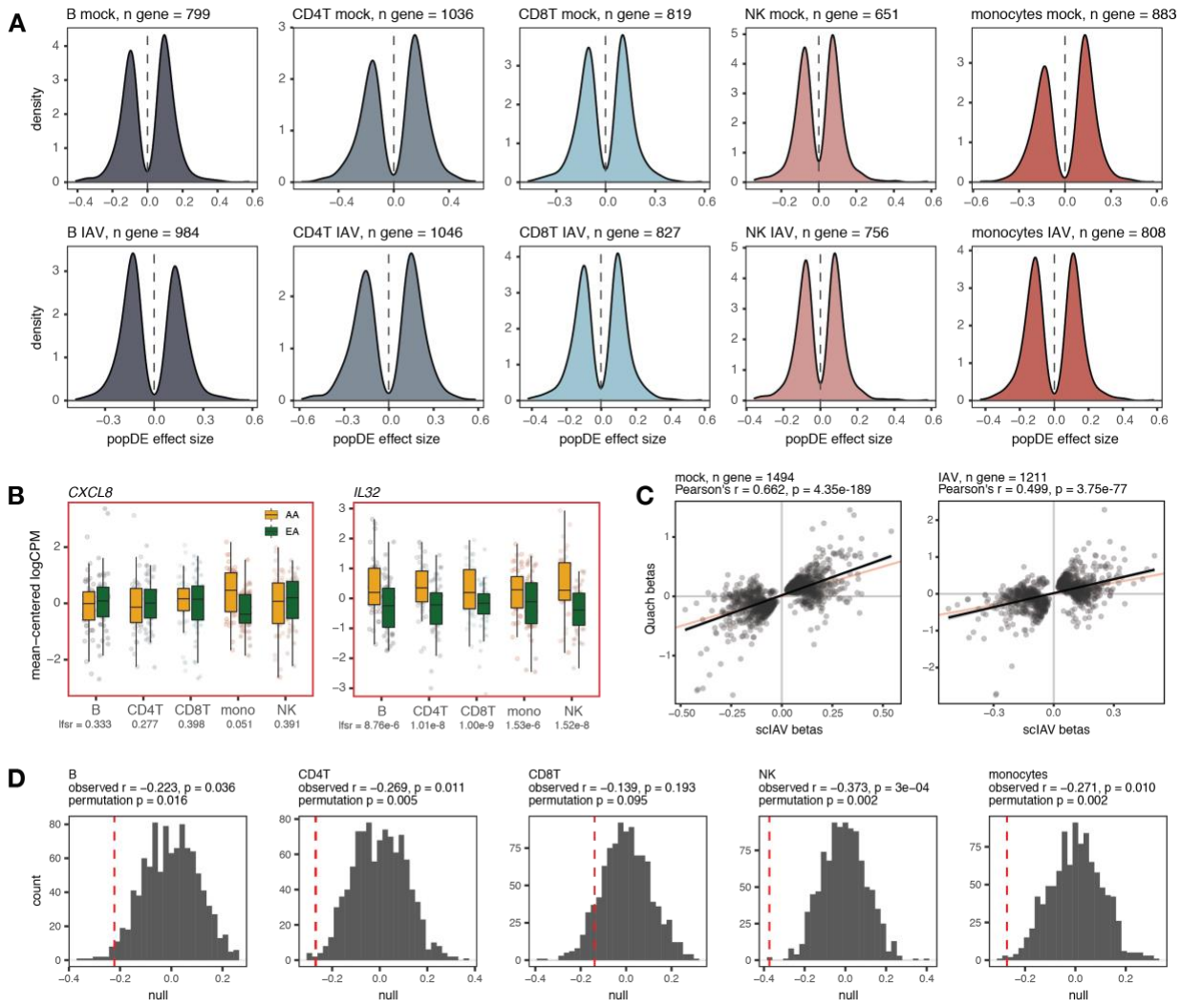


Fig. S3-2. Population-associated expression patterns. (A) PopDE effect size distributions among popDE genes (lfsr < 0.10) detected in each cell type in the mock condition (top row) and following IAV infection (bottom row). Overall, popDE effects are balanced with respect to sign across cell types and conditions. (B) Examples of cell type-specific (*CXCL8*, monocytes lfsr = 0.051, lfsr > 0.25 in all other cell types) and shared (*IL32*, lfsr < 8.8×10^{-6} in all cell types) popDE genes (AA in yellow, EA in green) in the IAV-infected condition. (C) Correlation between the popDE effect size estimates in monocytes in our single-cell IAV data (x-axis) and the Quach et al. data (Quach et al. 2016) (y-axis) among significant popDE genes (lfsr < 0.10) in our single-cell IAV dataset. Black line shows the best-fit line from a linear model, orange line shows the x = y line. P-values obtained from linear regressions. These correlations are strong, particularly when considering the inherent differences in experimental design between the two studies that are expected to contribute to variation in popDE effect sizes between the datasets (detailed in "Comparison with Quach et al. data"). (D) Observed Pearson correlation coefficients between the proportion of African genetic ancestry and IFN score following IAV infection (red dotted line) compared to null expectations when permuting the genetic ancestry vector across individuals (n

Fig. S3-2, continued. permutations = 1,000, density distribution in gray) for each cell type. Observed correlation coefficients and p-values from linear regressions are listed at the top of the plot, along with the p-values obtained from permutations. For four out of the five cell types tested (all except $CD8^+$T cells: p = 0.095), the observed Pearson correlation coefficient is significantly lower than most values obtained from permutation (B p-value = 0.016, $CD4^+$T = 0.005, NK = 0.002, monocytes = 0.002). Together, this analysis suggests that the association between genetic ancestry and IFN score in the IAV-infected condition is robust, although less so for $CD8^+$T cells.

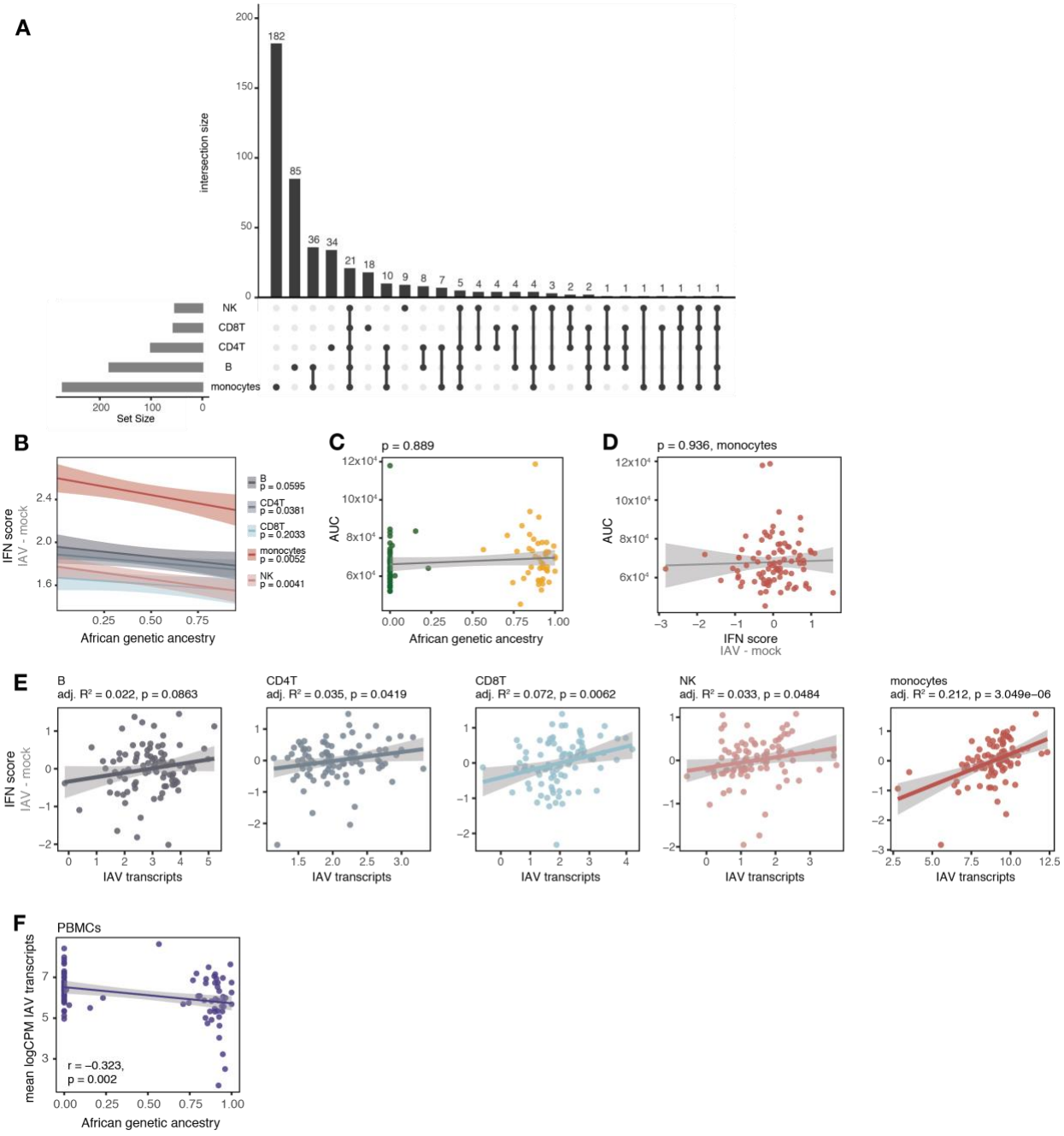


Fig. S3-3. Population-associated responses to IAV infection. (A) Sharing of popDR genes across cell types. (B) Correlation between the proportion of African genetic ancestry (x-axis) and IFN response (y-axis) across individuals (mean Pearson's r across cell types = -0.23, Fisher's meta-p = $6x10^{-5}$). (C) Correlation between the proportion of African genetic ancestry (x-axis) and baseline levels of IAV-specific serum IgG antibodies (y-axis). Anti-A/Cal/04/09 antibody titers were quantified using 4-fold serial dilutions for each individual's serum (total of eight dilutions per sample). Dilution and ELISA absorbance values were used to generate an area under the curve

Fig. S3-3, continued. (AUC, y-axis), which was used to summarize the levels of IAV (A/Cal/04/09)-specific serum IgG antibodies detected in each individual. (D) Correlation between IFN response (x-axis) and baseline levels of IAV-specific serum IgG antibodies (y-axis). (E) Correlation between IAV transcript expression (x-axis) and IFN response (y-axis) across individuals for each cell type. Higher IAV transcript expression is associated with stronger IFN responses in $CD4^+$ T cells, $CD8^+$ T cells, monocytes, and NK cells, with monocytes showing the strongest relationship (adj. $R^2 = 0.212$, $p = 3.1x10^{-6}$). (F) African genetic ancestry (x-axis) is negatively correlated with IAV transcript expression (y-axis) (Pearson's $r = -0.323$, $p = 0.002$) in PBMCs. To assess the robustness of this association, we tested whether outliers drive the overall correlation between genetic ancestry and intracellular IAV transcripts in the full sample. We 1) excluded individuals with standardized |z-score| values for IAV transcript measurements > 3; and 2) removed individuals with mean logCPM IAV $transcripts_{IAV - mock} < 4$. In both cases, the relationship between genetic ancestry (Quach et al. 2016; Nédélec et al. 2016) and IAV transcript levels was unchanged (z-score approach: Pearson's $r = -0.284$, $p = 0.008$; hard threshold approach: Pearson's $r = -0.258$, $p = 0.016$), suggesting that outliers do not drive the association described in the main text. In (B) - (F), p-values and best-fit slopes were obtained from linear regression models.

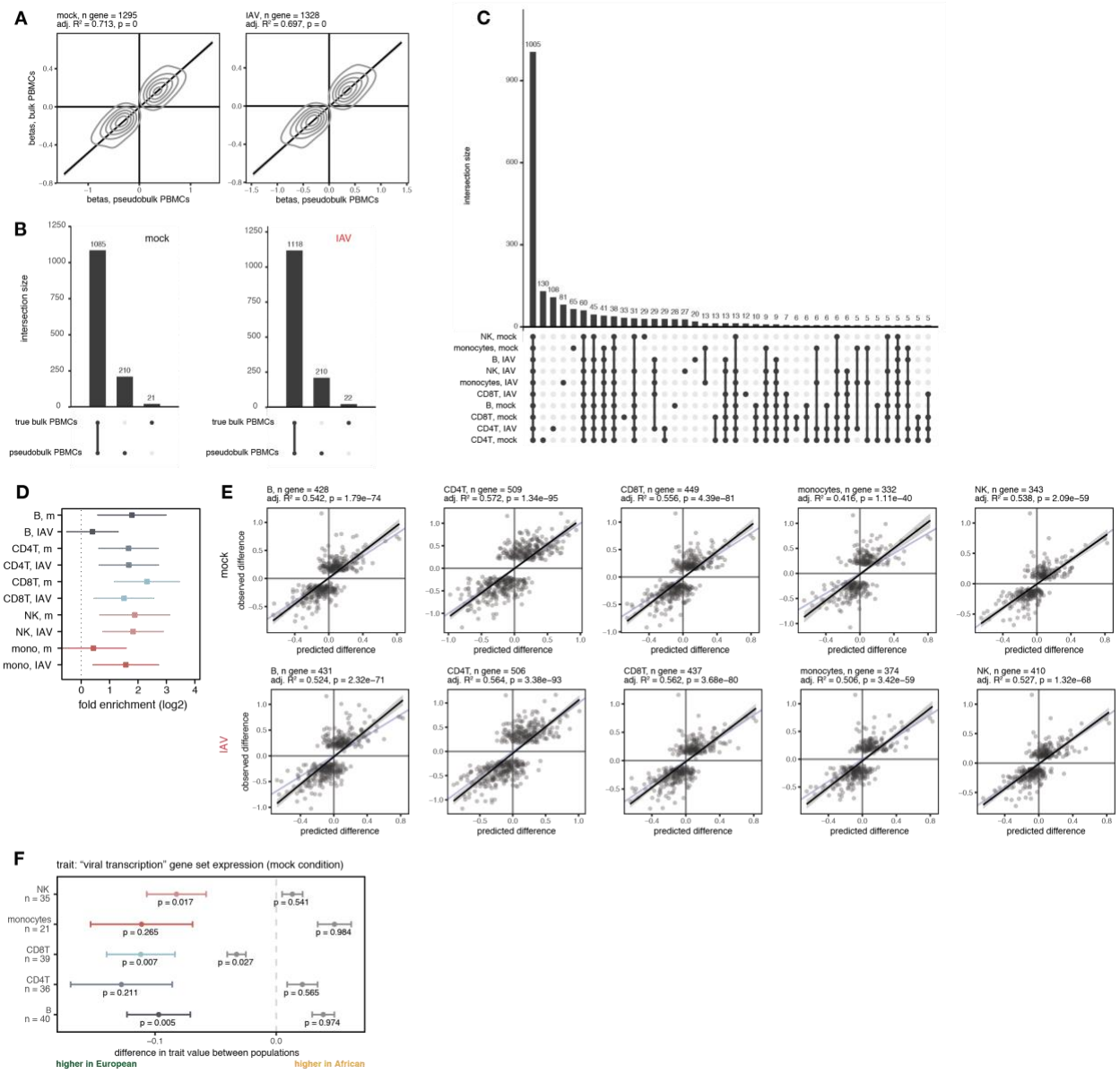


Fig. S3-4. *Cis*-genetic effects regulate gene expression variation. (A) Correlation between the eQTL mapping effect size estimates in the pseudobulk PBMC expression data (x-axis) and in the true bulk PBMC expression data (y-axis) among gene-SNP pairs with an eQTL effect in the pseudobulk data. P-values and best-fit slopes (black line) were obtained from linear regression models. (B) Sharing of eGenes in the pseudobulk PBMC expression data and the true bulk PBMC expression data for the mock (left) and IAV (right) conditions. (C) Sharing of eGenes across cell types and treatment conditions. (D) Enrichment of eGenes (lfsr < 0.10) among IFN-associated popDE genes (lfsr < 0.10 and in the Hallmark IFN-α/IFN-γ response pathway gene sets) identified in each cell type and condition ($\log_2$ fold enrichment with a 95% confidence interval, logistic regression; "m" = mock). (E) Correlation of the *cis*-predicted population differences in expression (x-axis) versus the observed population differences in expression (y-axis) among popDE genes with an eQTL across all cell types in the mock-exposed condition (top row) and IAV-infected

Fig. S3-4, continued. condition (bottom row). The black line shows the best-fit line from a linear model, and the blue line shows the x = y line. P-values were obtained from linear regression models. (F) Example term showing the effect of *cis*-SNP regression. In the mock condition, European-ancestry individuals display higher expression (median observed ancestry-associated difference < 0, colored points ± SE) of the genes belonging to the "viral transcription" term in the observed data. *Cis*-SNP regression (gray points ± SE) reduces this effect.

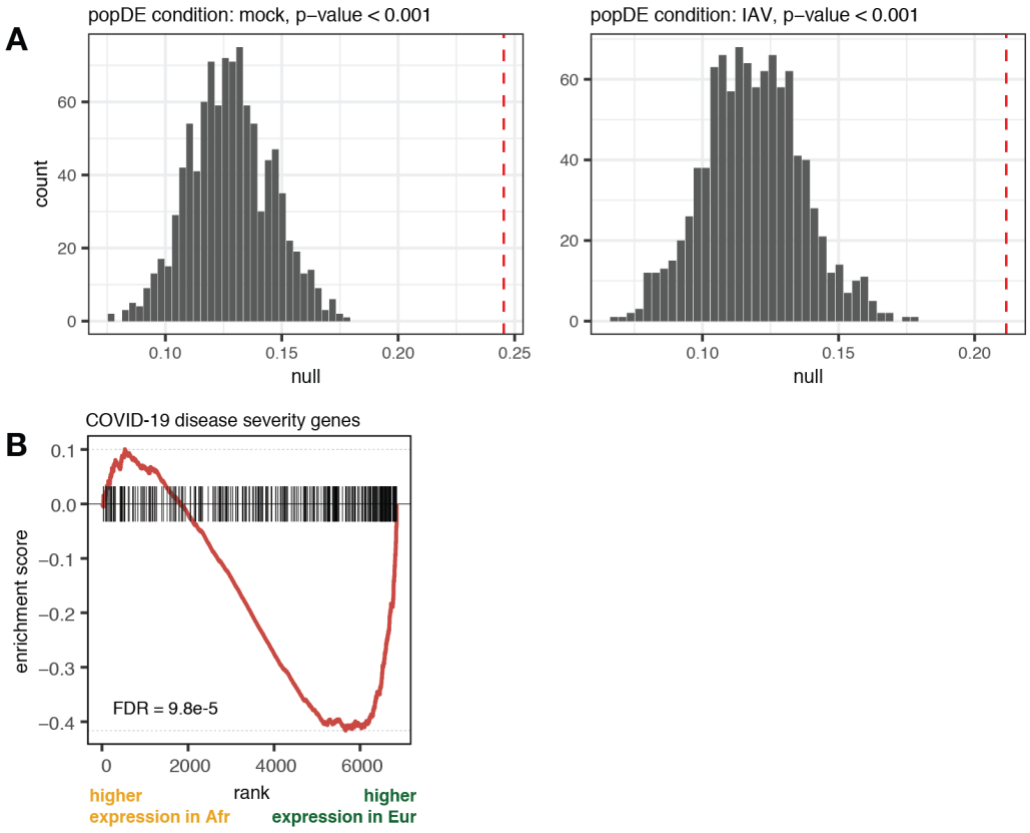


Fig. S3-5. COVID-19 severity-associated genes are enriched for genes differentially-expressed between populations. (A) Proportion of COVID severity-associated genes in monocytes that are popDE in monocytes (red dotted line) in both the mock (left) and IAV-infected (right) conditions compared to random expectation (gray, density distribution) when sampling the same total number of genes 1,000 times from all genes tested in the popDE analysis. (B) Barcode enrichment plot of genes positively associated with severity in monocytes, where popDE effect sizes at baseline (mock condition) are ranked from most positive to most negative (x-axis).

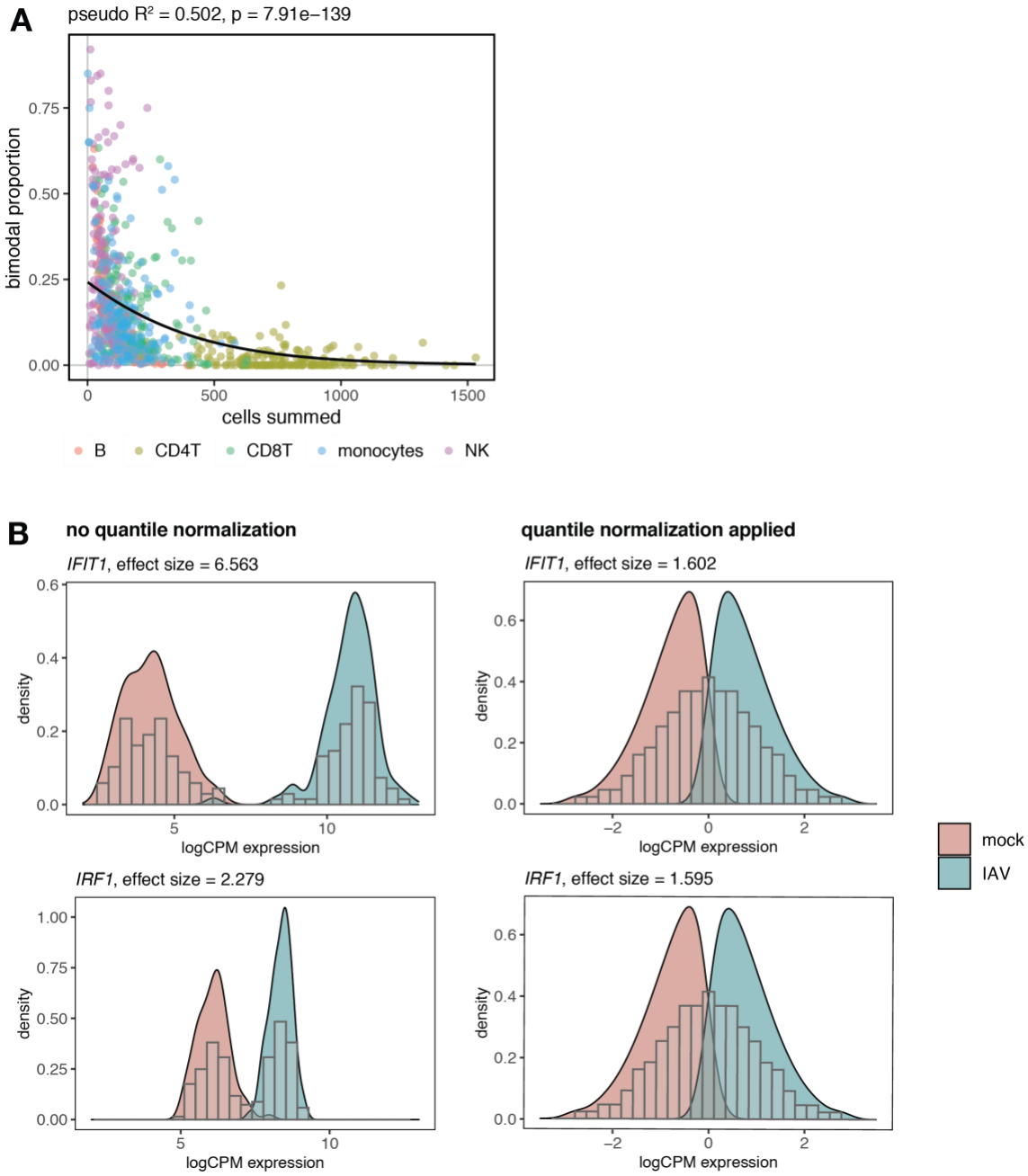


Fig. S3-6. Bimodal proportion and quantile normalization example. (A) Correlation between the number of cells summed to obtain pseudobulk gene expression estimates (x-axis) and the calculated bimodal proportion (y-axis) per sample for each cell type. Best-fit line (black) and p-value were obtained from a beta regression model. (B) Both *IFIT1* and *IRF1* display non-overlapping expression distributions between conditions prior to quantile normalization (left), although *IFIT1* shows a much stronger upregulation following IAV infection compared to *IRF1*. After quantile normalization (right), the transformed gene expression distributions for *IFIT1* and *IRF1* are very similar, and therefore, the estimated infection effect sizes are virtually equivalent.

Supplementary Tables for Chapter III

Table S3-1. Sample meta data. Available as an excel file online (Randolph et al. 2021). Description of cohort samples, including demographic information about donors, technical batch/variable information, and experimental information. Also included is a summary of the exome-sequencing results and kernel density estimation bandwidth parameters for the bimodal proportion calculations.

Table S3-2. Global infection effects. Available as an excel file online (Randolph et al. 2021). IAV infection differential expression effect for each gene in each cell type is reported, including effect size estimates, p-values, FDRs (Benjamini-Hochberg and permutation-based), and t-statistics.

Table S3-3. Global infection DE enrichments. Available as an excel file online (Randolph et al. 2021). Gene ontology enrichments for the differential expression effects are reported for each cell type. Term-specific p-values, FDRs (Benjamini-Hochberg-adjusted), enrichment scores, and leading edge gene sets are included.

Table S3-4. Ranked specificity scores and enrichments. Available as an excel file online (Randolph et al. 2021). Specificity score calculated for each gene. Also included are the gene ontology enrichments for ranked specificity scores. Term-specific p-values, FDRs (Benjamini-Hochberg-adjusted), and enrichment scores are reported.

Table S3-5. PopDE effects. Available as an excel file online (Randolph et al. 2021). Population differential expression effect for each gene in each cell type-condition combination is reported, including local false sign rate (lfsr), effect size estimate, and standard deviation of the effect size estimate.

Table S3-6. PopDE effect enrichments. Available as an excel file online (Randolph et al. 2021). Hallmark enrichments for the population differential expression effects are reported for each cell type-condition combination. Term-specific p-values, FDRs (Benjamini-Hochberg-adjusted), enrichment scores, and leading edge gene sets are included.

Table S3-7. PopDR effects. Available as an excel file online (Randolph et al. 2021). Population differential response effect for each gene in each cell type is reported, including local false sign rate (lfsr), effect size estimate, and standard deviation of the effect size estimate.

Table S3-8. eQTL effects. Available as an excel file online (Randolph et al. 2021). eQTL effect for the lead *cis*-SNP per gene is reported across cell types, including local false sign rate (lfsr), effect size estimate, and standard deviation of the effect size estimate.

Table S3-9. GO enrichments for popDE genes with an eQTL. Available as an excel file online (Randolph et al. 2021). Gene ontology enrichments for the popDE genes with evidence of an eQTL. Term-specific p-values and FDRs (Benjamini-Hochberg-adjusted) are included.

Table S3-10. WOS effects. Available as an excel file online (Randolph et al. 2021). Results from re-analysis of the Su et al. data (Su et al. 2020). WOS effect (i.e., COVID-19 severity score association) for each gene is reported across cell types, including local false sign rate (lfsr), effect size estimate, and standard deviation of the effect size estimate.

Table S3-11. Gene expression principal components regressed in the eQTL analysis. PCs regressed and number of significant eQTL per cell type and condition are reported.

Cell type	Regressed PCs (mock)	Regressed PCs (IAV)	No. genes under 0.10 FDR (mock)	No. genes under 0.10 FDR (IAV)
$CD4^+$ T	1 to 4	1 to 2	1377	1176
B	1 to 6	1 to 3	152	196
NK	1 to 2	1 to 2	68	76
monocytes	1 to 10	1 to 7	265	251
$CD8^+$ T	1 to 6	1 to 4	204	178
pseudobulk PBMCs	1 to 6	1 to 3	2095	1809
true bulk PBMCs	1 to 2	1 to 3	100	105

Supplementary Tables for Chapter IV

Table S4-1. Gene expression principal components regressed in the eQTL analysis. PCs regressed and number of significant eQTL per cell type and condition are reported.

Cell type	Regressed PCs (control)	Regressed PCs (COVID)	No. genes under 0.10 FDR (control)	No. genes under 0.10 FDR (COVID-19)
$CD4^+$ T	1 to 4	1 to 11	1458	380
B	1 to 7	1 to 8	339	94
NK	1 to 6	1 to 8	131	106
$CD14^+$ monocytes	1 to 6	1 to 11	447	910
$CD16^+$ monocytes	1 to 5	1 to 4	37	59
$CD8^+$ T	1 to 4	1 to 4	479	155